Is New Economic Geography Right?
Evidence from Price Data

Jessie H. Handbury*
Columbia University

David E. Weinstein*
Columbia University and NBER

November 14, 2010

Abstract

The agglomeration force behind the New Economic Geography literature initiated by Krugman is based on the notion that larger markets should have a lower variety adjusted price index. Despite his Nobel Prize, there have been no tests of this idea. This paper represents the first such test. Using a rich dataset covering 10-20 million purchases of grocery items, we find that after controlling for store and shopping effects: 1) Aggregate grocery prices are lower in larger cities; 2) Residents of larger cities have access to substantially more varieties than residents of smaller cities; and 3) These forces combine to substantially lower variety adjusted prices in large cities. In short, Krugman is right.

*We wish to thank Donald Davis, Mine Senses, Jonathan Vogel, and Joan Monras for excellent comments.

1. Introduction

In awarding Paul Krugman the Nobel Prize in economics for his work for international trade and economic geography, the committee characterized his contribution to economic geography as follows: "The new economic geography initiated by Krugman broke with … tradition by assuming *internal* economies of scale and imperfect competition. Agglomeration is then driven by *pecuniary externalities* mediated through market prices as a large market allows greater product variety and lower costs [emphasis in original]."[1] It is interesting to note, however, that, despite this award, there have been no tests of whether larger markets actually do have greater product variety and whether differences in the number of available varieties are sufficient to lower costs. This paper is the first to fill this gap by testing empirically whether large markets are actually characterized by lower prices of traded goods, and more importantly, whether this, in conjunction with greater product variety lowers costs for consumers. In other words, is the fundamental mechanism underlying new economic geography models correct?

One might suspect that we should have an answer to this question by now. Over the last two decades, there has been a burgeoning literature that has aimed to test the implications of New Economic Geography (NEG) models (see the recent excellent surveys by Brakman, Garretsen, and van Marrewijk [2009], Fujita and Mori [2005], and Combes, Mayer, and Thisse [2008]). A common theme of these tests has been that they have examined predictions of the theory – *e.g.*, home market effects, multiple equilibria, patterns of spatial agglomeration, and wage distribution – but they left untouched the question of whether the agglomeration mechanism Krugman postulated to be behind these predictions was at work.

This paper deviates from prior work by using highly disaggregated barcode data covering hundreds of thousands of goods purchased by 33,000 households in 49 cities in the US to

[1] Prize Committee of the Royal Swedish Academy of Sciences (2008).

measure product variety and the variety-adjusted costs faced by consumers in different locations. This permits the precise estimation of the pecuniary externalities hypothesized by Krugman.

NEG models generate externalities from a very simple structure. Trade costs are positive between cities, so locally produced goods must be cheaper than goods produced elsewhere. Further, if trade costs are prohibitive, larger cities will have more varieties than smaller ones. These two forces will combine to generate a lower price index in larger cities because more goods are produced locally in production centers[2]. This lower price index induces people to move into cities enhancing the scale economy.

The notion that prices are lower in larger cities has not been supported by data examined in previous work. In a seminal paper estimating a general equilibrium model of city size, Roback (1982) identified a cost-of-living premium using U.S. housing prices and, more recently, DuMond *et al.* (1999) confirmed this result using aggregate U.S. cost-of-living indices. Similarly, Tabuchi (2001) finds that aggregate consumer price indices, land values, and housing costs are all higher in larger Japanese cities.

One reason that these studies have not been deemed fatal for the theory is that it is easy to modify NEG models to generate higher housing prices in cities. Helpman (1998), for example, presents an NEG model in which prices of tradable goods are lower in cities drawing residents in but as enough people move into the city the price of housing rises, causing the aggregate price level to equilibrate across locations. Thus, his model features higher land prices in cities, lower traded goods prices, but equal aggregate prices. NEG models can also generate higher aggregate

[2] In the original Krugman (1991) paper more varieties are produced in a large market but with non-infinite trade costs all varieties are available to consumers in both large and small markets. We will specifically refer to the fact that more varieties are available to consumers in large markets as "the variety effect." This can arise in the original Krugman (1991) paper if some goods are traded at infinite cost, as in Helpman (1998), or if there is quadratic linear demand and/or a fixed cost associated with distributing each good to each city, as in Ottaviano, Tabuchi, and Thisse (2002).

consumer price indices in cities. Suedekum (2006), for example, shows how the Krugman (1991) model can generate higher non-traded goods prices by including a non-traded home goods sector in every city.

Secondly, prior empirical studies suffer from the problem of not carefully comparing identical goods. For example, Bils and Klenow (2001) and Broda, Leibtag, and Weinstein (2009) find that wealthier households purchase substantially more expensive varieties of the same narrowly defined good – *e.g.*, milk, eggs, water – than poorer households even in the same store. Similarly, Broda and Weinstein (2008) find that most of the variation in average prices paid by wealthy and poor households for goods like milk, eggs, and water is due to differences in the varieties consumed not the prices of the underlying varieties. To the extent that stores in wealthier neighborhoods cater more to local clientele, this is likely to bias studies based on non-identical samples of goods to find higher prices in wealthier locations.

Our paper uses these insights to make a number of contributions. Consistent with earlier analyses, we show that the price index for *identical goods* rises with city size. However, we also demonstrate that if one controls for the household making the purchase and the amenities of the store in which the purchase is made this result is not robust. In fact, we find that prices for the same good purchased are actually *lower* in larger cities once we control for these forces. This is the first empirical confirmation of Krugman's conjecture about how the prices of traded goods vary with city size. Moreover, if we control for the fact that prices in larger cities embody land prices, we find that prices net of land costs are even more negatively correlated with city size as Helpman (1998) hypothesized.

Second, our study is the first to document that the variety of traded goods available for consumption is substantially higher in larger cities. The elasticity of the number of products with

respect to city size is a whopping 0.2-0.3 (depending on the specification) which means that there is enormous variation in the number of available varieties across cities. For example, residents of New York (population 9.3 million) can choose between 97,000 different types of groceries, whereas residents of Des Moines (population 456,000) only have access to 32,000 varieties. This greater availability of varieties in larger cities means that variety-adjusted costs are likely to be substantially lower in large cities, as NEG models predict, setting the stage for our econometric exercise seeking to quantify the importance of the availability of varieties for welfare.

To do this estimation, we construct a variety-adjusted exact price index for each city in our sample. Our results show that, while the prices of identical goods are higher in larger cities, the variety effect more than offsets this price effect resulting in similar variety-adjusted costs across cities. Since the prices of identical goods are lower in larger cities when we control for purchaser characteristics, the variety-adjusted costs are substantially lower for a consumer in a large city than for a consumer sharing the same characteristics in a small city. For example, a household that moved from Des Moines to New York and purchased goods from the same type of stores in the same types of neighborhoods in the two cities would realize a 9 percent drop in the overall cost of its grocery purchases. This suggests that the price effects hypothesized by Krugman are indeed important.

The rest of the paper is structured as follows. Section II describes the data we use and develops some stylized facts about purchasing behavior and variety availability that motivates our modeling choices. Section III develops the price index theory and econometrics underlying our estimation strategy. We present our results in Section IV, and Section V concludes.

II. Data

The primary dataset that we use is taken from the AC Nielsen HomeScan database. This data was collected by AC Nielsen in a demographically representative sample of approximately 33,000 households in 52 markets across the U.S. in 2005. Households were provided with Universal Product Code (UPC) scanners to scan in every purchase they made regardless of whether these purchases were made in a store with scanner technology[3]. We have the purchase records for grocery items, which include purchase quantity, price, date, store name[4], product "module" (*i.e.*, a detailed description of the type of good, e.g. diet soft drink), as well as demographic information for each household making the purchase. For much of the analysis, we will be working with "brand-modules" which correspond to all the UPCs within a module that are marketed under a particular brand, *e.g.*, "Fanta-diet carbonated sodas." We will also use product "groups," a more aggregate product categorization provided by AC Nielsen. Each product "module" fits within a unique product "group" (*i.e.*, the "diet carbonated sodas" and the "regular carbonated sodas" product modules both fit within the carbonated sodas product group). Detailed descriptions of these data and the sampling methods used can be found in Broda and Weinstein (2010).

A major advantage of these data relative to the data used in previous studies comparing cost-of-living indices across cities is that we can compare prices of identical goods to directly

[3] In cases where panelists shop at stores without scanner technology, they report the price paid manually. Since errors can be made in this reporting process, we discard any purchase records for which the price paid was greater than twice or less than half the median price paid for the same UPC, approximately 250,000 out of 16 million observations. Einav, Leibtag, and Nevo (2008) address other concerns with the credibility of the Homescan data and find that the overall accuracy of the data is in line with other surveys of this type.

[4] We do not use store name dummies for the smallest stores in our dataset because all of the store's average price differences in a city would be assigned to its amenities. We therefore included close to 100 store dummies for stores that had over 100,000 dollars in sales in our sample. The majority of these are located in 15 or more cities, and all in at least 2 cities. For smaller stores, we used "channel ID" dummies that corresponded to whether the store was grocery store, mass-merchandiser, drug store, convenience store, or club store. Thus for small stores, we assume that stores within these types of categories are homogeneous across cities.

test the prediction that traded goods are on average cheaper in agglomeration locations. We can further use the household level demographic data and store names to separate how much of the inter-city price differential arises from factors that are not considered in the NEG models, such as differences in demographics, shopping behavior, and/or store amenities. Although the AC Nielsen dataset contains data for 52 markets, we classify cities at the level of Consolidated Metropolitan Statistical Area (CMSA) where available, and the Metropolitan Statistical Area (MSA) otherwise. For example, where AC Nielsen classifies urban, suburban, and ex-urban New York separately, we group them all as New York-Northern New Jersey-Long Island CMSA. There are two cases in which AC Nielsen groups two MSAs into one market. In these cases, we count the two MSAs as one city, using the sum of the population and manufacturing output and the population-weighted mean land value. We use population, income distribution, and racial and birthplace diversity data from the 2000 U.S. Census, manufacturing output data from 2007 U.S. Economic Census[5], and 2000 land values from Davis and Palumbo (2007). Matching the AC Nielson data with the Davis and Palumbo data reduces the number of cities in our analysis to 37.

Before turning to the theory and estimation, we characterize two stylized facts of the raw data that motivate our approach:

Stylized Fact 1: Consumers Appear to have "Ideal Types" of Each Variety

Our dataset includes every food UPC purchased by each household in our sample. Table 1 shows that the typical household purchases only 341 distinct UPC's during a continuous 250

[5]The main results reported below hold when we use either population or food manufacturing output as our measure of city size and all three measures are highly correlated. The data for food manufacturing output contains only the reported output of firms whose primary business is food manufacturing, and not restaurants and retail stores that also produce barcoded products that appear in our sample. The tables below, therefore, report results using population, our preferred measure of city size.

day period[6]. If we focus on one-person households, we see that these households buy, on average, only two different UPCs in a module conditional on making a purchase in that module. This suggests that they do not purchase a very diversified set of goods as love of variety models postulate.

One possible explanation for this behavior is that package sizes are too large to allow low levels of consumption of many goods. For example, consumers may want to consume every variety of soda available to them but they are constrained by the fact that it is difficult to purchase less than, say, 12 ounces of soda. If this were true and consumers had a constant elasticity of substitution (CES) style taste for variety without quality differences, one would not expect to see many repeat purchases of the same good by the same household. However, the data indicates that, on average, a household that purchases at least one unit of a UPC actually purchases 4 units of the same UPC. Moreover, if we broaden the definition a good to include all the UPC's in a brand within a module (e.g. count 6-packs of diet Coca-Cola and one-liter bottles of diet Coca-Cola as the same good), a household purchases 6 units of each good they purchase. This implies that the limited number of varieties each household purchases cannot be explained by packages being simply too large for consumers to purchase several different varieties in a 250 day period.

The flip side of these repeat purchases is that households that purchase a particular variety of a good tend not to purchase other varieties of the same good. If we again define a variety as a particular brand-module, we find that, in 60 percent of the cases, consumers only purchase one brand-module in a module. In other words, the raw data suggests that consumers,

[6] We define a period as "continuous" if it does not include any periods of more than 14 days over which we do not observe the household making a purchase.

on average, repeatedly purchase the same types of goods. Consumer behavior at the module level seems to follow more of an ideal-type structure than a love of variety structure.

Nevertheless, the heterogeneity in tastes across consumers seems to suggest that the love of variety model is not a bad depiction of *aggregate* consumer behavior. We can see this by examining the degree of overlap in the set of varieties purchased across households. If we randomly select a sample of households from the same city corresponding the smallest sample size collected any city (116 households), we find that typically 42 percent of the UPCs purchased by a household in a city are not purchased by *any* of the other 115 households in the sample.

These features of the data suggest that one should think of households as having heterogeneous ideal-type preferences, as opposed to the identical CES preferences that form the theoretical foundation for the variety-adjusted exact prices indices used in this paper. This discrepancy, however, is not a problem for our analysis if we think of consumers as having a logit demand system. In particular, Anderson, De Palma, and Thisse (1987) have demonstrated that a CES demand system can arise from the aggregation of ideal-type logit consumers. We will therefore follow Anderson, De Palma, and Thisse (1987) and use the CES structure to evaluate *aggregate* welfare even though we know that the discrete choice model is a better depiction of reality at the household level.

Stylized Fact 2: Consumers in Larger Cities Consume More Varieties

One simple way to see if households in larger cities have access to more varieties is to count them. Because AC Nielson tends to sample more households in larger cities, we need to make sure that the sample size is constant across cities. We therefore restrict ourselves to only looking at cities in which AC Nielsen sampled at least 450 households and compare the number

of varieties purchased by a random sample of 450 households in each of these cities. We plot the aggregate number of *different* UPCs purchased by all of these households against the size of the city in which the households live in figure 1. The results show a clear positive relationship between the variety of UPCs purchased in a city and the population of the city. This is certainly suggestive of the notion at the center of NEG models that the number of varieties available in a location rises with number of inhabitants in that location. However, it leaves open the question of whether these differences are likely to matter for consumers; that is, does the city size variety effect drive down the consumer price index in larger cities relative to small? We now turn to assessing this question.

III Empirical Strategy

Our main objective is to examine the relationship between city size and consumer costs by answering three separate, but related questions: 1) are traded goods prices lower in larger cities?; 2) are more varieties available in larger cities?; and 3) do the city size effects on prices and variety abundance combine to result in the "price index effect" predicted by the Krugman (1991) model?

III A. The Variety-Adjusted Price Index

Feenstra (1994) developed the variety-adjusted price index for the CES utility function. Here, we will modify it so that it can be used with our data. We begin by modeling the aggregate utility function as a nested CES utility function. To do this, we need to establish some notation. Let $g \in \{1,\ldots G\}$ denote the set of product groups, which we define in the same way as AC Nielsen, *i.e.*, sectors like "Crackers", etc. As in Broda and Weinstein (2010), the product group

of "Crackers", would contain brand-modules like "Nabisco-Premium-Flaked Soda Crackers" and "Pepperidge Farm Goldfish-Cheese Crackers." We let $b_g \in \{1,\ldots B_g\}$ denote the set of "brand-modules" within a product group.

We define U_{bgc} the set of all UPCs that have positive sales in city c in brand-module b, in product group g; U_{bg} the set of all UPCs that have positive sales *nationally* in brand-module b, in product group g; B_{gc} the set of all brands that have positive sales in city c in product group g; and B_g the set of all brands that have positive sales *nationally* in product group g.

It turns out that it is useful to measure the importance of a UPC's sales relative to several benchmarks. First, let s_{ubg} be the *national* expenditures on UPC u as a share of *national* expenditures on brand b in product group g, and let s_{ubgc} be a *city's* expenditures on UPC u as a share of the *city's* expenditures on brand b. We next define the *national* market share of city c's available UPCs in that particular brand-module (*i.e.*, $u \in U_{bgc}$) as

$$s_{bgc} \equiv \sum_{u \in U_{bgc}} s_{ubg} .$$

Thus, s_{bgc} tells us the expenditure share of UPCs within a brand that are available in a city using *national* weights. In other words, if the UPCs that constitute a large share of a brand's national sales are available locally, then s_{bgc} will be large. It will also be useful to define an analogous measure computed using *local* weights:

$$\tilde{s}_{bgc} \equiv \sum_{u \in U_{bgc}} s_{ubgc} .$$

Just as we defined shares of UPCs within brands, we also need to measure the importance of various brands. We do this in an analogous manner below;

$$s_{gc} \equiv \sum_{b \in B_{gc}} s_{bg} \text{ and } \tilde{s}_{gc} \equiv \sum_{b \in B_{gc}} s_{bgc} ,$$

where s_{bg} is the *national* market share of brand-module *b* in product group *g*, and s_{bgc} is the local market share of brand-module *b* in product group *g*.

We can now modify Broda and Weinstein's (2010) Proposition 1 as follows:

For $g \in G$, if $B_{gc} \neq \varnothing$, then the exact price index for the price of the set of goods *G* in city *c* relative to the nation as a whole that takes into account the differences in variety in the two locations is given by,

$$\mathrm{EPI}(\mathbf{p}_c, \mathbf{x}_c, U) = \prod_{g \in G} \left[\mathrm{CEPI}_{cg} \left(s_{gc} \right)^{\frac{1}{1-\sigma_g^a}} \prod_{b \in B_g} \left(s_{bgc} \right)^{\frac{w_{bgc}}{1-\sigma_g^w}} \right]^{w_{gc}},$$

where w_{bgc} and w_{gc} are log-ideal CES Kazuo Sato (1976) and Yrjo Vartia (1976) weights defined as follows:

$$w_{gbc} = \frac{\dfrac{\tilde{s}_{bgc} - s_{bg}}{\ln \tilde{s}_{bgc} - \ln s_{bg}}}{\displaystyle\sum_{b \in B_g} \left(\frac{\tilde{s}_{bgc} - s_{bg}}{\ln \tilde{s}_{bgc} - \ln s_{bg}} \right)} \text{ and } w_{gc} = \frac{\dfrac{\tilde{s}_{gc} - s_g}{\ln \tilde{s}_{gc} - \ln s_g}}{\displaystyle\sum_{g \in G} \left(\frac{\tilde{s}_{gc} - s_g}{\ln \tilde{s}_{gc} - \ln s_g} \right)},$$

σ_g^a is the elasticity of substitution across brand-modules in product group *g*, σ_g^w is the elasticity of substitution among UPCs within a brand-module, s_g is the share of product group *g* in national expenditures, and CEPI_{cg} is the conventional exact price index associated with the CES utility function for product group *g* and city *c* relative to the national sample.

This index has been used extensively in the literature (see Feenstra [forthcoming] for a survey of its use), so we will only briefly discuss its properties here. While the CEPI_{cg} measures the prices of products available in city *c* relative to their average price nationally, the second and third terms within the brackets adjust the exact price index to account for the fact that not all varieties, respectively defined as brands and UPCs, available nationally are available in city *c*. Each variety adjustment is calculated using an elasticity of substitution parameter and the national expenditure share of unavailable varieties. One of the properties of this index is that the variety adjustments and, therefore, the price level will fall as more UPCs and brands are available in a city, and will fall faster when these additional varieties and brands are popular nationally. This property is due to the fact that the index gives more weight to the availability of

non-common varieties that comprise a larger share of expenditure in the market where they are available, in our case, the national market. The elasticity of substitution parameter weights varieties by how differentiated they are. As a result, the availability of more differentiated varieties matters more than the availability of varieties that are close substitutes to existing varieties. We obtain the elasticities of substitution computed for UPCs within a brand-module and across brand modules within a product group from Broda and Weinstein (2010).

III B. Adjusting Prices for Amenities and Buyer Characteristics

Our first challenge in identifying the "price index effect" is to measure the prices of goods consumed in more than one city – *i.e.*, "common goods." One of the complicating factors associated with measuring these prices is that store amenities and shopping behavior might vary systematically across locations. If, for example, bigger cities feature nicer stores, we might systematically bias our results against the NEG model. Similarly, wealthier households pay more for the same goods in the same stores (presumably because the opportunity cost of time spent shopping is higher). This suggests that we should purge the data of these effects.

Let P_{uzrh} be the price of UPC u, purchased in zip code z, in store r, by household h. We will refer P_{uzrh} as the "unadjusted price." We also define p_{uzrh} as $\ln(P_{uzrh})$. The simplest way to purge the price data of amenities and other effects that are not in the model is in a two step process, the first of which is to run a regression of the form:

$$(1) \qquad p_{uzrh} = \alpha_u + \delta_r + \sum_i \beta_i Z_{ih} + \gamma_1 \ln\left(Pop_z\right) + \gamma_2 \ln\left(Income_z\right) + \gamma_3 \ln\left(Land_z\right) + \varepsilon_{uzrh},$$

where Z_h denotes a series of household characteristics; Pop_z denotes the population of the city in which the UPC was purchased; $Income_z$ is the per capita income in the zip code where the UPC

was purchased; $Land_z$ is the price of land in the city where the UPC was purchased; and Greek symbols denote parameters to be estimated.

Equation (1) enables us to make a number of adjustments to the prices to control for various characteristics that can affect prices but are outside of the canonical Krugman model.

We construct the log demographic-adjusted price as

$$\ln\left(\tilde{p}^{D}_{uzrh}\right) \equiv \ln\left(p_{uzrh}\right) - \sum_{i} \hat{\beta}_i Z_{ih} \,. \tag{2}$$

$\tilde{p}_{uzrh}$ corresponds to prices purged of store effects and demographic characteristics of the buyers. This will be useful when we want to control for the fact that wealthier people tend to locate in larger cities.

We then purge prices of store effects in the log store-adjusted price defined as:

$$\ln\left(\tilde{p}^{DS}_{uzrh}\right) \equiv \ln\left(p_{uzrh}\right) - \sum_{i} \hat{\beta}_i Z_{ih} - \hat{\delta}_r \,. \tag{3}$$

For stores with more than $100,000 in sales, which includes national chains such as *Wal-Mart*, *Kroger*, and *CVS*, we estimate δ_r for each chain and purge prices of chain-specific store amenities. For smaller stores, however, we restrict δ_r to be the same for all stores of the same type, where type is defined in one of seven categories. Grocery, drug, mass merchandiser, supercenter, club, convenience, and other. Here, we are accounting for the fact that products might be distributed through different channels in larger cities than smaller cities, *e.g.* through convenience and grocery stores as opposed to super centers and club stores.

Broda, Leibtag, and Weinstein (2009) found that stores in more expensive neighborhoods charge more for the same UPCs presumably because they face a more inelastic demand curve. We account for this possibility defining the following income-adjusted price

(4) $$\ln\left(\tilde{p}_{ucrh}^{DSI}\right) \equiv \ln\left(p_{ucrh}\right) - \sum_i \hat{\beta}_i Z_{ih} - \hat{\delta}_r - \hat{\gamma}_2 \ln\left(Income_z\right)$$

These prices will be useful when we want to consider the shift in utility that would occur if a household with a particular set of demographic characteristics living in a particular type of neighborhood moved to a comparable neighborhood in another city.

Finally, since even the prices of tradables contain a non-traded component, we may also wish to consider prices that have also been purged of land values as well. We define land-value adjusted prices as

(5) $$\ln\left(\tilde{p}_{ucrh}^{DSILV}\right) \equiv \ln\left(p_{ucrh}\right) - \sum_i \hat{\beta}_i Z_{ih} - \hat{\delta}_r - \hat{\gamma}_2 \ln\left(Income_z\right) - \hat{\gamma}_3 \ln\left(Land_z\right)$$

These prices will give us a sense of how the prices of purely traded goods vary across cities before land rents are included in the retail prices.

III C. Measuring Variety Availability in Cities

Measuring s_{gc} and s_{bgc} provides a particular challenge because we do not observe all of the varieties available in each city. Our estimates of s_{bgc} and s_{gc} are likely to be affected by a downward bias on our estimates of U_{bgc} and U_{gc}. The bias is likely to emerge from two sources. First, some cities have smaller samples of households than other cities, so there will be a natural tendency for the counts of UPCs purchased in each city to vary with the sample size in the city. Second, even if the sample sizes were identical in every city, the sample count of UPCs will be below the true count of UPCs because some goods will not be purchased by our sample of households but will be purchased by other households in the city.

Our approach to this problem is to break it into two components. We begin by developing two methodologies for estimating the number of UPCs and brand-modules available in a city, the

size of the sets U_{bgc} and U_{gc} respectively, using purchase records for only a sample of households in a city. The first is parametric and the second is non-parametric, but we will show that both methodologies yield very similar estimates for how the number of varieties and brands varies across cities, which is a critical component in understanding whether the NEG model is valid.

Unfortunately, there is not a theoretical underpinning for parametrically estimating s_{gc} and s_{bgc}. Nevertheless, these can be estimated using a non-parametric approach. We therefore rely on the fact that, in the standard NEG models (with no quality variation across varieties), s_{gc} and s_{bgc} are proportional to the number of varieties available and the fact that both parametric and nonparametric estimation yield very similar estimates of the counts of varieties to motivate our nonparametric estimation of s_{gc} and s_{bgc}.

The methodology that we will employ for estimating the number of products available in a city was initially developed by biostatisticians for estimating the number of distinct species in a region. Since this methodology is largely unfamiliar for economists, it is useful to go through some of the intuition here before presenting the full statistical model. Initially, let us assume that each household selects only one UPC out of the S available UPCs available in the market. If we also assume that each UPC is purchased with a probability π that is identical for all UPCs in the market, then it follows that $\pi = 1/S$.

Our task, now, is to estimate S using the number of different UPCs purchased by a sample of H households. To do this, we make one additional assumption: stores have sufficient inventories of goods so that the purchase of a UPC by one household does not reduce the probability of another household buying the same UPC. If household purchases are independent in the cross-section, then the probability that we observe one of the H households in our sample selecting a particular UPC is equal to one minus the probability that none of the H households

selects the UPC, or 1- $(1\text{-}1/S)^H$. The number of different UPCs that we expect to observe in the purchase records of the H households is simply the sum of these probabilities across all of the S available UPCs,

(6) $$S(H) = S\left(1-\left(1-1/S\right)^H\right)$$

It would be straightforward to obtain an estimate for the number of varieties in the market in this simple case. By equating $S(H)$ to the sample UPC count, $\tilde{S}(H)$, we can derive an estimate for the number of available UPCs, $\hat{S}$, which satisfies equation (6). Note that the distribution of $S(H)$ should follow the negative exponential function,

(7) $$S(H) = S\left(1-e^{-\ln\left(1-\frac{1}{S}\right)H}\right).$$

This simple approach cannot be applied to the data for two reasons. First, households purchase more than one UPC in the course of a year. And second, some UPCs, like milk, are likely to be purchased at higher frequencies than other UPCs, like salt. Hence the probability that we observe the purchase of a UPC will vary across UPCs. We can deal with the first problem by allowing the purchase probability π to differ from $1/S$. The second problem, however, is more complicated because solving it requires us to know not only the purchase frequencies of every observed UPC but also the frequency of purchases for UPCs that we do not observe in our sample.

In order to understand how we can solve this problem, let's consider another simple case in which there are two possible purchase frequencies, π_1 and π_2. We will refer to goods purchased at frequency π_1 as being in the first "incidence group" and goods purchased at frequency π_2 as being in the second incidence group. Without a loss of generality, we assume that the probability that any household purchases one of the UPCs indexed 1,…, S_1 is π_1 and the

probability that any household purchases one of the UPCs indexed S_1+1,...,S is $\pi_2 < \pi_1$, where $S = S_1 + S_2$ and $S_1\pi_1 + S_2\pi_2 = 1$. The probability that we will observe any one of the UPCs indexed 1,…, S_1 in a sample of H households' purchases is now $(1-(1-\pi_1)^H)$ and the probability that we will observe any one of the UPCs indexed S_1+1,...,S is $(1-(1-\pi_2)^H)$. We can now write the total number of different UPCs that we would expect to observe in the purchases of H households as

$$\begin{aligned} S(H) &= S_1\left(1-\left(1-\pi_1\right)^H\right) + S_2\left(1-\left(1-\pi_2\right)^H\right) \\ &= S\left[\alpha_1\left(1-\left(1-\pi_1\right)^H\right) + \alpha_2\left(1-\left(1-\pi_2\right)^H\right)\right], \end{aligned} \tag{8}$$

where $\alpha_1 = S_1/S$ and $\alpha_2 = S_2/S$ are shares of UPCs that fall in the two incidence groups. In this case, if we could obtain estimates for α_1, π_1, α_2, and π_2, we could estimate the total number of UPCs available in the market according to the following formula:

$$\hat{S} = \tilde{S}(H) \Big/ \left[\alpha_1\left(1-\left(1-\pi_1\right)^H\right) + \alpha_2\left(1-\left(1-\pi_2\right)^H\right)\right]. \tag{9}$$

Unfortunately, we cannot solve equation (9) for $\hat{S}$ using only the sample value for $\tilde{S}(H)$ (as we did with equation (6)). We now also need to obtain estimates of α_1, π_1, α_2, and π_2. Furthermore, obtaining estimates of the α's and π's is not straightforward because they need to be obtained from the mixture distribution governing the probability of jointly observing all of the UPCs in the sample. We will describe how we estimate this distribution in the context of the more general case that we consider in our analysis.

Following Mao, Colwell, and Chang (2005) we assume, as we did in the examples above, that groups of UPCs share the same selection probability. Suppose that each UPC u has a probability of $\pi_{c,u}$ of being selected by any one household in city c, and that there are K different incidence groups, or values that $\pi_{c,u}$ can take in each city c, such that $\pi_{c,u} = \pi_{c,k}$ for each UPC u in category k. We define $\alpha_{c,k}$ as the proportion of UPCs in city c that are selected with probability

$\pi_{c,k}$. For example, when residents of a particular city choose to buy a particular good, one might observe one purchase probability for dry goods, another one for perishable goods, a third one for cleaning supplies, etc.

We can now write the following equation relating the shares of each incidence group, the probability a UPC in that group is selected, and the total number of UPCs in the city as

$$S\sum_{k=1}^{K}\alpha_{c,k}\pi_{c,k}=1.$$

The number of UPCs we expect to observe in a sample of H_c households is

$$S_c(H_c)=S_c\sum_{k=1}^{K}\alpha_{c,k}(1-(1-\pi_{c,k})^{H_c}) \quad (10)$$

,

where Colwell, Mao, and Chang (2004) note that equation (10) can be re-written as:

$$S_c(H_c)=S_c\sum_{k=1}^{K}\alpha_{c,k}(1-\exp(C_{c,k}H_c)), \text{ where } C_{c,k}=-\ln(1-\pi_{c,k}).$$

It is straightforward to see that equation (7) is a specific case of equation (10) in which the selection probability is identical for all UPCs, *i.e.*, $K = 1$ and $\pi = 1/S$. Equation (10) can therefore be referred to as the "generalized negative exponential" (GNE) model.

Mao, Colwell, and Chang (2005) show how one can use maximum likelihood estimation to estimate S_c. The starting point is to realize that the number of households purchasing product u in city c, $h_{c,u}$, follows a binomial distribution with probability:

$$P\left(h_{c,u}\right)=\varphi\left(h_{c,u};\pi_{c,u}\right)=\binom{H_c}{h_{c,u}}\left(\pi_{c,u}\right)^{h_{c,u}}\left(1-\pi_{c,u}\right)^{\left(H_c-h_{c,u}\right)} \quad (11)$$

,

where H_c is the total number of households in the sample for city c,

$$\binom{H_c}{h_{c,u}}\equiv\frac{H_c!}{h_{c,u}!\left(H_c-h_{c,u}\right)!},$$

and once again $\pi_{c,u} = \pi_{c,k}$ for each UPC, *u*, in category *k*. Let $\{h_{c,u}\}_{u \in U_c}$,where U_c is the set of UPCs observed in the city *c* sample, be the observed *counts* of each UPC purchased by our sample of households in a city. Now, we can define the binomial mixture distribution as follows:

$$\Phi\left(h_{c,u}\right) = \sum_{k=1}^{K} \alpha_{c,k} \varphi\left(h_{c,u}; \pi_{c,u}\right).$$

This distribution tells us the probability of observing $h_{c,u}$ purchases of any UPC *u* in our data, regardless of its incidence group, given the size and purchase probabilities of each of the incidence groups.

Mao, Colwell, and Chang (2005) derive a maximum likelihood methodology for estimating the α's and π's for a given *K* using data on the number of samples (in our case, households) in which each variety is observed. The variable $n_{c,j}$ is defined to be the number of products that are purchased by *j* households in the dataset for city *c*, *i.e.*, for which $h_{c,u}$ equals *j*. In other words, if 100 UPCs are purchased by no households, 50 UPCs are purchased by 1 household, and 25 UPCs are purchased by 2 households, then we would have $n_{c,0} = 100$; $n_{c,1} = 50$; and $n_{c,2} = 25$. The joint likelihood of the total number of products available in the city, S_c, and the parameters of the mixture distribution is

$$L\left(S_c, \left\{\alpha_{c,k}, \pi_{c,k}\right\}_{k=1}^{K}\right) = \frac{S_c!}{\prod_{j=0}^{H_c}\left(n_{c,j}!\right)} \prod_{j=0}^{H_c} \Phi(j)^{n_{c,j}}$$

Note that from equation (9), we know that the number of available products, S_c, is a function of the number of observed products, $S_c(H_c) = \sum_{j=1}^{H_c} n_{c,j}$, and the parameters of the mixture distribution, $\left\{\alpha_{c,k}, \pi_{c,k}\right\}_{k=1}^{K}$. Therefore, we only need to estimate the parameters of the mixture distribution to derive an estimate for the number of available products. To do so, Mao, Colwell,

and Chang (2005) maximize a conditional likelihood function. Let $\tilde{\varphi}(h_{c,u};\pi_{c,u})$ be a zero-truncated binomial density, *i.e.* the probability that a product is purchased by *j* households conditional on it being purchased by more than one household, and $\tilde{\Phi}(j)=\sum_{k=1}^{K}\alpha_{c,k}\tilde{\varphi}(j;\pi_{c,k})$ be the mixture distribution over these densities for *K* incidence groups. If we denote the total number of UPCs that are purchased by at least one household in the sample $n_{c,+}$, the conditional likelihood function is

$$L\left(\left\{\alpha_{c,k},\pi_{c,k}\right\}_{k=1}^{K}\right)=\frac{n_{c,+}!}{\prod_{j=1}^{H_c}\left(n_{c,j}!\right)}\prod_{j=1}^{H_c}\tilde{\Phi}(j)^{n_{c,j}}$$

We estimate the parameters of mixture distributions for a range of values for *K* using the conditional likelihood function. Each distribution implies a different estimate for the total number of UPCs available in a city. We then compute the Akaike Information Criterion (AIC) for each value of *K* in each city. We choose between the distributions by selecting the number of incidence groups for all cities equal to the *K* that maximizes the sum of the AICs across all cities.[7]

Once we have our estimates the α's and π's, we can use equation (10) to obtain an estimate of the total number of varieties as follows:

$$\hat{S}_c=\tilde{S}_c(H_c)\left[\sum_{k=1}^{K}\hat{\alpha}_{c,k}(1-(1-\hat{\pi}_{c,k})^{H_c})\right]^{-1} \tag{12}$$

[7] We also assume that the sampling is sufficient so that we observe some UPCs purchased in each incidence group. We need this assumption because we cannot say anything about the number of goods in an incidence group that are purchased with such low a probability that no one in the sample ever buys them. In other words, just as we cannot answer how many angels can dance on the head of a pin, we can only discuss the number of available varieties for observable classes of goods.

where variables with circumflexes represent parameter estimates and $\tilde{S}_c(H_c)$ is equal to the sample count of distinct varieties in the city. It is useful to note that when the number of sampled households in the city, H_c approaches infinity, the fact that the α's sum to one implies that our count of the number of distinct products purchased by these households becomes our estimate of the number of varieties.

While the above approach is parametric in the sense that we assume the distribution of the count of varieties depends on the underlying probabilities and shares of the various incidence groups, there is a nonparametric approach that is often used and based on the estimation of "accumulation curves" (see Mao, Colwell, and Chang [2004, 2005]). We define an accumulation curve, $S(n)$, as the expected number of distinct UPCs purchased by a sample of n households. The total number of different varieties available equals the asymptote of the accumulation curve.

There is a well-developed methodology for estimating these curves. We first randomly order the households in our sample for a given city. We then count the number of unique UPCs purchased by the first household in the random ordering for a city and denote this count $S(1)$. Next, we take the data for the second household in our ordering and add it to the data for the first household in a hypothetical sample and count the number of unique UPCs in this sample, denoting it $S(2)$. We continue to add households to the hypothetical sample creating a series ($S(1)$, $S(2)$, $S(3)$,…, $S(H_c)$), where H_c is the total number of households in our sample for city c.

One of the problems of this approach is that each accumulation curve will be sensitive to the random order in which households are added to the curve due to both random error and sample heterogeneity. Colwell and Coddington (1994) note that the random error can be reduced by randomizing the sample order R times and generating an accumulation curve, $S_{c,r}(n)$, for each

random ordering indexed by *r*. The random error-adjusted accumulation curve is the mean of the species accumulation curves over the different randomizations,

$$\overline{S}(n) = \sum_{r=1}^{R} S_r(n) .$$

We can adjust for random error by using mean accumulation curves over 50 randomizations, *i.e.*, $R = 50$.[8]

An important feature of accumulation curves is that their value in the limit when *n* approaches infinity is an estimate of the total number of different goods available in the city.[9] Unfortunately, theory does not tell us what the distribution will be for these accumulation curves. Therefore, we will follow Jimenez-Valverde *et al.* (2006) by estimating the parameters of various functional forms and use the AIC goodness-of-fit test to choose between a range of functional forms that pass through the origin and have a positive asymptote.

We can use this accumulation curve methodology to estimate s_{gc} and s_{bgc}. Just as we can build an accumulation curve corresponding to the count of the different goods represented in a sample, we can also build a curve corresponding to the national market share of the different goods represented in a sample. We estimate the asymptotes of the two different *share*

[8] To determine how sensitive the adjusted accumulation curve is to sample heterogeneity, Colwell and Coddington (1994) suggest comparing the adjusted accumulation curve with the accumulation curve that we would expect to draw if all of the UPCs purchased by the households in the city sample were randomly assigned to households. We use the random placement method to calculate this expected accumulation curve and its variance as:

$$\tilde{S}(n) = S_{TOT} - \sum_{u=1}^{S_{TOT}} (1-\frac{n}{N})^{n_u}, \text{ and } \sigma^2(n) = \sum_{u=1}^{S_{TOT}} (1-\frac{n}{N})^{n_u} - \sum_{u=1}^{S_{TOT}} (1-\frac{n}{N})^{2n_u}$$

where S_{TOT} is the total number of UPCs recorded in the sample, n_u is the total number of households that purchase that UPC *u*, and *N* is the total number of households in the sample. If the expected curve, $\tilde{S}(n)$, rises more steeply from the origin than the mean curve, $\overline{S}(n)$, the sample heterogeneity is greater than that which could be explained by random sampling error alone. Coleman *et al.* (1982) state that the hypothesis of random placement will hold if the expected curve does not deviate by more than one standard deviation from the mean curve for more than one third of the sample sizes, *n*, and the deviations are distributed randomly across *n*. We checked that this is true in our data.

[9] One might think that it would be more appropriate to use the total number of households in the city instead of the asymptote to compute the number of goods available. This would be true if every good available in a city were purchased by at least one household, but not otherwise. In practice, since the number of households in a city is quite large, it does not matter whether one focuses on the asymptote or sets *n* equal to the number of households.

accumulation curves for each product group in each market. The first curve is used to estimate the national market share of the brand-modules that are available in a city within a product group, s_{gc}. The second is used to estimate the national market share of the UPCs that are available in a city within each brand-module that is available in the city, s_{bgc}.

Unfortunately, we are unable to estimate one value of s_{bgc} for each brand-module b in each product group g for each city c because it is highly unlikely that we observe every brand-module available in a city, *i.e.* every b in B_{gc}. Instead, we use a common s_{bgc} for all brand-modules within each product group which we denote $\overline{s}_{gc}$. This within-brand UPC share is an estimate for the national sales of UPCs available in a city divided by the national sales of brand-modules available in the city, *i.e.* the weighted average within-brand national market share of the UPCs available in a city conditional on the brand-module being available in the city. We calculate this estimate using an accumulation curve of the national sales of the UPCs available in a hypothetical sample of H households as a share of the national sales of brand-modules available in the city as a whole. For the denominator, we use national product group sales multiplied by our estimate for s_{gc}. This simplifies the variety adjustment in Proposition 1, so we can calculate the EPI for city c relative to the national sample as:

$$\text{EPI}\left(\mathbf{p}_c, \mathbf{x}_c, U\right) = \prod_{g \in G} \left[\text{CEPI}_{cg} \left(s_{gc}\right)^{\frac{1}{1-\sigma_g^a}} \left(\overline{s}_{gc}\right)^{\frac{1}{1-\sigma_g^w}} \right]^{w_{gc}}.$$

By rearranging terms, we can decompose the exact price index into a common exact price index for a city $CEPI_c$ and a variety adjustment, VA_c which we define below:

$$\text{(13)} \qquad \text{EPI}\left(\mathbf{p}_c, \mathbf{x}_c, U\right) = CEPI_c \times VA_c = \prod_{g \in G} \left[\text{CEPI}_{cg} \right]^{w_{gc}} \times \prod_{g \in G} \left[\left(s_{gc}\right)^{\frac{1}{1-\sigma_g^a}} \left(\overline{s}_{gc}\right)^{\frac{1}{1-\sigma_g^w}} \right]^{w_{gc}}.$$

Equation (13) will be useful for discussing the role played by new varieties in our analysis.

IV. Results

We begin our analysis by exploring how prices vary with city size. Table 2 presents results from estimating equation (1). The first column of the table demonstrates that if we simply regress prices on city size, we find that the prices of (common) UPCs tend to be higher in larger cities but not significantly so. As we argued earlier, however, the prices of identical goods may vary systematically with the store in which they are purchased (presumably because some stores offer better amenities). Controlling for store amenities and purchaser characteristics tends to further weaken the link between city size and prices. In other words, most of the apparently higher prices of goods in larger cities is due to the fact that people in larger cities tend to be wealthier and shop in nicer stores.

Most surprisingly, we find that the relationship between city size and prices turns negative (although again not significantly so) once we control for household income and city land prices. These results are not conclusive in part because we have not aggregated the data appropriately to build a price index, but they strongly suggest that the idea that prices rise with city size is not a robust feature of the data.[10]

We now turn to estimating the number of varieties available in a city based on our structural GNE approach. With 49 cities, the GNE approach involves the estimation of several hundred parameters, so we do not report all the values here. The AIC indicates that UPCs tend to

[10] One possible problem with these results is that we measure the size of cities by their population, but it could be the case that food production might not be highly correlated with population. We only have data on urban food manufacturing for 34 of our 49 cities, so shifting to food manufacturing as our measure of city size reduces the number of cities in our sample by 30 percent. For the cities with food manufacturing data, the correlation between food manufacturing and population is 0.6, with the three largest population centers New York, Los Angeles, and Chicago being the three largest producers of manufactured food products. This probably understates the true correlation because many food products manufactured by retailers (whose output Bernstein and Weinstein (2002) have shown are highly correlated with population) are probably not counted in food manufacturing output. Nevertheless, we obtain qualitatively similar result if we measure city size with food manufacturing.

fall within 10 incidence groups in terms of their purchase frequency.[11] Table 3 summarizes these estimates across our sample of 49 cities. We see that in all cities there are few UPCs that are purchased with very high frequency – on average, one in ten thousand UPCs are purchased with a frequency of 0.5 by a household. This would correspond to about 8 UPCs in the typical city having a purchase probability of 0.5 over the course of a year by a typical household. However, we also see that the vast majority of UPCs have extremely low purchase probability. 49 percent of UPCs have a purchase probability of approximately 1 in a thousand. Thus the product space is characterized by a few UPCs with high purchase probabilities and a vast number of UPCs that are rarely purchased.

Another way of summarizing the estimates is to examine how the probability that a UPC is purchased varies with city size. We would expect that the probability that any consumer purchases any one UPC would go down as the range of available UPCs in the city increases. Fortunately, this is easy to examine given that our GNE structure enables us to compute the probability that a household purchases a UPC by simply calculating

$$\Pi_c = \sum_{k=1}^{K} \alpha_{c,k} \pi_{c,k} .$$

In figure 2, we see that the average probability of purchase decreases sharply with city size. A UPC sold in our smallest city, Des Moines, has three times the probability of being purchased by any individual household as a UPC sold in New York. The fact that households in larger cities

[11] The incidence groups do not map directly into the product groups or product modules. The estimated purchase frequency associated with each UPC within a product group or brand-module will vary with the popularity of the UPC, which may be correlated with its brand, container, size, and other characteristics. The UPCs in the high frequency incidence group tend to be the most popular varieties of products that are frequently purchased, *e.g.* 12-packs of Coke cans, which are purchased by almost a third of the households in our sample. Less popular varieties of soda tend to fall into the lower purchase frequency incidence groups. These incidence groups will also be populated with the most popular UPCs in less frequently purchased product categories, *e.g.* Fleischmann's fresh cake yeast along with the more obscure varieties of soda as well as the less popular varieties of yeast.

are much less likely to buy any individual UPC is strongly suggestive of the fact that that the range of UPCs available in a city is increasing with city size.

We test this directly by using the GNE parameter estimates to calculate an estimate for total number of varieties in each city and considering how this varies with city size. Figure 3 plots how the log number of estimated varieties in each city varies with city size. It is interesting that the relationship between city size and the number of varieties in a city is similar but much stronger than the one we observed in figure 1. The data seem to strongly suggest a relationship between the size of a city and the number of varieties available as hypothesized by Krugman. Residents of New York have access just over 97,000 different varieties of groceries, while residents of small cities like Omaha and Des Moines have access to fewer than 32,000.

We test this relationship between city size and variety abundance formally in table 4. Table 4 presents the results from regressing the estimated number of varieties in a city from the GNE procedure on the log of the population in the city. The first three columns of the table present regressions of the sample counts of varieties in each city on the log of the city's population. The next columns present regressions of the estimate of number of varieties based on the GNE asymptotes on city size, and the final three columns repeat these results for our Weibull estimates. As one can see, the elasticity of variety with respect to population is less using the GNE estimates because these correct for the correlation between sample size and population in the AC Neilson data. What is most striking, however, is that we observe a very strong and statistically significant relationship between the size of the city and the number of estimated varieties. Our estimates indicate that a city with twice the population as another one typically has 20 percent more varieties.

One concern that one might have with these results is that they might be biased because larger cities might have more diverse populations. In order to control for this we constructed a number of Herfindahl indexes based on the shares of MSA population with different income, race, and country of birth. These indices will be rising in population homogeneity. In addition, we include the per capita income in each city. As one can see from Table 4, controlling for urban diversity does not alter the results.

Finally, we were concerned that our results might be due to a spurious correlation between city population and urban land area. If there are a constant number of unique varieties per unit area, then more populous cities might appear to have more diversity simply because they occupy more area. To make sure that this force was not driving our results, we include the log of urban land area in our regressions. The coefficient on land area is not significant when we use either the GNE or Weibull estimates of varieties and sometimes comes in negatively, while the coefficient on population remains positive and very significant. These results indicate that controlling for land area and demographic characteristics does not eliminate the strong relationship between city size and the number of available varieties. The R^2 of around 0.5 to 0.6 indicates that city size is an important determinant of variety availability. This is the first time anyone has documented that the number of tradable goods varies systematically with city size as hypothesized by New Economic Geography models.

Unfortunately, our structural estimation techniques cannot be applied to the estimation of the shares because although we have a theory of the probability of purchasing a good, we do not have a theory for probability of spending a particular share of income on a good. As we mentioned earlier, this forces us to use a non-structural technique. However, before doing so, we will first provide some intuition for the methodology and then demonstrate that non-structural

methods yield almost identical results as structural methods on count data and fit share data extremely well.

We begin by plotting accumulation curves, which are a graph of the number of distinct varieties we obtain on average in a random sample of *n* households against the number of households. The intuition for the approach arises from the fact that if households in a city choose more disparate sets of goods than households in a second city, the accumulation curve for the first city will lie above that of the second city because *n* households in the first city will tend to consume more varieties than *n* households in the second city. One can then obtain an estimate for the total number of goods available in each city based on the asymptote of the accumulation curve for that city.

We plot these accumulation curves for the twelve cities for which we have the largest samples in Figure 4. These curves reveal that the four highest curves – corresponding to the cities with the greatest variety of goods purchased – are for New York, DC-Baltimore, Philadelphia, and Boston. These cities are all among the five largest cities in our sample. This raw data plot is yet another indicator that residents of larger cities consume a broader set of varieties than those in smaller cities.

We can examine this more formally by estimating the asymptotes of the accumulation curves. Since we are not sure how to model these accumulation curves, we try five different functional forms – Clench, Chapman-Richards, Morgan-Mercer-Flodin, Negative Exponential, and Weibull. We choose among these based on the Akaike Information Criterion. The Weibull was a strong favorite with the lowest AIC score in the majority of cities for which we modeled UPC count accumulation curves, and so we decided to focus on this functional form.

Figure 5 plots the raw data and the estimated Weibull accumulation curve for our largest city, New York. A typical sample of 500 random households buys around 45,000 unique UPCs, and a sample of 1000 households typically purchases around 65,000 different goods. Since there is some overlap between the consumption baskets of different households in the sample, the number of unique UPCs that are added to the hypothetical sample is decreasing with the size of the hypothetical sample which produces a monotonically rising, but concave UPC accumulation curve. As one can see from the plot, the estimated Weibull distribution fits the data extremely well. The estimated asymptote is 112,000 varieties, which is 35,000 more than we observe in our sample of 1500 New York households.

Figure 6 presents a plot of the log population in a city against the log of the estimated Weibull asymptote. As one can see, there is a clear positive relationship between the two variables: the accumulation curves imply that larger cities have more varieties than smaller ones. Despite the fact that the Weibull indicates more varieties for New York than the GNE, overall Weibull asymptotes tend to be around 20 percent lower than our GNE structural estimates. Visual inspection reveals that the relative relationship between the estimated number of varieties and city size using the GNE and plotted in figure 3 is almost identical to that obtained with the Weibull. We can see this even more clearly in figure 7 where we plot the Weibull asymptotes against the GNE asymptotes. The correlation between the two is 0.99 and the slope is 1.07 indicating that both methodologies yield essentially the same relationship between city size and the number of available varieties.

Just as we constructed accumulation curves for the number of different UPCs in cities in the previous section, we also construct these for the share of common brands in each product group, s_{gc}, and the average share of common UPCs in each brand in each product group, $\overline{s}_{gc}$. We

then estimate the asymptotes of the fitted Weibull curve to each of these. Figure 8 plots these estimates for one of the largest product groups in terms of sales and the number of varieties available nationally, bread and baked goods, for two cities with large samples in the AC Nielson data but very different populations, New York and Little Rock. Panel A shows the s_{gc} accumulation curve while panel B portrays the same curve for $\overline{s}_{gc}$. As one can see from the plots, the Weibull fits the share accumulation curves extremely well.[12] By examining the asymptote of the Weibull distribution in the first panel, we can see that New Yorkers have access to a set of bread brands available that constitute 79 percent of national expenditures on bread. By contrast, residents of Little Rock, with a population less than a tenth as large, have access to bread brands that constitute 74 percent of national expenditures. Panel B shows that the within-brand-module market share of UPCs available in a city, conditional on that brand-module being available in the city, is, on average, lower than the within-product group market share of brand-modules available in the city. Firms sell UPCs accounting for 53 and 47 percent of their national sales in New York and Little Rock, respectively. This indicates that knowing that a firm sells a product in a city does not necessarily mean that all varieties of that product are available there. In summary, New Yorkers have access to 5 percent more of the bread market, in terms of brands, than residents of Little Rock and, conditional on a brand being available in their city, New Yorkers also have access to 6 percent more of the market for each brand, in terms of UPCs, than do Little Rock residents.

To demonstrate how these differences in the market shares of UPCs and brands available in the two cities translate into a discount on the exact price index in New York relative to Little Rock, we can calculate the variety adjustments for bread in each city from equation 11. The

[12] The AIC favors the Weibull distribution when computing share accumulation curves as well.

elasticity of substitution between UPCs within brand-modules in the bread product group is 17.2, so the UPC variety adjustment for New York is $0.53^{1/(1-17.2)}$, or 1.04. The across-brand elasticity of substitution for the bread product group is 9.6, so the brand variety adjustment for New York is $0.79^{1/(1-9.6)}$, or 1.03. The fact that there are fewer bread varieties available in New York than nationally means that someone restricted to only consume those varieties available in New York would face a price index that is 7 percent higher than someone with access to all national varieties. A similar calculation shows that the variety adjustment for the bread product group in Little Rock is equal to 9 percent. The variety adjustment for New York relative to Little Rock is equal to 1.07/1.09, or 0.98, implying that people living in New York face 2 percent lower costs for bread because they have access to more varieties.

Figures 9A and 9B plot the average asymptotes of the share accumulation curves (estimated using a Weibull distribution) in each city against the log of the population in each city to show that the results for bread in New York and Little Rock are representative of the sample as a whole. As one can see from the figure, there is a strong positive relationship between our estimates for the national market share covered by varieties available in each city and the city's population. Nationally, consumers spend 5 percent more on brands and UPCs available in the largest cities than they do on those brands and UPCs available in the smallest cities. Once again we see that people in larger cities have access to the more and/or more popular varieties, while those in smaller cities have more limited access.

The asymptotes, in conjunction with our price data, enable us to estimate the exact price index for each product group. Table 5 presents our estimates for how the conventional exact price index, the variety adjustment, and the exact price index vary across cities. The first three columns in the table use unadjusted prices to compute the index. As we saw in table 2,

unadjusted prices are higher in larger cities, and these results translate into a common exact price index that rises with city size. These higher prices, however, are offset by the variety adjustment to prices arising from the greater availability of varieties in larger cities leading to an exact price index that is invariant to city size.

The unadjusted prices, however, are problematic because they do not correct for the store in which the goods are purchased or the type of household making the purchase. We therefore use adjusted prices in the subsequent columns. This yields two striking results. First, when we adjust for household characteristics and store quality in column 7, on average, the prices of common goods in larger cities are lower than prices in smaller cities. In the basic specification, this result is only significant at the 10 percent level. The main reason for this is that the prices of common goods also incorporate some of the higher land costs. If we purge prices of the impact of land costs, we see in column 13 that the CEPI falls more sharply with city size and this result is significant at the 5 percent level.

More important one for the theory, however, is that we observe that the exact price index falls with city size, and this relationship is significant whenever we use adjusted prices. About two-thirds of this effect appears to be due to the fact that individual goods prices are lower in larger cities, while the remainder is due to there being more, and more important, varieties available in larger cities. The economic significance of this result is fairly substantial. Variety adjusted prices in New York are 9 percent lower than those in the smallest city in our sample (Des Moines). This suggests a potentially very important role for new economic geography affecting the cost of living in cities.

V. Conclusion

Krugman (1991) hypothesized that market prices drive agglomeration and the new economic geography literature which has followed rests on the same foundations. Previous empirical work using the CPI or housing costs as a cost of living measure has found that the cost-of-living is, in fact, a dis-amenity of large cities and, as such, a force of dispersion in a model of economic geography. The results presented here suggest that Krugman's result, as well as both the price and variety mechanisms underlying it, are correct. Specifically, these results support the extensions to Krugman's model in which the two primary components of the cost of living index – housing costs and product prices - move in opposing directions as agglomeration occurs.

The analysis presented here has shown that the variety-adjusted cost of grocery products decreases with city size and preliminary robustness checks using one quarter of all purchase data for a sample of 300 households in 10 cities suggest that this pattern extends throughout the traded consumer goods sector. Further work on a larger set of household purchases could confirm that the patterns identified here are not specific to grocery products.

References

Anderson, S., De Palma, A., and Thisse, J-F. (1987) "The CES is a Discrete Choice Model?" *Economic Letters*, Vol. 24, No. 2, pp. 139-140.

Asahi, C., Hikino, S., and Kanemoto, Y. (2008) "Consumption Side Agglomeration Economies in Japanese Cities," CIRJE Discussion Papers, No. CIRJE-F-561, April.

Bernstein, J. and Weinstein, D. (2002) "Do Endowments Predict the Location of Production? Evidence from National and International Data," *Journal of International Economics*, Vol. 56, pp. 55-76.

Bils, M. and Klenow, P. (2001) "Quantifying Quality Growth," *American Economic Review*, Vol. 91, No. 4, pp. 1006-1030.

Brakman, S., Garretsen, H., and van Marrewijk, C. (2009) *The New Introduction to Geographical Economics*. Cambridge: Cambridge University Press.

Broda, C. and Weinstein, D. (2008) *Prices, Poverty, and Inequality: Why Americans are Better Off than You Think*, Washington, DC: The AEI Press.

Broda, C. and Weinstein, D. (2010) "Product Creation and Destruction: Evidence and Price Implications," *The American Economic Review* , June.

Broda, C., Leibtag, E., and Weinstein, D. (2009) "The Role of Prices in Measuring the Poor's Living Standards," *Journal of Economic Perspectives*, Vol. 23, No. 2, Spring, pp. 77-97.

Colwell, R. and Coddington, J. (1994) "Estimating Terrestrial Biodiversity through Extrapolation," *Philosophical Transactions: Biological Sciences*, Vol. 245, No. 1311, July, pp. 101-118.

Colwell, R., Mao, C., and Chang, J. (2004) "Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves," *Ecology*, Vol. 85, No. 10, pp. 2717-2727.

Combes, P., Mayer, T., and Thisse, J-F. (2008) *Economic Geography: the Integration of Regions and Nations*, Princeton: Princeton University Press.

Davis, M. and Palumbo, M. (2007) "The Price of Residential Land in Large US Cities," *Journal of Urban Economics*, Vol. 63, No. 1, pp. 352-384.

DuMond, J.M., Hirsch, B., and Macpherson, D. (1999) "Wage Differentials Across Labor Markets and Workers: Does Cost of Living Matter?" *Economic Inquiry*, Vol. 37, No. 4, October, pp. 577-598.

Dunlevy, J. (2006) “The Influence of Corruption and Language on the Pro-Trade Effect of Immigrants: Evidence from the American States.” *The Review of Economics and Statistics*, Vol. 88, No. 1, February, pp. 182-186.

Einav, L., Liebtag, E., and Nevo, A. (2008) “On the Accuracy of Nielsen Homescan Data,” USDA Economics Research Report Number 69.

Feenstra, R. (1994) “New Product Varieties and the Measurement of International Prices,” *American Economic Review*, Vol. 84, No. 1, March, pp. 157-177.

Feenstra, R. [forthcoming] *Product Variety and the Gains from International Trade.* MIT Press.

Fujita, M. and Mori., T (2005) "Frontiers of the New Economic Geography", *Papers in Regional Science* 84(3), 377-405.

Gould, D. (1994) “Immigrant Links to the Home Country: Empirical Implications for U.S. Bilateral Trade Flows,” *The Review of Economics and Statistics*, Vol. 76, No. 2, May, pp. 302-316.

Helpman, E. (1998) “The Size of Regions,” in Pines, D., Sadka, E., and Zilcha, I., eds., *Topics in Public Economics*, Cambridge: Cambridge University Press, pp. 33-54.

Jimenez-Valverde, A., Mendoza, S., Martin, J., Cano, J., and Munguira, M. (2006) “Comparing Relative Model Fit of Several Species-Accumulation Functions to Local Papilionoidea and Hesperioidea Butterfly Inventories of Mediterranean Habitats,” *Biodiversity and Conservation*, Vol. 15, pp. 163-176.

Krugman, P. (1991) “Increasing Returns and Economic Geography,” *The Journal of Political Economy*, Vo. 99, No. 3, June, pp. 483-499.

Mao, C., Colwell, R., and Chang, J. (2005) “Estimating the Species Accumulation Curve Using Mixtures,” *Biometrics*, Vol. 61, June, pp. 433-441.

Prize Committee of the Royal Swedish Academy of Sciences (2008) “Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2008,” http://nobelprize.org/nobel_prizes/economics/laureates/2008/ecoadv08.pdf

Ottaviano, G., Tabuchi, T., and Thisse, J. (2002) “Agglomeration and Trade Revisited,” *International Economic Review*, Vol. 43, No. 2, May, pp. 409-435.

Roback, J. (1982) “Wages, Rents, and the Quality of Life,” *The Journal of Political Economy*, Vol. 90, No. 6, December, pp. 1257-1278.

Roback, J. (1988) “Wages, Rents, and Amenities: Differences Among Workers and Regions,”

Economic Inquiry, Vol. 26, No. 1, January, pp. 23-41.

Sato, K. (1976) "The Ideal Log-Change Index Number," *The Review of Economics and Statistics,* Vol. 58, No. 2, pp. 223–228.

Suedekum, J. (2006) "Agglomeration and Regional Costs of Living," *Journal of Regional Science*, Vol. 46, No. 3, pp. 529-543.

Tabuchi, T. (2001) "On Interregional Price Differentials," *The Japanese Economic Review*, Vol. 52, No. 1, March, pp. 104-115.

Tabuchi, T. and Yoshida, A. (2000) "Separating Urban Agglomeration Economies in Consumption and Production," *Journal of Urban Economics*, Volume 48, pp. 70-84.

Vartia, Y (1976) "Ideal Log-Change Index Numbers," *Scandinavian Journal of Statistics,* Vol. 3, No. 3, pp. 121–126.

Figure 1

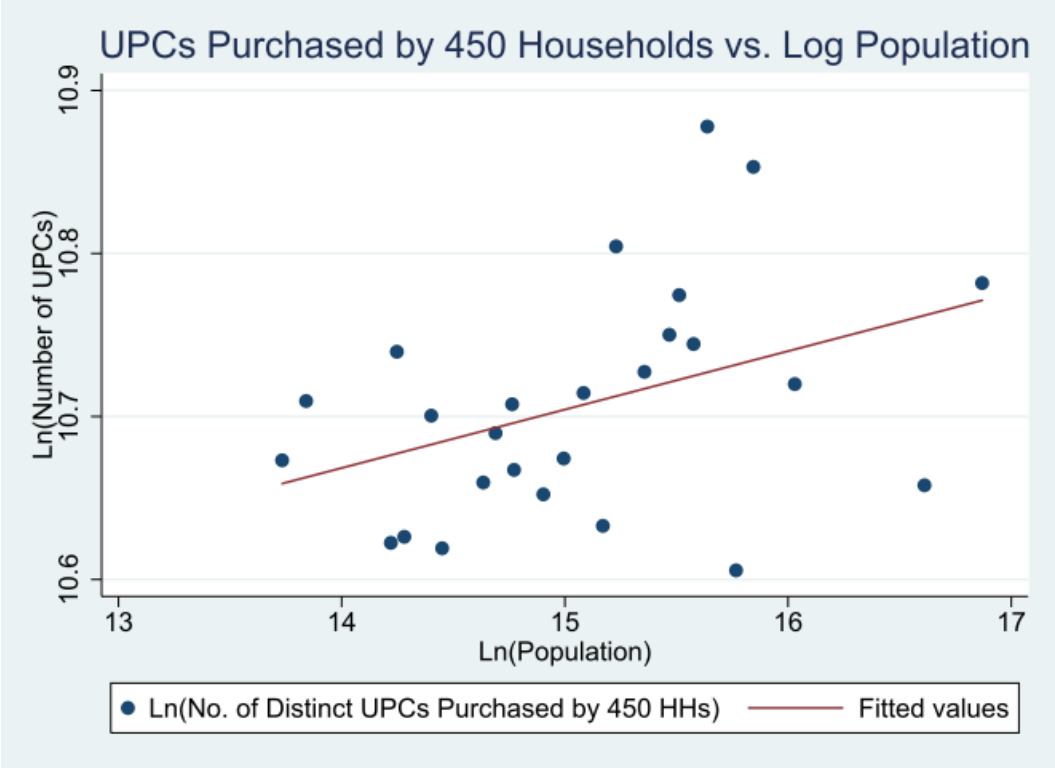
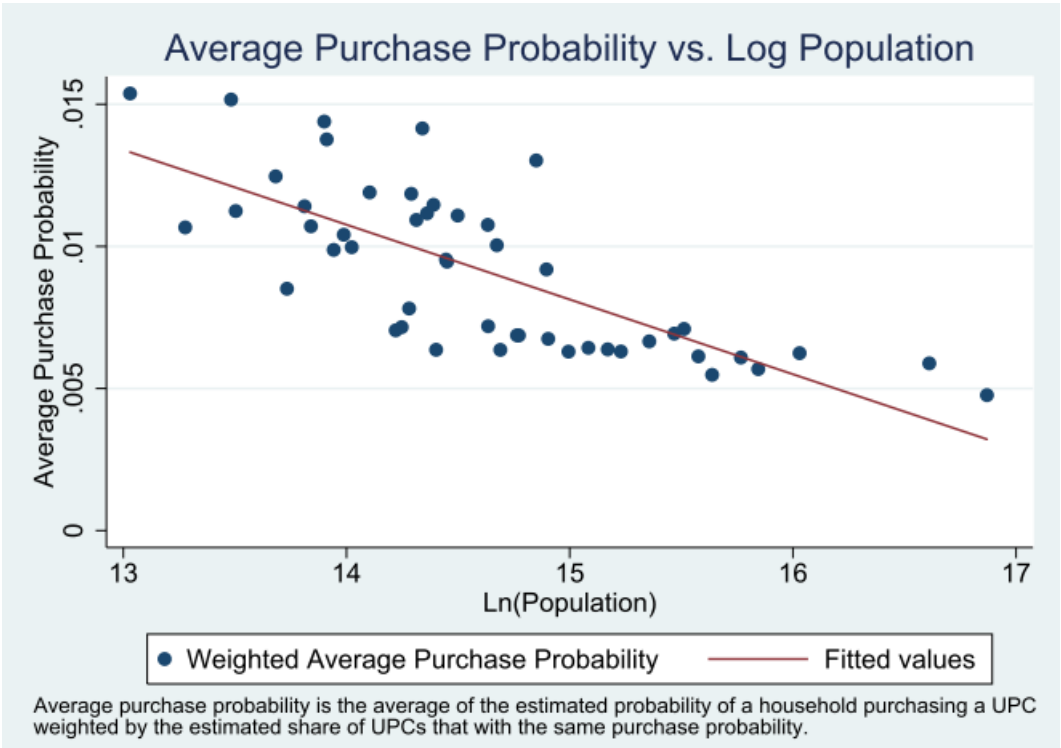


Figure 2



Average purchase probability is the average of the estimated probability of a household purchasing a UPC weighted by the estimated share of UPCs that with the same purchase probability.

Figure 3

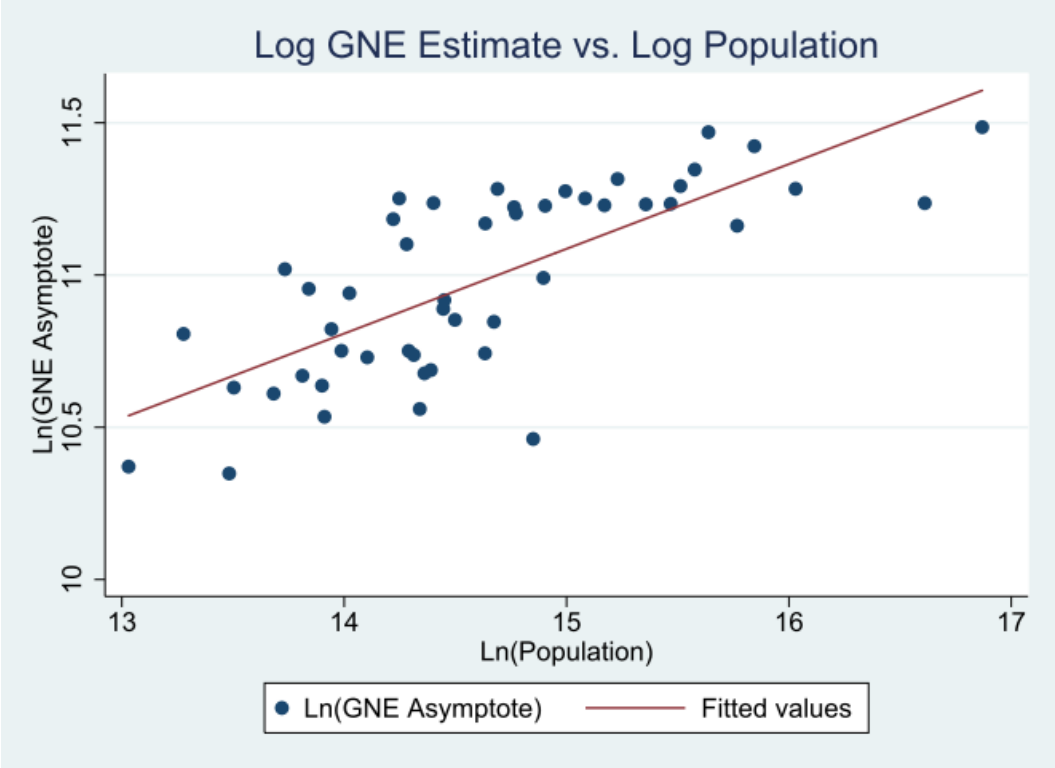


Figure 4

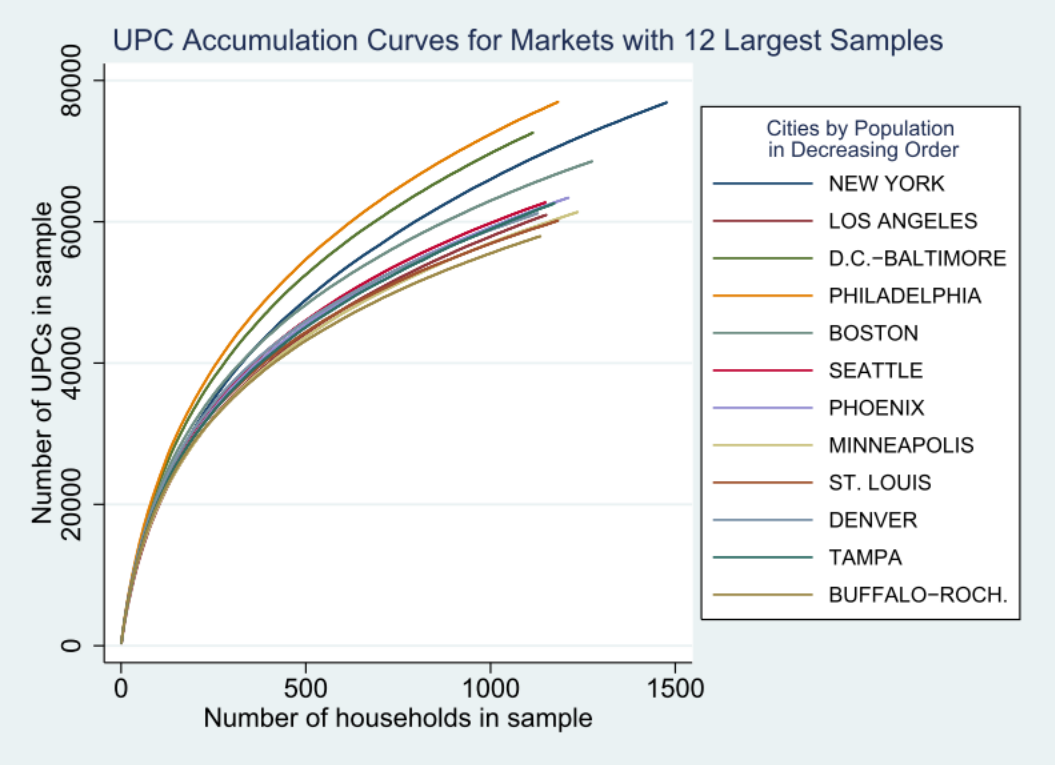


Figure 5

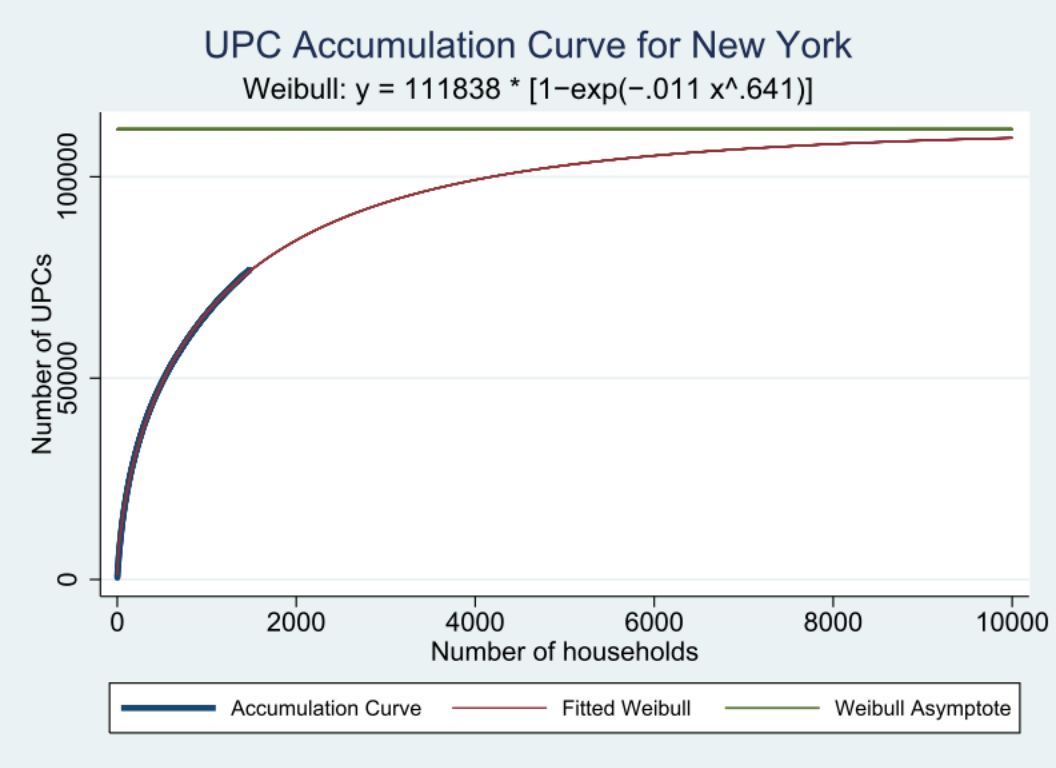


Figure 6

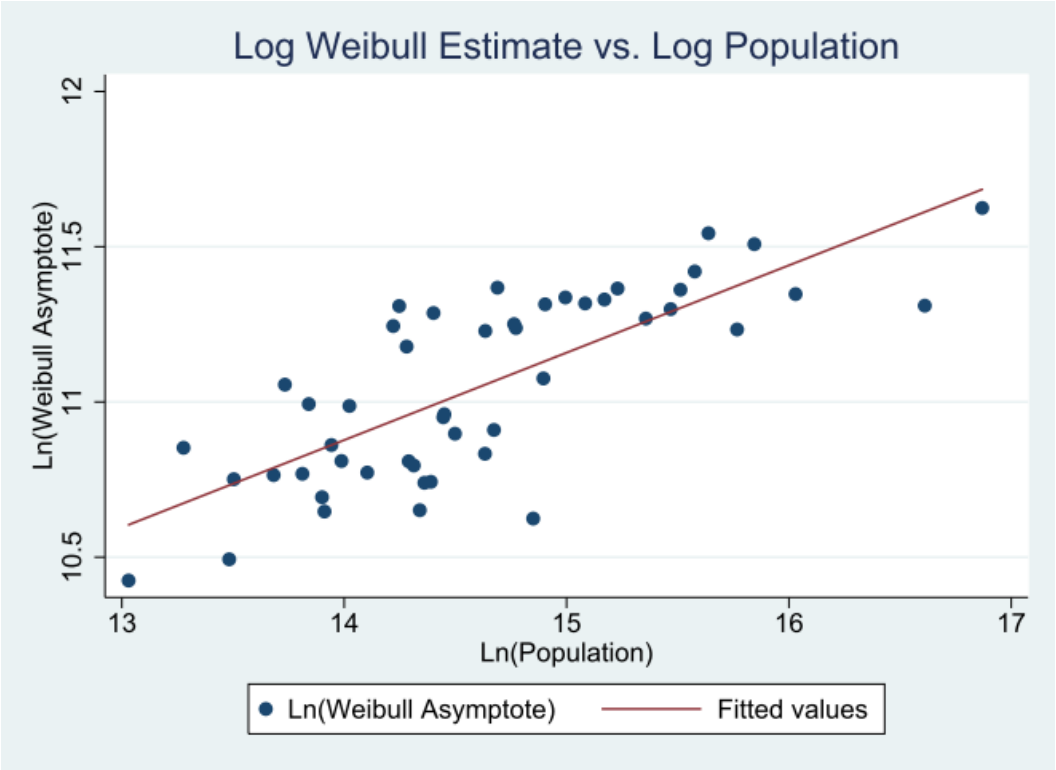


Figure 7

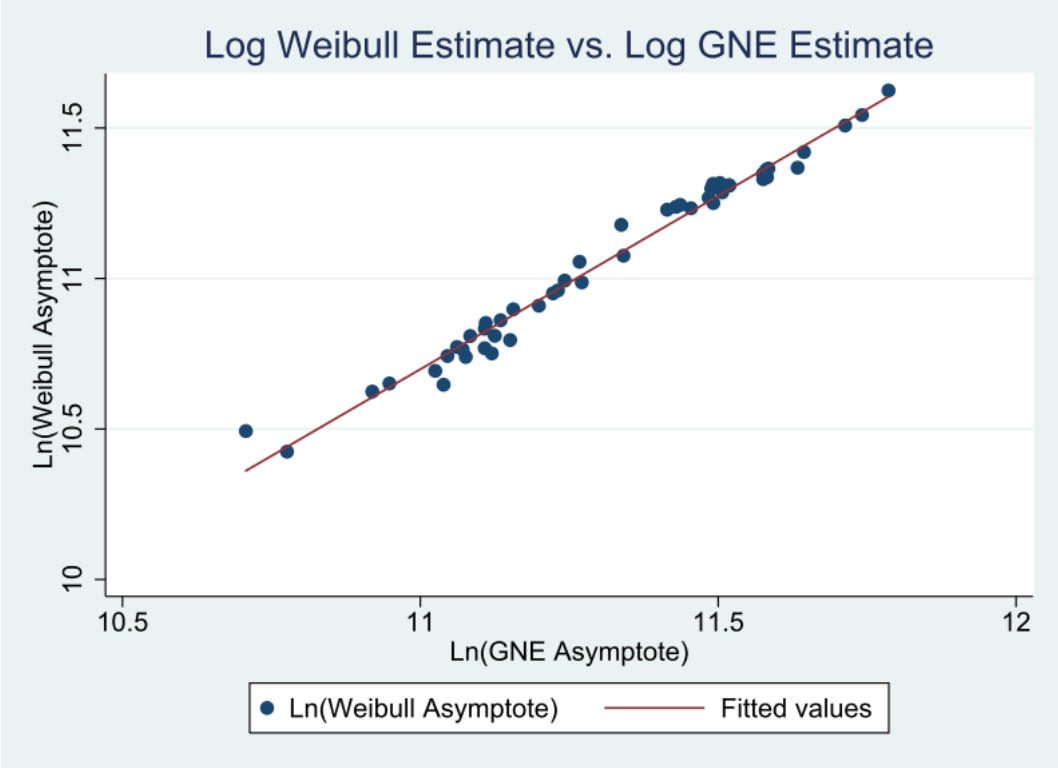

Log Weibull Estimate vs. Log GNE Estimate
Ln(Weibull Asymptote)
10
10.5
11
11.5
10.5
11
11.5
12
Ln(GNE Asymptote)
Ln(Weibull Asymptote)
Fitted values

Figure 8A

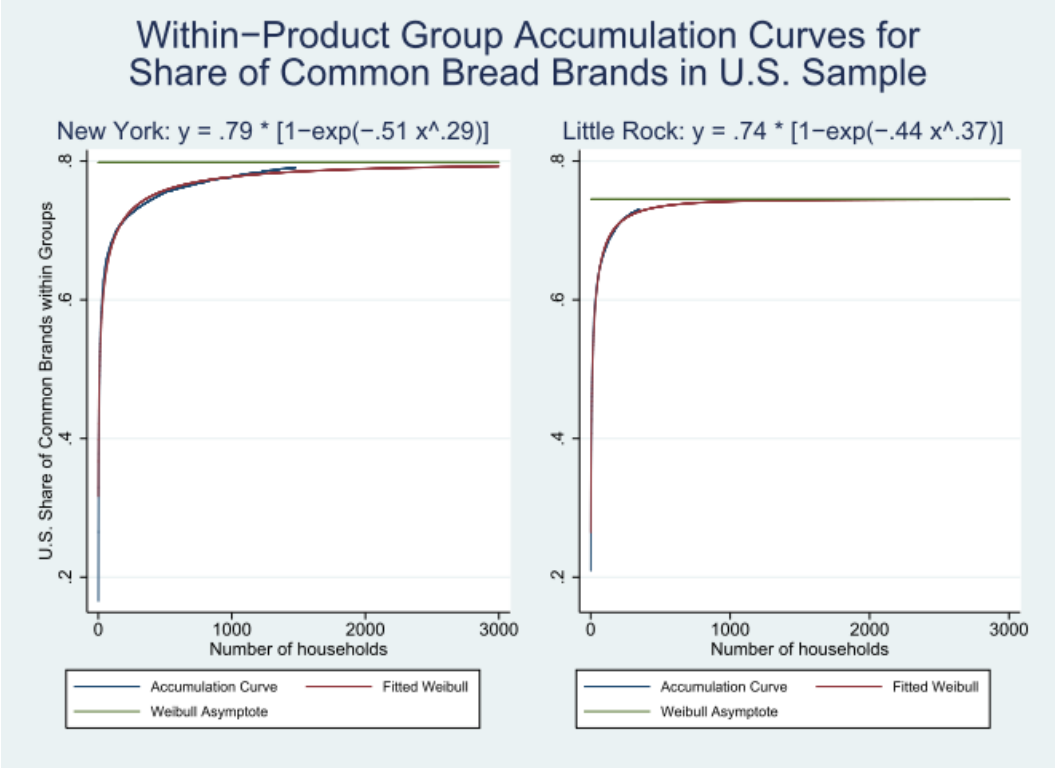


Figure 8B

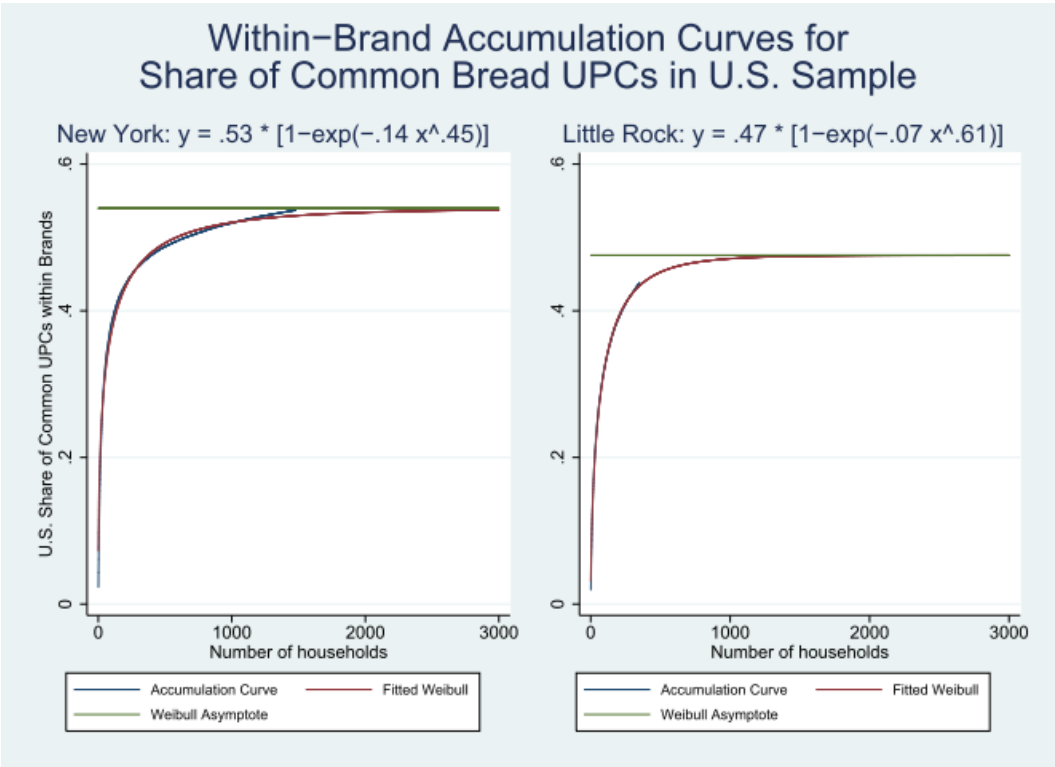
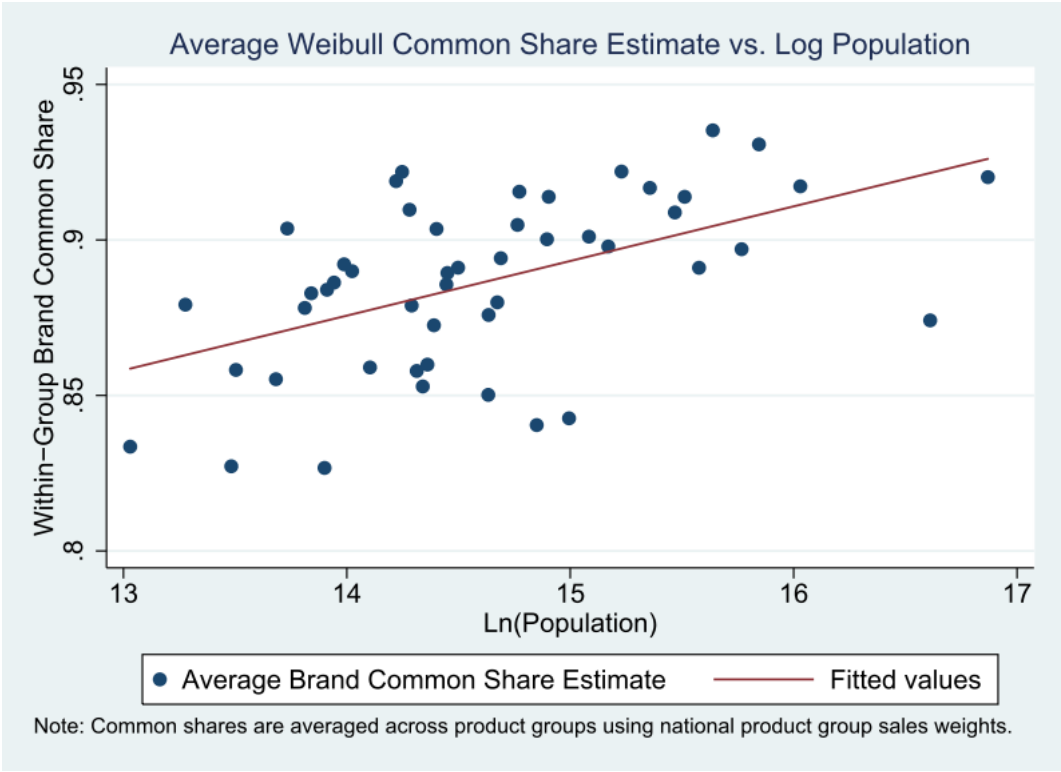
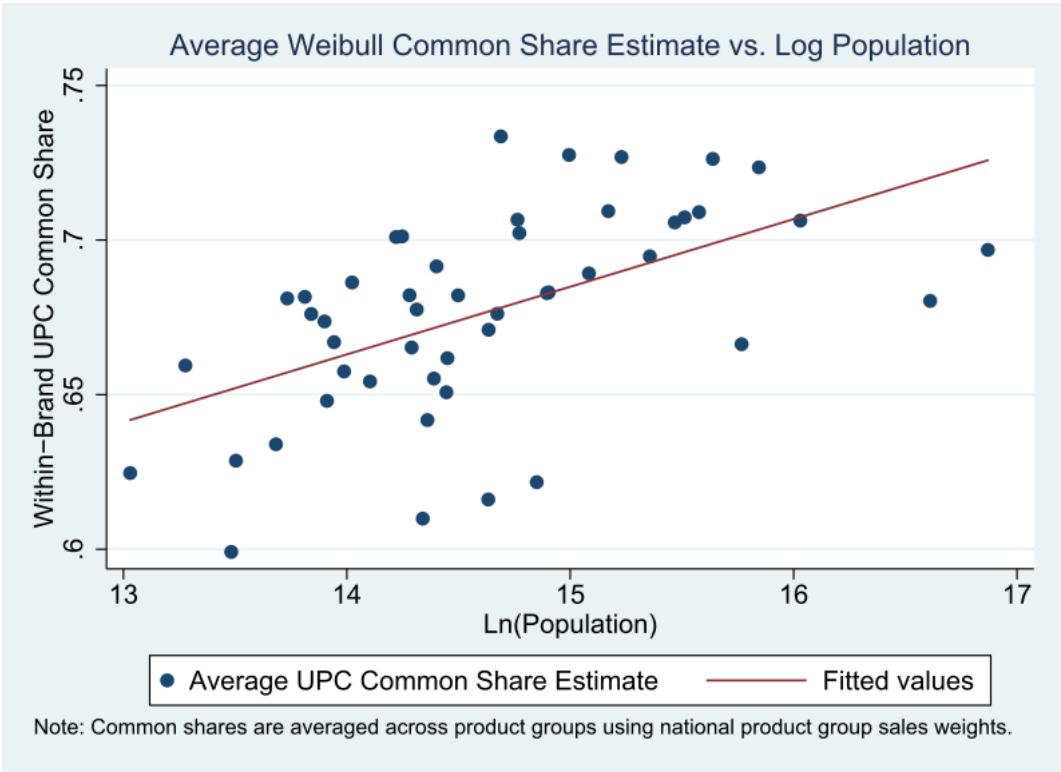


Figure 9A



Note: Common shares are averaged across product groups using national product group sales weights.

Figure 9B



Note: Common shares are averaged across product groups using national product group sales weights.

Table 1

Summary Statistics for Number of Purchases

	All Households		Households with: One Person		Two Person		Three Person	
Number of Households	18570		4594		8141		2627	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Number of Unique UPCs Purchased, Conditional on Purchasing One								
Total	364	341	243	228	354	339	430	414
Within Each Product Group	8	5	6	4	7	5	9	5
Within Each Product Module	3	2	2	1	3	2	3	2
Brand-Modules Purchased Conditional on Purchasing One Good in a Module								
Total	250	240	173	165	247	239	291	284
Within Each Product Group	5	3	4	3	5	3	6	4
Within Each Product Module	2	1	2	1	2	1	2	1
Number of Units Purchased of Each:								
UPC	4	1	4	1	4	1	4	1
Brand-Module	6	2	5	2	5	2	6	2
Probability of Purchasing Only One Good Per Module, Conditional on Purchasing One Good in a Module								
UPC	0.47		0.51		0.48		0.45	
Brand-Module	0.60		0.63		0.60		0.58	

Notes:
[1] Including only 250 consecutive days for each household, where consecutive is defined as any period without gaps between household purchases lasting 14 days or more. Any household without 250 days of consecutive 2005 purchase data is not considered in this table.

Table 2

	Ln($Price_{h,c,u}$)				
Ln($Population_c$)	0.00811	0.00439	0.00422	0.00329	-0.00283
	[0.00733]	[0.00408]	[0.00401]	[0.00395]	[0.00441]
Ln($Income_h$)	-	-	-	0.0104***	0.0101***
	-	-	-	[0.000969]	[0.000947]
Ln(Land $Value_c$)	-	-	-	-	0.0135***
	-	-	-	-	[0.00359]
Storename Dummies[2]	No	Yes	Yes	Yes	Yes
Household Demographic Dummies[3]	No	No	Yes	Yes	Yes
Observations	12,700,000	12,700,000	12,700,000	12,700,000	12,700,000
R-squared	0.001	0.090	0.091	0.092	0.093

Robust standard errors in brackets. Standard errors are clustered by city.
*** p<0.01, ** p<0.05, * p<0.1

Notes:

[1] $P_{i,c,h}$ = average price paid for UPC *i* by household *h* in city *c*.

[2] Regressions with storename dummies include a store type dummy instead of a storename dummy if the store name is missing or sample store sales are under $100,000.

[3] Household controls include dummies for household size, male and female head of household age, marital status, race, and hispanic.

[4] Random weight goods have been dropped from the sample.

[5] All regressions have UPC fixed effects and observations are weighted by total sample UPC sales.

Table 3
Summary Statistics for GNE Parameter Estimates

	Probability of Purchase ($\pi_{c,k}$)		Share of UPCs ($\alpha_{c,k}$)	
Incidence Group (k)	Mean	Standard Deviation	Mean	Standard Deviation
1	0.496	0.096	0.00009	0.00006
2	0.332	0.103	0.00041	0.00029
3	0.226	0.085	0.00116	0.00071
4	0.152	0.066	0.00300	0.00181
5	0.102	0.052	0.00757	0.00470
6	0.065	0.035	0.01953	0.01045
7	0.038	0.022	0.05083	0.01821
8	0.019	0.011	0.12115	0.01879
9	0.008	0.004	0.30803	0.03277
10	0.001	0.001	0.48823	0.03465

Table 4

Do larger cities have more UPC varieties?

	Sample Count		ln(GNE Asymptote)			ln(Weibull Asymptote)			
ln(Population)	0.312***	0.338***	0.281***	0.278***	0.300***	0.261***	0.281***	0.303***	0.275***
	[0.0432]	[0.0678]	[0.0971]	[0.0364]	[0.0572]	[0.0821]	[0.0340]	[0.0534]	[0.0769]
ln(Per Capita Income)	-	-0.155	-0.043	-	-0.137	-0.060	-	-0.129	-0.074
	-	[0.341]	[0.369]	-	[0.288]	[0.312]	-	[0.269]	[0.292]
Income Herfindahl Index	-	-0.952	-0.289	-	-0.630	-0.178	-	-0.321	0.000
	-	[3.132]	[3.246]	-	[2.641]	[2.745]	-	[2.466]	[2.568]
Race Herfindahl Index	-	0.064	0.115	-	0.074	0.109	-	0.111	0.135
	-	[0.411]	[0.417]	-	[0.347]	[0.353]	-	[0.324]	[0.330]
Birthplace Herfindahl Index	-	0.006	0.029	-	-0.012	0.004	-	-0.028	-0.017
	-	[0.282]	[0.285]	-	[0.238]	[0.241]	-	[0.222]	[0.225]
ln(Land Area)	-	-	0.087	-	-	0.060	-	-	0.042
	-	-	[0.106]	-	-	[0.0898]	-	-	[0.0840]
Constant	6.158***	7.474**	6.275*	6.912***	8.057***	7.240**	6.937***	7.962***	7.381**
	[0.632]	[3.391]	[3.704]	[0.533]	[2.860]	[3.132]	[0.498]	[2.670]	[2.930]
Observations	49	49	49	49	49	49	49	49	49
R-squared	0.53	0.53	0.54	0.55	0.56	0.56	0.59	0.60	0.60

*** p<0.01, ** p<0.05, * p<0.1

Standard errors in brackets.

Table 5

Are price indexes higher in larger cities? What if we adjust for varieties?

	Unadjusted Prices			*Demographic-Adjusted*			*Demographic-Store Adjusted*			*Demographic-Store-Income Adjusted*			*Fully-Adjusted*		
Household Effects	Included			Removed			Removed			Removed			Removed		
Store Effects	Included			Included			Removed			Removed			Removed		
Zip Income Effects	Included			Included			Included			Removed			Removed		
Land Value Effects	Included			Included			Included			Included			Removed		
	$CEPI_c$ [1]	VA_c [2]	EPI_c [3]	$CEPI_c$ [4]	VA_c [5]	EPI_c [6]	$CEPI_c$ [7]	VA_c [8]	EPI_c [9]	$CEPI_c$ [10]	VA_c [11]	EPI_c [12]	$CEPI_c$ [13]	VA_c [14]	EPI_c [15]
Log Population	0.0120**	-0.00823**	0.00488	-0.0134	-0.00528**	-0.0200*	-0.0159*	-0.00519**	-0.0227**	-0.0167*	-0.00517**	-0.0235**	-0.0213**	-0.00508**	-0.0283***
	[0.0051]	[0.0038]	[0.0074]	[0.0095]	[0.0025]	[0.011]	[0.0085]	[0.0025]	[0.0096]	[0.0085]	[0.0025]	[0.0096]	[0.0083]	[0.0025]	[0.0093]
Constant	0.825***	1.214***	1.023***	1.317***	1.143***	1.488***	1.355***	1.141***	1.527***	1.367***	1.141***	1.540***	1.436***	1.140***	1.612***
	[0.076]	[0.056]	[0.11]	[0.14]	[0.037]	[0.16]	[0.13]	[0.037]	[0.14]	[0.13]	[0.037]	[0.14]	[0.12]	[0.037]	[0.14]
Observations	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37
R-squared	0.138	0.12	0.012	0.053	0.117	0.089	0.091	0.113	0.137	0.1	0.112	0.146	0.158	0.108	0.208

*** p<0.01, ** p<0.05, * p<0.1

Standard errors in brackets.

Notes:

[1] The dependent variables in the above regressions are indices. These indices are calculated using unadjusted prices or prices that have been adjusted as indicated above.

[2] Random weight goods have been dropped from the sample.

[3] $EPI_c = CEPI_c VA_c$ which implies that $\log(EPI_c) = \log(CEPI_c) + \log(VA_c)$. Note that the dependent variables in the above regressions are in levels, not logs, so the coefficients on log population in the $CEPI_c$ and VA_c regressions do not add to the coefficient on log population in the EPI_c regression.