THE MONEY PUMP AS A MEASURE OF REVEALED PREFERENCE VIOLATIONS

FEDERICO ECHENIQUE, SANGMOK LEE, AND MATTHEW SHUM

Abstract. We introduce a measure of the severity of violations of the revealed preference axioms, the *money pump index (MPI)*. The MPI is the amount of money one can extract from a consumer who violates the axioms. It is also a statistical test for the hypothesis that a consumer is rational, when behavior is observed with error.

We present an application using a panel data set of food expenditures. The data exhibits many violations of the axioms. Mostly, the MPI for these violations is small. The MPI indicates that the hypothesis of consumer rationality cannot be rejected.

1. Introduction

The assumption that consumers are rational is one of the oldest and most controversial assumptions in economics. Conceptually, the empirical content of the rationality assumption has been very well understood since the works of Samuelson (1938), Afriat (1967), Richter (1966) and Varian (1982): revealed preference theory captures the empirical content of rational consumption behavior.

As a practical matter, however, revealed preference analysis is problematic due to the "all or nothing" nature of the exercise: a data set either satisfies the generalized axiom of revealed preference (GARP) or it does not. In practice, however, it is useful to gauge how *severely* consumers violate the axiom. Our paper presents a new measure of the severity of a violation of GARP. The measure is based on the idea that a consumer who violates GARP is subject to being exploited as a "money pump." We propose that the severity of a violation be measured by the amount of money that could be extracted from the consumer; we call this the *money pump index (MPI)*.

Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena CA 91125. Emails: {fede,sangmok,mshum}@hss.caltech.edu. We thank Dan Silverman for comments, and the editor and two anonymous referees for many useful suggestions on previous drafts.

INSERT FIGURE 1

The MPI implements a statistical test for the hypothesis of consumer rationality. A violation of GARP may be the result of measurement errors in prices, or in consumer choices. The MPI measures how likely it is that a violation of GARP is due to such errors, and essentially compatible with rational behavior.

We present an empirical application to household-level "scanner" panel data containing time-series of household-level food grocery purchases collected at checkout scanners in supermarkets. In contrast to many earlier studies, using more aggregate, or cross-sectional consumption data, our analysis revealed a substantial number of violations of GARP. Specifically, 396 out of the 494 households in our data set violate GARP at some point. However, most of these violations are not severe: our MPI is centered around 6% of a household's food expenditures, or about $12.80 when evaluated at the average monthly food expenditure of $213.

The magnitudes of the MPI appear intuitively very small (in both dollar and percentage terms). We proceed to formally use the MPI to test the hypothesis of consumer rationality. We test whether the MPI could be accounted for simply by measurement errors in the variables. In our empirical application, we are unable to reject the null hypothesis that the observed MPIs are consistent with rational behavior and measurement errors. In other words, the apparently small 6% MPI is also small in a statistical sense.

We correlate our measure with demographic variables. Most results are intuitive: Less educated, poorer, and older households make more severe violations of GARP than do highly educated, richer, and younger households. On the other hand, smaller households make more severe violations of GARP. Moreover, because the demand for many food grocery items reflect seasonal trends, we also check whether GARP violations are more severe when comparing consumption between peak vs. non-peak seasonal periods. We find no evidence of this, implying that consumption in our data can be modeled by stable preferences which exhibit no seasonal component.

Money pump. Our measure of the severity of a GARP violation is motivated by the idea that a violation of GARP exposes a consumer to being manipulated as a "money pump." For example, consider the situation in Figure 1(a). A consumer buys bundle x at prices p and x' at prices p'. Evidently, there is a violation of GARP (actually of WARP, the weak axiom of revealed preference) because x was purchased

INSERT FIGURE 2

when x' was affordable, and vice versa. Knowing these choices, a devious "arbitrager" who follows the opposite purchasing strategy (buying bundle x at prices p', and bundle x' at prices p), could profitably resell x to the consumer at prices p, and x' at prices p'. The total profit the arbitrager would make equals

$$mp = p \cdot (x - x') + p' \cdot (x' - x).$$

We use the magnitude mp, "money pump cost," to measure the severity of the violation of GARP. Specifically, our MPI is the money pump cost expressed as a percentage of expenditure.

The MPI is an intuitive measure of the severity of a violation of GARP. Consider the situations in Figures 1(a) and 1(b). Each figure presents a violation of GARP, but intuitively the violation in 1(b) is more severe than the one in 1(a). The money pump cost reflects this difference. Figures 2(a) and 2(b) represent the money pump cost: it is the sum of the translation of the p-budget line (from crossing x to crossing x'), and the translation of the p'-budget line (from crossing x' to crossing x). The money pump represents the severity of the violations, and it is expressed in monetary terms, so the numerical value of a violation has a clear interpretation. As we explain below, it also functions as a formal test statistic.

The idea that arbitragers can "pump money" from irrational consumers is not new, and it has been used as a reason for why one should not observe irrational behavior. For our purposes, however, the devious arbitrager is a fictional character. There is a debate on whether GARP violators would be driven out of the market because of the actions of arbitragers who exploit them as money pumps (see, for example, Mulligan (1996), Rabin (2002) and Laibson and Yariv (2007)). We do not take a stand on the issue: our use of the money pump is purely an application of the idea captured by Figure 1.

Panel data. Many of the recent studies of GARP employ repeated cross-sectional data; tests of GARP implemented on such data require some sort of aggregation or "matching" of similar households across different cross-sections. In contrast, our panel data allows us to study how household purchases vary with prices over time, without the need to aggregate or "match" consumers. By focusing on supermarket purchases, we also observe a higher frequency of price changes relative to expenditure, compared to standard cross-sectional consumption data sets. As is well-known (see e.g. Blundell, Browning, and Crawford (2003)), a large variability in expenditure relative to prices can result in GARP having low power.

2. Related literature

The literature on testing the revealed preference axioms is large, and contains both classical papers as well as more recent contributions. Afriat (1967) and Varian (1982) are seminal contributions to the methodology of revealed preference tests; Varian (2006) provides a survey. Empirical applications of revealed preference tests have employed both field as well as experimental data.

In principle, tests of WARP/GARP require repeated observations of a decision-making unit (individual or household) across different pricing regimes. However, many of the empirical investigations of GARP using field data employ data from cross-sectional household-level surveys (such as the Consumer Expenditure Survey in the US, and the Family Expenditure Survey in the UK). Thus, an important challenge addressed in these papers is how to "match" households across different time periods to form a synthetic panel. Blundell, Browning, and Crawford (2003) and Blundell, Chen, and Kristensen (2007) address this issue by estimating an "Engel curve" relating a household's consumption to prices, expenditure and household demographics, and test GARP by comparing the predicted consumption behavior of households with similar demographics and expenditure levels across different pricing regimes. Hoderlein and Stoye (2009) take a more agnostic approach, and use results from the copula literature to obtain bounds on the percentage of households which violate WARP in two separate cross-sections of survey data.

In the present paper, we avoid these difficulties by using a long household-level scanner panel dataset, where the purchase decisions of given households over a two-year period are observed. To our knowledge, testing the revealed preference axioms using scanner data is new in the literature.

At the same time, a large literature testing revealed preference using experimental data has also developed. These have employed both laboratory experiments (recent contributions include Andreoni and Miller (2002), Sippel (1997) and Fevrier and Visser (2004)), as well as field experiments utilizing unique subject pools (psychiatric patients in Battalio et al. (1973), children in Harbaugh, Krause, and Berry (2001), tufted capuchin monkeys in Chen, Lakshminarayanan, and Santos (2006)).

It is fair to say that most of the empirical literature, using both field and experimental data, finds relatively few violations of GARP. Therefore, the power of GARP as a test of rationality is a real concern; these issues have been discussed in, *inter alia*, Bronars (1987), Blundell, Browning, and Crawford (2003), Andreoni, Gillen, and Harbaugh (2011), Beatty and Crawford (2011). Experiments suffer less from this problem because they are often carefully designed to avoid power issues (see e.g. Andreoni and Miller (2002)).

At the same time, revealed preference tests are quite stark, allowing for either rational or irrational consumers. In practice, one would like to accommodate a grey area where "small" violations of GARP may not indicate a worrying degree of irrationality (or may indicate imperfections in the data). In the existing literature, various researchers have proposed ways to quantify the degree of violations from GARP, including Afriat (1967), Varian (1985, 1990), Gross (1995), and Heufer (2008).[1]

In terms of assessing the severity of violations of GARP, MPI is closest in spirit to the *efficiency index* originally proposed by Afriat (1967) and subsequently modified by Varian (1990). Jerison and Jerison (2011) is a recent contribution, relating Afriat's efficiency index to a measure of the asymmetry of the Slutsky matrix. Choi et al. (2011) uses Afriat's index on data from a large-scale field experiment; they find some of the same qualitative empirical results as we do using MPI on scanner data (specifically the results we report in Section 4.3).

We review these developments in Section 3.2 below.

3. Methodology

3.1. **Money Pump Index.** Suppose that we observe the purchases of a single consumer when she faces different prices. Observation k ($k = 1, \dots, K$) consists of a consumption bundle $x^k \in \mathbb{R}^l_+$ that the consumer bought at prices $p^k \in \mathbb{R}^l_{++}$.

Let X be the set of all observed consumption bundles. That is, $X = \left\{x^k : k = 1, \dots, K\right\}$. The revealed preference relation on X is the binary relation R defined as $x^k \, R \, x^l$ if $p^k \cdot x^k \geq p^k \cdot x^l$. The strict revealed preference relation is the binary relation P defined as $x^k \, P \, x^l$ if $p^k \cdot x^k > p^k \cdot x^l$.

[1] Apesteguia and Ballester (2010) axiomatize a measure of deviations from rationality. It applies in general choice environments with finitely many choices. It does not use the special structure of Walrasian budgets.

The data satisfy the *weak axiom of revealed preference* (WARP) if whenever $x^k \, R \, x^l$ it is false that $x^l \, P \, x^k$.

The data satisfy the *generalized axiom of revealed preference* (GARP) if there is no sequence $x^{k_1}, x^{k_2}, \ldots, x^{k_n}$ such that

$$x^{k_1} \, R \, x^{k_2} R, \ldots, R x^{k_n} \text{ while } x^{k_n} \, P \, x^{k_1}. \tag{1}$$

A violation of GARP is identified with a sequence $x^{k_1}, x^{k_2}, \ldots, x^{k_n}$. We say that n is the *length* of the sequence.

Given a sequence $x^{k_1}, x^{k_2}, \ldots, x^{k_n}$ for which (1) holds, we can compute the *money pump cost* associated to this sequence as

$$\sum_{l=1}^{n} p^{k_l} \cdot (x^{k_l} - x^{k_{l+1}}), \quad \text{(taking } k_{n+1} = k_1\text{)}.$$

Our money pump cost is measured in dollars. In order to compare this cost across consumers with different budgets, we normalize the cost by each household's total expenditure. Specifically, the money pump index MPI equals the money pump cost as a proportion of total expenditure: if (1) holds for the sequence $x^{k_1}, x^{k_2}, \ldots, x^{k_n}$, we compute the *MPI* of the sequence as

$$MPI_{\{(x^{k_1}, p^{k_1}), \ldots, (x^{k_n}, p^{k_n})\}} = \frac{\sum_{l=1}^{n} p^{k_l} \cdot (x^{k_l} - x^{k_{l+1}})}{\sum_{l=1}^{n} p^{k_l} \cdot x^{k_l}}, \quad \text{(taking } k_{n+1} = k_1\text{)}. \tag{2}$$

Note that MPI is measured for each violation of GARP. In the empirical application in Section 4.2, we sometimes report a household-level MPI by computing the mean and median MPI across the different violations of GARP for a given household.

Remark 1. *Calculating money pump costs can be a huge computational task. For the data we present in Section 4, $K = 26$; so there are*

$$\sum_{k=2}^{26} \binom{26}{k} (k-1)! \; \approx \; 4.39239 \times 10^{25}$$

potential cycles, which are unique up to rotations. There are fast algorithms for checking if GARP has been violated (see Varian (1982)), but they do not suffice to calculate MPI.[2]

[2]Warshall's algorithm, suggested by Varian for checking GARP, can be used to calculate an approximation of MPI.

3.2. **Comparison with Afriat's efficiency index.** We briefly review and compare our approach to the "efficiency indices" proposed by Afriat (1967) and Varian (1990) to quantify violations from GARP. Given $e \in [0,1]$, let R_e and P_e be the binary relations defined by $x^k \ R_e \ x^l$ if $ep^k \cdot x^k \geq p^k \cdot x^l$, and $x^k \ P_e \ x^l$ if $ep^k \cdot x^k > p^k \cdot x^l$. Clearly, if $e = 1$, then R_e is the original revealed preference relation, so if R_1 satisfies GARP then the data are consistent with rationality. At the other extreme, R_0 satisfies GARP trivially. *Afriat's efficiency index (AEI)* is defined as the supremum over all the numbers e such that (R_e, P_e) satisfies GARP.

The ideas behind AEI are similar to our MPI (perhaps unavoidably so, as they try to measure the same phenomenon); but AEI and MPI differ in their interpretations. MPI is the monetary magnitude that can be extracted from a consumer that violates GARP. AEI can be interpreted as a "margin of error" (Varian, 1990) that we allow the agent to make in his consumption choices, or as a tolerance for wasted expenditure.

The behaviors of the MPI and AEI can be quite different, and they can give opposite conclusions on the same data. We present two simple examples to illustrate this point. The first example is in Figure 3, which shows two pairs of observations $\{(z,p),(x',p')\}$ and $\{(x,p),(x',p')\}$ which both violate WARP. The MPI for these observations are:

$$\begin{aligned} MPI_{\{(z,p),(x',p')\}} &= \frac{p \cdot (z - x') + p' \cdot (x' - z)}{p \cdot z + p' \cdot x'} \\ &> \frac{p \cdot (x - x') + p' \cdot (x' - x)}{p \cdot x + p' \cdot x'} = MPI_{\{(x,p),(x',p')\}}, \end{aligned}$$

since, as drawn in Figure 3, $p \cdot x = p \cdot z$ and $p' \cdot (x' - z) > p' \cdot (x' - x)$. We conclude that the violation in the data (z,p), (x',p') is more severe than in (x,p), (x',p').

Calculating the AEI for these pairs of violations yields

$$\begin{aligned} AEI_{\{(z,p),(x',p')\}} = \max\left\{\frac{p' \cdot z}{p' \cdot x'}, \frac{p \cdot x'}{p \cdot z}\right\} &= \frac{p \cdot x'}{p \cdot z} \\ &= \frac{p \cdot x'}{p \cdot x} = \max\left\{\frac{p' \cdot x}{p' \cdot x'}, \frac{p \cdot x'}{p \cdot x}\right\} = AEI_{\{(x,p),(x',p')\}} : \end{aligned}$$

that is, the AEIs are the same for both violations.

As the example illustrates, the crucial difference between AEI and MPI lies in how it treats each difference $p^k \cdot x^k - p^k \cdot x^l$ in a violation of GARP. The MPI simply adds up the differences, and the resulting measure of a violation of GARP is the dollar amount that can be extracted by running the money pump implied by the violation of GARP. The AEI, on the other hand, seeks to "break" the violation of GARP at

INSERT FIGURE 3

its weakest link. Thus, once $e < 1$ deflates a value of expenditure $p^k \cdot x^k$ to the point where there is no violation of GARP, the remaining differences $p^k \cdot x^k - p^k \cdot x^l$ play no role in the measure.[3]

We illustrate the point with a second concrete example. The example exhibits a small violation of GARP according to AEI, but a large one according to MPI. In fact, the violation is negligible according to AEI, while the parameter δ below can be chosen so that MPI is arbitrarily large. Let $p^1 = (\frac{1}{\delta}, \delta)$; $p^2 = (\delta, \frac{1}{\delta})$; $p^3 = (1, 1)$; $x^1 = (\delta^2, 0)$; $x^2 = (\frac{\delta^2}{1+\delta^2}, \frac{\delta^2}{1+\delta^2})$; and $x^3 = (1, 0)$. Then $p^1 \cdot x^1 = p^1 \cdot x^2 = \delta$; $p^2 \cdot x^2 = p^2 \cdot x^3 = \delta$; $p^3 \cdot x^3 = 1$; while $p^3 \cdot x^1 = \delta^2$. So the e-revealed preference relation R_e satisfies GARP for any $e < 1$; AEI therefore equals 1.[4] On the other hand,

$$MPI = \frac{(\frac{1}{\delta}, \delta)((\delta^2, 0) - (\frac{\delta^2}{1+\delta^2}, \frac{\delta^2}{1+\delta^2})) + (\delta, \frac{1}{\delta}) \cdot ((\frac{\delta^2}{1+\delta^2}, \frac{\delta^2}{1+\delta^2}) - (1, 0))}{\delta + \delta + 1} + \frac{(1, 1) \cdot ((1, 0) - (\delta^2, 0))}{\delta + \delta + 1}$$

which $\to 1$ as $\delta \to 0$. So if δ is small, MPI exhibits a large violation of GARP, while the violation is mild according to AEI. Thus the AEI and MPI can give the opposite conclusion on the same data.

The observations also suggest that *MPI and AEI are more likely to differ on longer sequences that violate GARP than on shorter sequences.* As we shall see in Section 4.2, that is indeed the case in our data.

3.3. **Statistical tests using the MPI: how large is "large?"** We formulate a statistical basis for testing whether a violation of WARP could be explained by either simple mistakes on the part of the consumer, or measurement errors in variables. Such an approach to testing WARP was pioneered by Varian (1985, 1990).

[3]Varian modifies AEI by allowing e to vary across the different price vectors, looking at a vector (e_k). Varian's measure is the closest distance to the unit vector $(e_k = 1)$ of a (e_k) with no violations of GARP. This version of AEI suffers from the same problem. When we minimize the distance $\|(e_k) - \mathbf{1}\|$, if we fix the value of one e_k such that there is no violation of GARP, we want to set the other entries in the vector equal to 1.

[4]Similarly, Varian's version of AEI would count these data as basically consistent with GARP, as there are vectors (e^1, e^2, e^3) arbitrarily close to $(1, 1, 1)$ such that the corresponding relations satisfy GARP.

We assume that there is measurement error in prices, such that

$$p = q + \varepsilon,$$

where p are observed prices, q are true but unobserved (latent) prices, and ε is an i.i.d measurement error drawn from a normal distribution with mean μ and variance σ^2, independently across households and prices. The assumption that observed prices contain measurement error is natural for our empirical application below, because differences between shelf prices (which are observed) and transactions prices (which we do not completely observe) constitute the most important source of measurement error in supermarket datasets.[5]

Recall that the money pump cost for observations $((p^1, x^1), (p^2, x^2))$ violating WARP is defined as

$$\begin{aligned}\hat{T}_{MP} &\equiv \left[(p^1 - p^2) \cdot (x^1 - x^2)\right] \mathbb{1}_{\{((p^1,x^1),(p^2,x^2)) \text{ violate WARP}\}} \\ &= \left[(q^1 - q^2) \cdot (x^1 - x^2) + (\varepsilon^1 - \varepsilon^2) \cdot (x^1 - x^2)\right] \mathbb{1}_{\{((q^1+\varepsilon^1,x^1),(q^2+\varepsilon^2,x^2)) \text{ violate WARP}\}}.\end{aligned}$$

Even with the normality assumption on the measurement errors (ε), we cannot derive the distribution of $\hat{T}_{MP}$ under the null of rationality – ie, that $(q^1, x^1), (q^2, x^2)$ satisfy WARP, and that the observed violation of WARP was generated only due to measurement error in prices. However, taking a cue from Varian (1985), we can bound $\hat{T}_{MP}$ by another random variable T_{MP}, and perform the desired hypothesis test based on the distribution of T_{MP} rather than $\hat{T}_{MP}$. Specifically, because $(p^1 - p^2) \cdot (x^1 - x^2) \geq 0$ when $((p^1, x^1), (p^2, x^2))$ violates WARP, we have:

(3)

$$\begin{aligned}\hat{T}_{MP} &= \left[(q^1 - q^2) \cdot (x^1 - x^2) + (\varepsilon^1 - \varepsilon^2) \cdot (x^1 - x^2)\right] \mathbb{1}_{\{((q^1+\varepsilon^1,x^1),(q^2+\varepsilon^2,x^2)) \text{ violate WARP}\}} \\ &= \max\{0, (q^1 - q^2) \cdot (x^1 - x^2) + (\varepsilon^1 - \varepsilon^2) \cdot (x^1 - x^2)\}.\end{aligned}$$

To proceed further, we require an additional assumption. Suppose that we are given consumption bundles x^1 and x^2 that were chosen at some observed prices, p^1 and p^2, respectively. Let us assume that x^1 and x^2 are rational choices at the true (but unobserved) prices; say q^1 and q^2. Then it follows from Afriat's theorem (see Afriat

[5] Varian (1985) assumed measurement error in choices rather than prices. This difference is insubstantial, as the test we describe below would also work for errors in choices. For our empirical application, assuming measurement error in prices seems more natural.

(1967) or Varian (1982)) that there are positive numbers λ^1 and λ^2 such that

$$\lambda^1 q^1 \cdot (x^1 - x^2) + \lambda^2 q^2 \cdot (x^2 - x^1) \leq 0.$$

The numbers λ^1 and λ^2 have a natural interpretation as Lagrange multipliers. They are the marginal utility of a relaxation of the budget constraint; the "marginal utility of income." We make an additional assumption that the two marginal utilities are equal:

Assumption *EMUI (equal marginal utility of income):* $\lambda^1 = \lambda^2$.

The EMUI assumption has a clear economic meaning: at the point at which the purchases were made, an additional dollar in income would result in the same increase in utility. It seems a reasonable assumption for our application to supermarket purchases. Assuming EMUI, we obtain that

$$(q^1 - q^2) \cdot (x^1 - x^2) \leq 0.$$

Using this inequality in Eq. (3), we get

$$\begin{aligned}\hat{T}_{MP} &= \max\{0, (q^1 - q^2) \cdot (x^1 - x^2) + (\varepsilon^1 - \varepsilon^2) \cdot (x^1 - x^2)\} \\ &\leq \max\{0, (\varepsilon^1 - \varepsilon^2) \cdot (x^1 - x^2)\} \equiv T_{MP},\end{aligned}$$

The upshot is that the observed money pump cost $\hat{T}_{MP}$ can be bounded from above by T_{MP} which, given the distributional assumptions on ε, follows a truncated normal distribution with mean zero and variance $2 \cdot \|x^1 - x^2\|^2 \cdot \sigma^2$.

Given a nominal size α, then, we can find a critical value C_α satisfying $P(T_{MP} > C_\alpha) = \alpha$; we set $C_\alpha = F^{-1}_{T_{MP}}(1 - \alpha)$, where $F_{T_{MP}}$ denotes the cumulative distribution function of T_{MP}. However, because $\hat{T}_{MP} \leq T_{MP}$, the "true size" of the test is

$$P(\hat{T}_{MP} > C_\alpha) \leq P(T_{MP} > C_\alpha) = \alpha;$$

so there is a downward size distortion: When MPI is large enough to warrant a rejection, we can do so with at least the desired confidence $1 - \alpha$. However, the test tends to underreject, which raises power problems. We return to the power issue below in discussing the empirical application.

Tests for longer cycles. We focused on a test for violations of WARP, but the same calculation holds for testing general violations of GARP. The condition derived from Afriat's theorem holds for general sequences of demands. Given K observations of consumptions and prices, and generalizing the $EMUI$ assumption accordingly, we obtain

$$\hat{T}_{MP} = \left[\sum_{k=1}^{K} p^k \cdot (x^k - x^{k+1})\right] 1_{\langle (p^k, x^k) \rangle_{k=1}^{K} \text{ violates GARP}\}}$$
$$\leq \max\left\{0, \sum_{k=1}^{K} \varepsilon^k \cdot (x^k - x^{k+1})\right\} \equiv T_{MP};$$

where T_{MP} follows a truncated normal with mean zero and variance $\sum_{k=1}^{K} \|x^k - x^{k+1}\|^2 \cdot \sigma^2$.

4. Empirical Results: incidence and severity of GARP violations

4.1. **Data Description.** We use a household-level scanner panel dataset, the so-called "Stanford Basket Dataset", which contains grocery expenditure data for 494 households from four grocery stores in an urban area of a large US Midwestern city, between June 1991 and June 1993 (104 weeks). This dataset was collected by Information Resources, Inc. (IRI), and has also been used in, among others, Bell and Lattin (1998), Shum (2004), and Hendel and Nevo (2006a,b).

We focus in this paper on households' expenditures on food categories, of which there are fourteen: bacon, barbecue, butter, cereal, coffee, cracker, eggs, ice-cream, nuts, analgesics, pizza, snack, and sugar. We restrict attention to food because we do not expect consumers to change their food expenditure very dramatically in response to changes in income; most foods are basic necessities, and the role for "luxurious" spending on food is arguably more limited than for other types of goods.[6]

We observe 103,345 transactions of 4,082 unique items: i.e unique Universal Product Codes (UPC). Each transaction records the household identity, UPC, transaction week, consumed units, price per unit (shelf prices), and the coupon value (if used).

For the prices in the empirical analysis, we used the "shelf prices", which are the prices posted in the supermarket at the time of purchase. While the majority (86%)

[6]By focusing on food expenditures, our approach requires an assumption that food items are separable in households' preferences, so that purchases of non-foods affect food consumption only through the income left over from such purchases. Hence, our test is implicitly a joint test of rationality and separability for food. However, separability is ubiquitous as an assumption in applied demand analysis, and has been universally assumed in applied work to reduce the dimensionality of demand system (a point emphasized by Deaton and Muellbauer (1980) and Blundell (1988), among others).

of transactions take place at shelf prices, the actual transaction price may differ from the shelf price, primarily due to the household's use of coupon discounts. We did not incorporate coupons into our analysis explicitly, because we only observe coupons when a household uses them, and do not observe coupon availability when either an item is not purchased, or the coupon is not used. This partial observability of coupon discounts is problematic for revealed preference analysis, because the GARP inequalities depend on a counterfactual calculation: how much would a consumption basket cost at prices at which another consumption basket was purchased.[7] For these reasons, we use shelf prices in our analysis, and consider coupon discounts as the primary source of measurement error in the price data.

In order to obtain consistent consumption data over goods, we aggregate transactions by brand name and category: when distinct items have a common brand name, their transactions are aggregated. Hence, each "product" in the sample is a food product with a distinct brand name, and we aggregate across all sizes/presentational forms of each product. Analogously, aggregate prices and discounts at the product level are obtained by averaging the prices and discounts of each size, weighted by the amount consumed. To minimize stockpiling and inventory issues, we also aggregate households' expenditures for each good over time, to a four-week period.

Even after this aggregation, not all brands are consumed for every time period; some products enter, some exit, and many others are simply unpopular items which are infrequently purchased by the households in the dataset. Since GARP requires price observations over every time period, we use only brands for which price data are available for every period. For this reason, we drop 12,976 (or 12.5 %) of the purchases from the dataset.

4.2. **GARP and MPI.** Exploiting the panel nature of our dataset, we next consider GARP violations for each household separately. This allows for arbitrary unobserved heterogeneity at the household-level. The MPI is measured for each violation of GARP; we consider the mean or the median MPI across the household's violations.

Table 1 presents a summary of our results. Out of 494 households, 395 (roughly 80%) of them violate WARP (GARP for sequences of length 2) for at least some pairs of observations. Hence, a significant proportion of households do exhibit violations of WARP; our result is in contrast to much of the previous empirical literature, which

---

[7]Moreover, there are a few cases where discounts (coupon values per unit) exceed shelf prices.

INSERT TABLE 1

INSERT FIGURE 4

fails to find many violations. Given Remark 1, we only check for violations of GARP that involve cycles of limited length: lengths 2, 3 and 4.

In Table 1, moving from left to right, we include progressively longer cycles in testing GARP. When we include cycles of length 3 and 4, thereby searching a substantially larger number of possible cycles (5,525 and 95,225, compared to 325), the overall number of households violating GARP increases only by 1 (from 395 to 396). In theory GARP may be violated when WARP is satisfied. However, Table 1 shows that WARP closely approximates GARP in practice; only one household satisfies WARP while violating GARP. Moreover, the median and mean level of MPI change only slightly as we search over longer cycles.

On the other hand, the severity of the violations, in terms of MPI (see Equation 2), is moderate or small: the mean and median MPI, taken across all households, are only about 6% of total expenditure.

To break this down further, we calculated, for each household, the MPI of each violation of WARP, and obtained the *household-specific* median level of MPI, across all the cycles for this household, which violated WARP. In Figure 4, we plot the cumulative distribution function of this household-specific median MPI, across the 395 households which exhibit some violation of WARP. The function clearly rises very steeply for values of the MPI $< 10\%$, but is largely flat thereafter. This indicates that a large majority of households have very small violations of WARP, and only a handful of households have larger violations, exceeding 10% of expenditures. Thus, large violations do occur, but they are infrequent.

Are these violations of GARP severe? The finding that MPI is small is reinforced by the results of using MPI as a statistical test; we follow the method described in Section 3.3. The observed standard deviation of price discounts by coupons is $\hat{\sigma} = 1.1143$ (measured in cents per unit). Taking this as the value for the standard deviation in measurement error, and at a nominal size of $\alpha = 5\%$, we find no rejections of GARP at all; that is, none of the observed MPIs is large enough to warrant a rejection at this nominal size.

Since, by construction, our test will underreject relative to the nominal size, we reduced the value we used for the standard deviation of measurement error even

INSERT TABLE 2

INSERT TABLE 3

further. Only by reducing this standard deviation by a factor of five, to $\frac{1}{5}\hat{\sigma} = 0.2229$ are we able to reject around 5% of the observed MPIs. Hence, the value of σ which we would need in order to reject GARP at usual significance levels is substantially lower than the standard deviation of price discounts in our data, which is a reasonable proxy for measurement error. Despite the tendency of our procedure to underreject, this is rather convincing evidence that the MPI fails to reject GARP.

Table 1 also exhibits an empirical comparison of MPI with AEI. A large value of AEI indicates a small violation, so we report instead the values of (1-AEI). The correlation between MPI and (1-AEI) may be instructive. The correlation is 0.7834, 0.6250, and 0.5197 for cycles of length 2, 3, and 4, respectively. As suggested by our discussion in Section 3.2, the AEI and MPI differ more the longer is the length of the cycle in the violation of GARP. [8]

4.3. **Demographic Variables.** Next, we consider the demographic determinants of rational (or irrational) consumption behavior, as measured by the MPI. Table 2 shows the population distributions of the demographic variables.[9] The panelists are generally older, and their education levels are higher than the general U.S. population.

Since MPI has a positive value only when consumptions violate WARP, we consider censored Tobit regressions of MPI on demographic variables. Table 3 shows the regression results with 156,000 (=480 × 325) observations: 480 households with $\binom{26}{2}$ possible pairs.

MPI is higher for older, poorer and less educated households, than for younger, richer and more educated households. MPI is also higher for smaller households.

5. Further results

5.1. **Seasonality and stability of preferences.** A consumer may fail GARP because his preferences change: they are not "stable." Given two observations, (x, p)

[8]Using an approximation, we can compute these correlations for longer cycles. The correlation continues to decrease: 0.24 and -0.44 for cycles of length 5 and 6, respectively. The negative correlation at length 6 is purely driven by two specific observations.

[9]The demographic data is missing for 14 households. We drop these from our data set, and work with 480 households.

INSERT FIGURE 5

INSERT TABLE 4

and (x', p'), it is possible that x was a rational choice for a different utility function than x'. We argue that unstable preferences would be reflected in large MPI. Therefore, our empirical findings support the hypothesis that preferences are stable.

Consider a consumer who uses one utility for some purchases, and another utility for other purchases. We argue that this consumer's MPI is positive, for arbitrarily small changes in prices. In fact, MPI is larger when the difference in demands under both utilities is larger, thus implying that when preferences are unstable, MPI can be interpreted as a measure of this instability.

Specifically, consider Figure 5(a). Suppose that a household follows two distinct utility functions. These two utility functions give rise to two different demand functions: $d_1(p, I)$ and $d_2(p, I)$. Fix prices p, and suppose that we observe $x = d_1(p, I)$; see Figure 5(a). The second utility, on the other hand, would give demand $\hat{x} = d_2(p, I)$. Now, by continuity of demand, if we choose prices p' close to p (as in the figure) then $x' = d_2(p', I)$ is close to $\hat{x}$. But this implies a violation of WARP.

The money pump cost of the violation of WARP in Figure 5(a) is $(p' - p) \cdot (x' - x)$. We can look at the money pump cost for an arbitrarily small change in prices. In particular, fix a direction of change in price ∇p, and consider an infinitesimal price change in p in the direction of ∇p. So $p' = p + \varepsilon \nabla p$; for $\varepsilon > 0$. As ε shrinks to 0, x' converges to $\hat{x}$, so the money pump cost approaches $\varepsilon \nabla p \cdot (\hat{x} - x)$; see Figure 5(b). So a small price change gives an increase in money pump cost, as long as the change in prices forms an acute angle with the difference in the demand functions. Note also that a larger difference in demands results in a larger money pump cost, for a given direction of change of prices.

Given this interpretation of the money pump cost as a measure of an agent's changes in preferences, we next look and see whether the MPI (money pump cost as the proportion of expenditure) reflects seasonal trends in demand for certain types of grocery items, because these trends may be attributable to changes in preferences over time. Specifically, we focus on the case of ice cream demand, for which the seasonal peak in demand is during the summer months. If this seasonality is in fact due to changing preferences, then we should expect to see larger MPIs in cycles involving peak and non-peak periods, than in cycles involving only non-peak periods.

Such evidence is presented in Table 4. For this exercise, we aggregate consumption up to 'ice-creams' vs. 'all other foods'. For each pair of periods, we count the number of households violating WARP, and compute the average of their median MPIs. Numbers (or parenthesized numbers) in the table are the numbers of households (or average MPIs), which are averaged over the pairs of periods falling into a corresponding pair of seasons.

Surprisingly, we find no evidence of seasonality. For instance, MPIs are 1.52% between summer and winter months (a peak/non-peak comparison), versus 2.16% between the winter and spring months (two non-peak periods). This suggests that, while seasonality may indeed be present, prices may also be moving in a fashion such that agents' resulting consumption choices do not violate GARP.[10]

More generally, we also repeated this exercise at the disaggregate level, without aggregating across different goods. Overall, we found no systematic patterns between MPIs, and the periods across which we were considering cycles. Again, this suggests that WARP violations are not related to time or seasonality trends.[11]

5.2. **Power of GARP and Bronars index.** The vast majority of empirical studies of revealed preference find very few violations of GARP, a stylized fact which is referred to as the "low power" of GARP. To address this issue, "power indexes" have been developed to quantify the extent to which particular datasets may be useful for testing GARP. A power index seeks to measure the extent to which a collection of observed budget sets can detect violations of GARP. For example, if we have two budgets B and B' that are nested (e.g. $B \subseteq B'$) then no choices by a consumer could reveal a violation of GARP. The earliest and most well-known power index is that of Bronars (1987). Specifically, the Bronars index measures the number of GARP violations under the behavioral assumption that consumer made purely random choices on their observed budget sets.

---

[10] Indeed, Chevalier, Kashyap, and Rossi (2003) provide evidence that prices on grocery items tend to be lower during peak demand periods for these items (see also Nevo and Hatzitaskos (2005) for further discussion). Such "countercyclical" price variation may mask any seasonal variation in preferences, and lead to no violations of revealed preference.

[11] A similar empirical question of demand stability has been addressed in the agricultural economics literature using revealed preference methods; eg. Chalfant and Alston (1988) and Jin and Koo (2003).

Typically, in cross-sectional data, income variability is much higher than price variability; as a result, budget sets tend to be nested, and no choices that exhaust the consumer's budget can violate GARP. In contrast, in our application using scanner data, we found a large fraction of the households violate GARP. Nevertheless, we still computed the Bronars index for the budgets observed in our data. Firstly, we draw random budget shares of goods from a continuous uniform distribution. The generated budget shares are then multiplied by observed total expenditure. We obtain the random consumption of each commodity by dividing its budget share by the price. We subsequently check GARP violation for each observation.[12]

We find, surprisingly, that the Bronars index indicates low power. We repeat this procedure 100 times and find that amongst 494 there are on average only 3-4 households violating GARP for each generated panel data set, much lower than the number of GARP violations observed in the actual choice data.

Our finding illustrates a basic problem with the Bronars index. A power index should reflect the probability that GARP is rejected when it should be rejected: but Bronars calculates the probability of rejecting GARP under purely random behavior. It is well known that random behavior can be very close to rational (see Becker (1962)). One would instead want to measure the degree to which GARP is rejected under some alternative, clearly irrational, model of behavior. Bronars does not provide such a model; indeed, it seems very difficult to find an acceptable alternative benchmark to rationality under which to measure power – there is, in a sense, only one way to be rational, but many more ways to be irrational.[13]

6. Conclusion

We present a new measure of the severity of a violation of GARP, the *money pump index (MPI).* The measure is based on the idea that a consumer who violates the axiom is subject to being exploited as a money pump. The MPI has a simple

[12]This computational method is the so-called "Method 2", in Andreoni, Gillen, and Harbaugh (2011).

[13]There are alternative power indexes formulated by Famulari (1995) and by Andreoni, Gillen, and Harbaugh (2011). Andreoni, Gillen, and Harbaugh's test, in particular, rests on a clever "reversion" of AEI to measure how far an observation that satisfies GARP is from not satisfying it. Since we are not mainly concerned with power, we have not calculated these alternatives measures on our data set. They should be important to assess any empirical finding that fails to detect violations of GARP.

interpretation as the certain dollar amount that can be extracted from a consumer who behaves irrationally.

We carry out an empirical application, using a scanner panel data set of households' food purchases at supermarkets. Almost all of the households in our sample violate the revealed preference axioms over the two year sample period; on average, however, the MPI calculated for these violations is small, suggesting that the violations of revealed preference are not severe. This is supported by formal statistical testing, by which the hypothesis of consumer rationality cannot be rejected.

References

Afriat, Sidney N. 1967. "The Construction of Utility Functions from Expenditure Data." *International Economic Review* 8 (1):67–77.

Andreoni, James and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica* 70 (2):737–753. URL `http://www.jstor.org/stable/2692289`.

Andreoni, Jim, Ben Gillen, and William T. Harbaugh. 2011. "The Power of Revealed Preference Tests: Ex-Post Evaluation of Experimental Design." Mimeo, UCSD.

Apesteguia, Jose and Miguel A. Ballester. 2010. "A Measure of Rationality and Welfare." Mimeo, Universitat Pompeu Fabra.

Battalio, Raymond, John Kagel, Robin C Winkler, Edwin B. Fischer, Robert Basmann, and Leonard Krasner. 1973. "A Test of Consumer Demand Theory using Observations of Individal Consumer Purchases." *Western Economic Journal* 11 (4):411–428.

Beatty, Timothy K. M. and Ian A. Crawford. 2011. "How Demanding Is the Revealed Preference Approach to Demand?" *The American Economic Review* 101 (6):2782–2795.

Becker, Gary S. 1962. "Irrational Behavior and Economic Theory." *The Journal of Political Economy* 70 (1):1–13.

Bell, David and James Lattin. 1998. "Shopping Behavior and Consumer Preference for Store Price Format: Why 'Large Basket' Shoppers Prefer EDLP." *Marketing Science* 17:68–88.

Blundell, Richard. 1988. "Consumer Behaviour: Theory and Empirical Evidence–A Survey." *The Economic Journal* 98 (389):16–65.

Blundell, Richard, Martin Browning, and Ian Crawford. 2003. "Nonparametric Engel Curves and Revealed Preference." *Econometrica* 71 (1):205–240.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves." *Econometrica* 75 (6):1613–1669.

Bronars, Stephen G. 1987. "The Power of Nonparametric Tests of Preference Maximization." *Econometrica* 55 (3):693–698.

Chalfant, James A and Julian Alston. 1988. "Accounting for Changes in Tastes." *Journal of Political Economy* 96 (2):391–410.

Chen, Keith, Venkat Lakshminarayanan, and Lauri Santos. 2006. "How Basic are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior." *Journal of Political Economy* 114:517–537.

Chevalier, Judith A., Anil K Kashyap, and Peter E. Rossi. 2003. "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data." *American Economic Review* 93 (1):15–37.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2011. "Who is (more) Rational?" Mimeo, UC Berkeley.

Deaton, Angus S. and John Muellbauer. 1980. *Economics and Consumer Behavior*. Cambridge, UK: Cambridge University Press.

Famulari, Melissa. 1995. "A household-based, nonparametric test of demand theory." *The Review of Economics and Statistics* 77 (2):372–382.

Fevrier, Philippe and Michael Visser. 2004. "A Study of Consumer Behavior using Laboratory Data." *Experimental Economics* 7:93–114.

Gross, John. 1995. "Testing data for consistency with revealed preference." *The Review of Economics and Statistics* 77 (4):701–710.

Harbaugh, William T., Kate Krause, and Timoth R. Berry. 2001. "GARP for kids: On the development of rational choice behavior." *American Economic Review* 91 (5):1539–1545.

Hendel, Igal and Aviv Nevo. 2006a. "Measuring the Implications of Sales and Consumer Stockpiling Behavior." *Econometrica* 74:1637–1674.

———. 2006b. "Sales and Consumer Inventory." *RAND Journal of Economics* 37:543–561.

Heufer, Jan. 2008. "A Geometric Measure for the Violation of Utility Maximization." *Ruhr Economic Papers* .

Hoderlein, Stefan and Jörg Stoye. 2009. “Revealed Preferences in a Heterogeneous Population.” Mimeo, Dep. of Economics, Brown University.

Jerison, David and Michael Jerison. 2011. “Real Income Growth and Revealed Preference Inconsistency.” Mimeo, MIT and SUNY Albany.

Jin, Hyun J. and Won W. Koo. 2003. “The effect of the BSE outbreak in Japan on consumers’ preferences.” *European Review of Agricultural Economics* 30:173–192.

Laibson, David and Leeat Yariv. 2007. “Safety in Markets: An Impossibility Theorem for Dutch Books.” Mimeo, California Institute of Technology.

Mulligan, Casey. 1996. “A Logical Economist’s Argument Against Hyperbolic Discounting.” University of Chicago, mimeo.

Nevo, Aviv and Konstantinos Hatzitaskos. 2005. “Why Does the Average Price of Tuna Fall During Lent?” NBER Working Papers 11572, National Bureau of Economic Research, Inc.

Rabin, Matthew. 2002. “A perspective on psychology and economics* 1.” *European Economic Review* 46 (4-5):657–685.

Richter, Marcel K. 1966. “Revealed Preference Theory.” *Econometrica* 34 (3):635–645.

Samuelson, Paul A. 1938. “A note on the pure theory of consumer’s behaviour.” *Economica* 5 (17):61–71.

Shum, Matthew. 2004. “Does Advertising Overcome Brand Loyalty? Evidence from Breakfast Cereals.” *Journal of Economics and Management Strategy* 13:241–272.

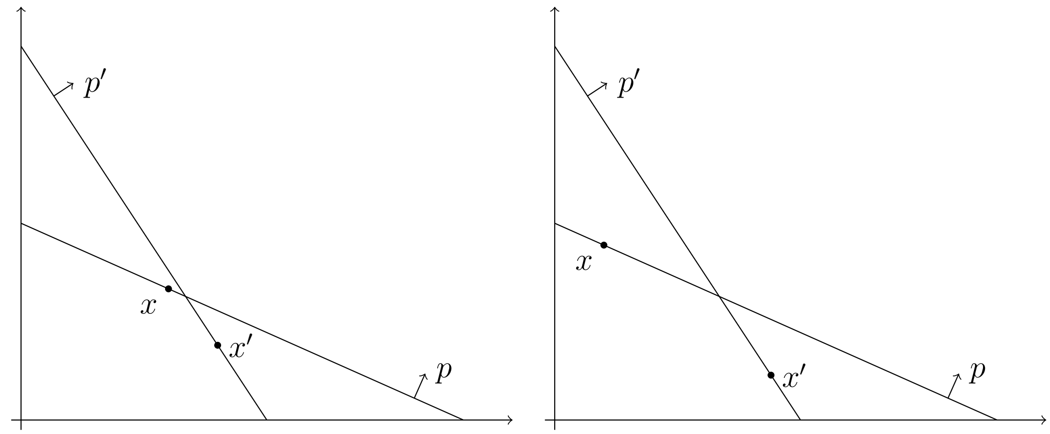
Sippel, Reinhard. 1997. “An Experiment on the Pure Theory of Consumer’s Behavior.” *Economic Journal* 107:1431–1444.

Varian, Hal R. 1982. “The Nonparametric Approach to Demand Analysis.” *Econometrica* 50 (4):945–974.

———. 1985. “Non-Parametric Analysis of Optimizing Behavior with Measurement Error.” *Journal of Econometrics* 30:445–458.

———. 1990. “Goodness-of-fit in optimizing models.” *Journal of Econometrics* 46 (1-2):125 – 140.

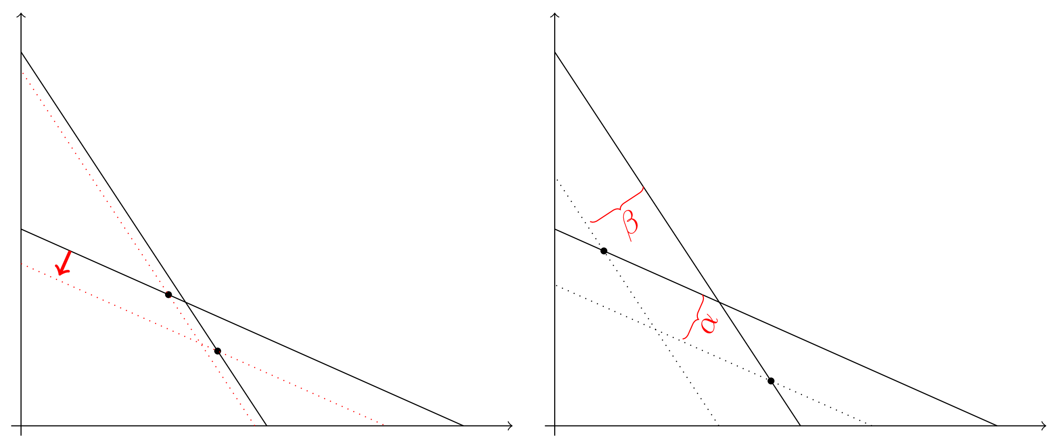
———. 2006. “Revealed preference.” *Samuelsonian economics and the twenty-first century* :99–115.



(a) (x, p) and (x', p') violate GARP (in fact WARP).

(b) A more "severe" violation of GARP.

FIGURE 1. Two observations: (x, p) and (x', p').



(a) Two observations: (x, p) and (x', p').

(b) $mp = \alpha + \beta; \alpha = p \cdot (x - x'), \beta = p' \cdot (x' - x)$.

FIGURE 2. Money pump costs for Figure 1.

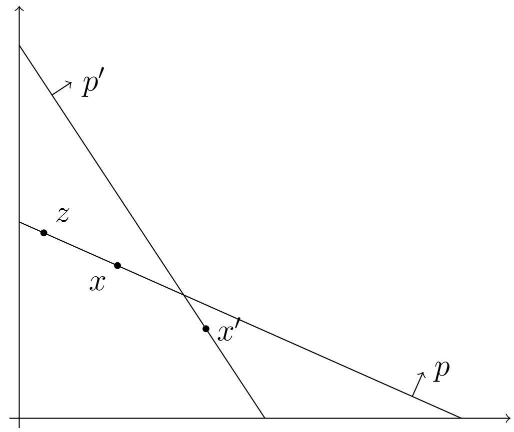


FIGURE 3. Two violations of WARP: (x, p), (x', p') and (z, p), (x', p').

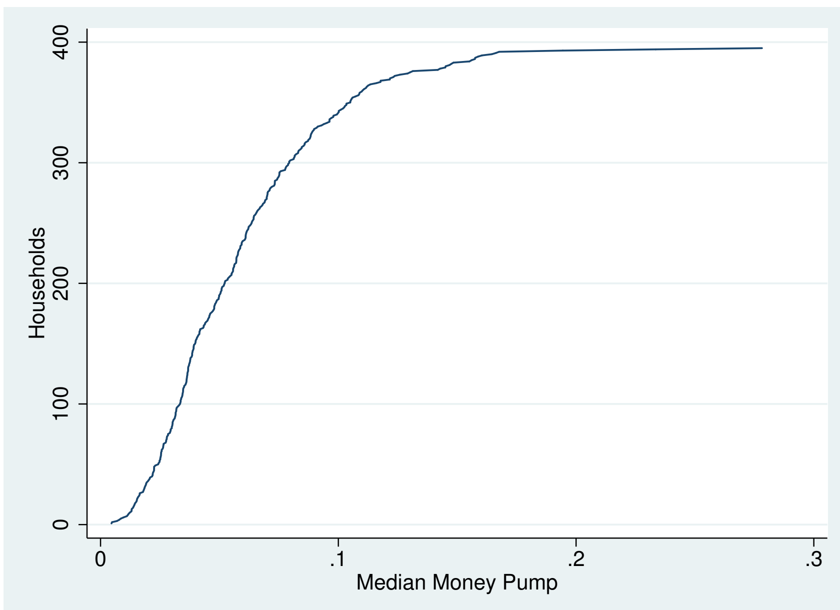
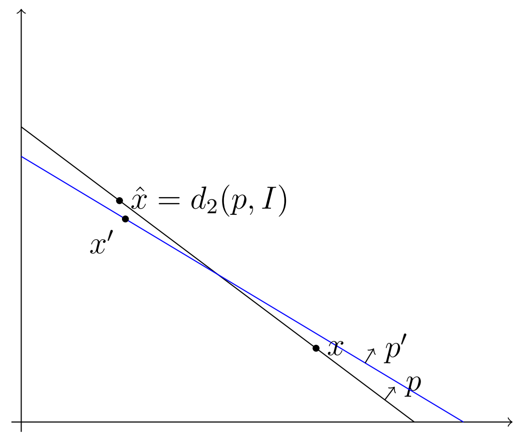
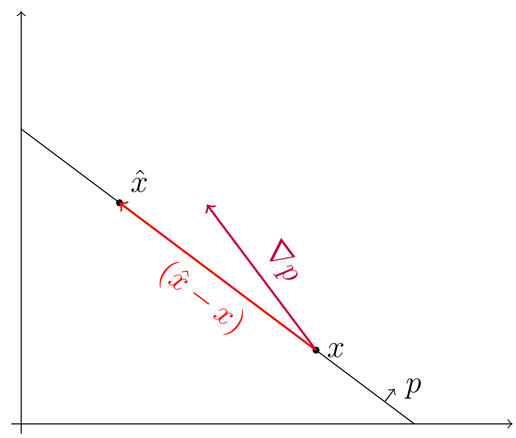


FIGURE 4. Cumulative distribution of households' median MPI



(a) $x = d_1(p, I)$, and $x' = d_2(p', I)$.



(b) $mp = \varepsilon \nabla p \cdot (\hat{x} - x)$.

FIGURE 5. Unstable preferences.

Cycle Lengths Included	{2}	{2,3}	{2,3,4}
Total Numb. Households	494	494	494
Households Violating GARP	395	396	396
Median[a] MPI	5.97%	5.95%	5.91%
Mean[a] MPI	6.22%	6.12%	6.09%
Median[b] (1-AEI)	2.40%	2.09%	2.03%
Mean[b] (1-AEI)	2.63%	2.38%	2.31%
Possible Cycles	325	5525	95225
Median Numb. Violations	2	3	3
Mean Numb. Violations	3.028	4.833	6.008

[a]: numbers in these rows are the median/mean MPI's across all violations of GARP among the sample households.

[b]: numbers in these rows are the mean/mean values of (1-AEI) across all violations of GARP among the sample households.

TABLE 1. Money Pump: calculated by Equation 2, averaged over households violating GARP

Variables:		# Households
Family Size	Mid Size (3,4 members)	187
	Large Size (> 4 members)	65
Income	Mid Annual Income ($\in [\$20,000, \$45,000]$)	200
	High Annual Income ($> \$45,000$)	141
Age[a]	Mid Age	201
	Old Age	157
Education[b]	High School	197
	College	255
Total Households:		480

Table 2. Demographic Variables

[a]Middle-aged households are those in which the average of the spouses' ages is between 30 and 65; in old-aged households, this average exceeds 65.
[b]If both spouses are present in a household, the average education of both spouses is reported.

Variable	Regression 1	Regression 2	Regression 3
MidAge			0.0112
			(1.92)
OldAge			0.0126
			(1.74)
MidFamily	-0.0161***	-0.0118*	-0.0085
	(-3.55)	(-2.49)	(-1.66)
LargeFamily	-0.0281***	-0.0243***	-0.0190*
	(-4.04)	(-3.45)	(-2.52)
MidIncome		-0.0181***	-0.0169**
		(-3.44)	(-3.02)
HighIncome		-0.0154*	-0.0142*
		(-2.57)	(-2.16)
High School	-0.0154	-0.00874	-0.00768
	(-1.84)	(-1.02)	(-0.89)
College	-0.0162	-0.00666	-0.00441
	(-1.95)	(-0.76)	(-0.49)
Constant	-0.452***	-0.450***	-0.463***
	(-30.63)	(-30.57)	(-27.56)
σ^a	0.195***	0.195***	0.195***
	(36.65)	(36.66)	(36.66)
N	156000	156000	156000

^{a}: estimated standard deviation of errors in Tobit regression.

t statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

TABLE 3. MPI explained by demographic variables: Results from Tobit (censored) regressions

	Spring	Summer	Fall	Winter
Spring	1.3333	0.6944	1.0556	0.9375
	(1.90%)	(1.84%)	(2.00%)	(2.16%)
Summer	·	1.2000	0.8333	1.2292
		(2.13%)	(1.65%)	(1.52%)
Fall	·	·	1.0667	0.9583
			(1.82%)	(1.78%)
Winter	·	·	·	1.6429
				(1.87%)

Each cell contains number of households which violate WARP between any two months, as classified by the season of the months. (Corresponding average MPIs for these violations are in parentheses.)

TABLE 4. No evidence of changing preferences: Ice-cream vs. Other foods.