HOUSEHOLD LABOR SUPPLY AND TAXES: A NONPARAMETRIC, REVEALED PREFERENCE APPROACH

Melissa Famulari*

Department of Economics
University of California – San Diego
La Jolla, CA 92093-0508
mfamulari@ucsd.edu

December 13, 2006

Abstract

This paper uses the Generalized Axiom of Revealed Preference (GARP) to test the static labor supply model. The advantages of GARP are that it tests for utility maximization without making parametric assumptions and accommodates piecewise linear and nonconvex budget sets. Further, GARP tests are an important preliminary step before estimating a parametric labor supply model because GARP helps identify which assumptions are particularly troubling. Tests of GARP are applied to the hours of work and disaggregated commodity choices of individual households' from the 1982-85 Consumer Expenditure Survey. The leisure and commodity choices of demographically similar households resoundingly reject GARP. To find the source of the violations, GARP tests on commodity choices alone are performed and the joint hypotheses of utility maximization and common preferences within demographic group are not rejected. This suggests that the source of the rejections is the labor supply decision. This paper then highlights the role that measurement error in hours, combined with calculating the hourly wage as earnings divided by hours, plays in directly leading to violations of revealed preference. Leisure and commodity choices are generally consistent with GARP when using predicted wages and the tax schedule to model the budget constraint. Therefore, as long as researchers instrument for the wage in the nonlinear budget constraint setting, the search for less restrictive empirical models of static labor supply is likely to be worthwhile.

* I have been working on this paper for years and so would like to thank my colleagues at the Bureau of Labor Statistics, the University of Texas-Austin and the University of California-San Diego. This paper greatly benefited from comments made by Kate Antonovics, Julie Cullen, Roger Gordon, and Preston McAfee.

I. INTRODUCTION

Preference parameters from labor supply models are crucial to understanding the incentive effects of government transfer programs, social insurance programs, and the tax system. Though economists have long recognized the importance of accurate estimates of income and substitution effects, providing a convincing set of empirical estimates has proved difficult. Empirical studies frequently estimate negative compensated labor supply elasticities, violating the fundamental prediction of labor supply models.[1]

Why do we get these puzzling results? Estimation is greatly complicated by the piecewise-linear and nonconvex budget constraints created by the U.S. tax system and transfer programs. The fundamental difficulty is that the return to an additional hour of work is a function of the number of hours the individual chooses to work. This endogeneity of the wage occurs because Federal income taxes are progressive, social security taxes are capped at a specific income level, and transfer payments depend on income earned. A further complication of piecewise-linear and nonconvex budget constraints is that responses to changes in the budget constraint may depend critically upon an individual's location on the budget set e.g., whether the individual is at a kink in the budget constraint or not. Addressing these empirical issues has been a major focus in applied labor economics.[2]

Given the complicated nature of the budget constraint, particularly simple functional forms are used to model preferences and many restrictive assumptions are imposed on the empirical models including a normally distributed error (even though there are large spikes in the weekly and annual hours distributions), linearity in income and substitution effects (which does not allow for backward bending labor supply), and separability of consumption and leisure. As noted by

---

[1] See for example, Bloomquist (1996), Macurdy, Green, and Paarsch (1990), Columbino and Del Bocca (1990), Van Soest, Woittiez and Kapteyn (1990), and Triest (1990).

[2] For reviews of the literature see Blundell and Macurdy (1999) and Heim and Meyer (2003). For reviews of older literature see Pencavel (1986), Killingsworth and Heckman (1986), and Moffit (1990).

Moffitt (1990), Macurdy, Green, Paarsch (1990) and Heim and Meyer (2003), the apparent rejection of static labor supply models may be due to mis-specified empirical models.

It is possible to test for optimizing behavior without imposing any parametric specification of preferences and without simplifying the budget constraint: determine whether individual choices and budget constraints are consistent with the Generalized Axiom of Revealed Preference (GARP).[3] An important advantage of this nonparametric approach is that the budget constraint can be piecewise linear and nonconvex with little complication for tests of GARP. The key building block of tests of GARP is determining the set of commodity bundles that are on or inside the individual's budget constraint. Even if the budget constraint is quite complicated, it is straightforward to determine the set of commodity bundles that are affordable at current prices.

I apply tests of GARP to the disaggregated consumption expenditures and hours of work reported by households in the 1982-85 Bureau of Labor Statistics' Current Expenditure Survey (CE). These are the only U.S. data which report individual hours worked per week, weeks worked per year, annual earnings, and disaggregated household expenditures on commodities. Commodity quantities are determined by dividing household expenditures by intertemporal-interarea price indexes. Gross hourly wages are determined by dividing household reported annual earnings by annual hours of work. The budget constraint incorporates the parameters of the Federal, State, and City tax codes as well as Social Security payroll taxes for each of the four years, 18 states, and 25 cities in the data. To avoid assuming that all households have the same utility function, households are grouped into 23 different subsamples based on their demographic characteristics and nonparametric tests are performed on each of the subsamples.

The joint hypotheses of optimizing behavior and common preferences among households with similar demographic characteristics are resoundingly rejected, for both single person

[3] If an individual chooses a commodity bundle *i* when bundle *j* is affordable, then a preference for *i* over *j* has been established. A preference for *i* over *j* can also be established indirectly, e.g., if *i* is preferred to *k* and *k* is preferred to *j* then *i* is indirectly revealed preferred to *j*. GARP states that if a preference for *i* over *j* has been established, then when we do see the individual consuming bundle *j*, it must be because bundle *i* is not affordable at *j*'s prices.

households and families. These results contrast sharply with tests of GARP on these same households' commodity choices which are reasonably consistent with the joint hypotheses that households with similar demographic characteristics have common preferences and preference maximize. Thus, the failure of GARP is attributable to the hours of work decision which suggests that *any* parametric model of static labor supply is likely to be rejected by the data.

GARP tests are then used to identify the source of the rejection of the static model. It has long been known that hourly wages calculated as the ratio of earnings to hours worked will bias downward the wage coefficient in a labor supply regression when hours are measured with error. A key insight provided by revealed preference tests is that measuring the hourly wage in this manner leads to more than a bias when hours are measured with error: it leads to data that are inconsistent with a utility function.[4] Tests of GARP on simulated data are used to determine whether measurement error in hours worked could explain the rejections of the static labor supply model. The simulations start with data that pass GARP. Measurement error is then randomly added to the hours of work quantity and GARP is re-tested. If measurement error in hours worked is twice the standard deviation estimated in the Bound et. al. (1994) validation study *and* the hourly wage is calculated as earnings divided by hours worked, it is possible to generate as many violations of GARP in the simulations as were found in the CE data. A wage instrument greatly improves consistency with GARP in the simulations, even when hours of work are reported with error.

Finally, I return to the CE data to see whether using a wage instrument in the GARP tests leads to data that is consistent with the static labor supply model. Each worker in the CE is assigned a predicted hourly wage based on their demographic characteristics. The consumption and hours of work choices of the households in each of the 23 demographic groups are far more

---

[4] Borjas (1980) and Eklof and Sacklen (2000) examine the effects division bias on compensated labor supply elasticities. Both papers find that the estimated compensated labor supply elasticity is not simply a smaller positive number but is negative, a finding that is consistent with the insight provided by these revealed preference tests.

consistent with revealed preference axioms when using the wage instrument and the tax schedule to define the budget set. Since the underlying data are generally consistent with maximization of a common utility function, for 22 or the 23 groups, proceeding to estimate a parametric static labor supply model that instruments for the wage and allows for nonlinear, nonconvex taxes is reasonable.

The paper is organized as follows. Section two describes the nonparametric tests for preference maximization, GARP and WARP, for the static labor supply model. Section three describes how I allow for taxes in these tests. In section four I discuss the issues involved in rejecting GARP. Section five describes the data used in these tests of GARP. Section six presents the GARP test results for the 23 demographic groups and for the simulated data.

II. NONPARAMETRIC TEST FOR PREFERENCE MAXIMIZATION

Assume there are *m* commodities, *N* individuals and for each individual *i* we observe a price vector $P_i = \{P_{1i}, P_{2i}, ..., P_{mi}\}$, a quantity vector $X_i = \{X_{1i}, X_{2i}, ..., X_{mi}\}$, wages W_i, and hours worked H_i. The difference in *i's* expenditures if she purchases *j's* consumption bundle is denoted $P_i(X_i - X_j) = \sum_k P_{ki} X_{ki} - \sum_k P_{ki} X_{kj}$, for goods *k* = 1,...,m. Let *T* be total time available. Leisure can be determined from the relation $L_i = T - H_i$. Define the direct revealed preference operator (DRP) as follows:

(1) *i* DRP *j* if and only if $P_i(X_i - X_j) + W_i(L_i - L_j) >= 0$.

However, since $L_i = T - H_i$ and $L_j = T - H_j$, (1) can be rewritten as:

(2) *i* DRP *j* if and only if $P_i(X_i - X_j) + W_i(H_j - H_i) >= 0$.

If the consumption-hours bundle *i* is chosen when both *i* and *j* are affordable, then a preference for *i* over *j* has been revealed directly. The Weak Axiom of Revealed Preference (WARP) states:

i DRP *j* implies $P_j(X_j - X_i) + W_j(H_i - H_j) < 0$. WARP states that if *i* DRP *j*, then when we observe *j* being consumed, it must be that *i* is not affordable at *j*'s wages and prices.

In addition to a *direct* revealed preference relationship, we can define a revealed preference (RP) operator:

(3) *i* RP *j* if and only if $P_i(X_i - X_s) + W_i(H_s - H_i) >= 0$,

$$P_s(X_s - X_r) + W_s(H_r - H_s) >= 0, ..., P_z(X_z - X_j) + W_z(H_j - H_z) >= 0$$

A preference for *i* over *j* is revealed if a chain of DRP relationships links *i* to *j*, e.g., if *i* DRP *s*, *s* DRP *r*, and *r* DRP *j* then *i* is revealed preferred to *j*. The Generalized Axiom of Revealed Preference (GARP) states *i* RP *j* implies $P_j(X_j - X_i) + W_j(H_i - H_j) < 0$. GARP states that if *i* RP *j*, then when we observe *j* being consumed, it must be because at *j*'s wages and prices, *i* is not affordable. Varian (1982) proves that if a set of data is consistent with GARP, then that data can be "rationalized" by a nonsatiated, continuous, concave, monotonic utility function.

These tests of GARP are static because they are based on current expenditures and earnings. Household *i* has negative nonlabor income if $P_i X_i - W_i H_i < 0$ (the household is saving) or positive nonlabor income if $P_i X_i - W_i H_i > 0$ (the household is dissaving). It is clearer where the static nature of these GARP tests comes in if I re-write expression (2) as *i* DRP *j* if and only if $(P_i X_i - W_i H_i) - (P_i X_j - W_i H_j) > 0$. Thus, *i* DRP *j* if (at *i*'s prices) *i* can maintain the same or greater level of savings (dissaving) when purchasing *j*'s consumption bundle and working *j*'s hours. Why household *i* is saving or dissaving is not addressed.

III. TAXES AND REVEALED PREFERENCE

Given an individual's budget constraint, it is trivial to determine the direct revealed preference relationships between her choice and any other consumption-hours bundle. If the alternative consumption-hours bundle is either on or inside the individual's budget constraint, then her choice is directly revealed preferred to the alternative; if the alternative consumption-hours bundle is outside the individual's budget constraint, then her choice is not directly revealed preferred to the alternative. Note that the budget constraint can be piecewise linear and nonconvex without increasing the complexity of the test for GARP. Taxes operate solely through redefining which consumption-hours choices are "affordable"; the test for consistency with revealed preference axioms proceeds as usual.

Even in the nonparametric context, however, introducing nonconvexities has a cost: consistency with GARP is no longer a sufficient condition for the existence of a nonsatiated, continuous, concave, monotonic utility function. A common utility function may not rationalize observed choices even if the data are consistent with GARP.[5] However, because consistency with GARP is a necessary condition for preference maximization, determining whether a data set passes GARP is an important first step before parametric estimation.

I modify the DRP relationship to allow for taxes in two different manners: one based on a constant marginal tax rate, DRP(mt), and the other using the tax schedule, DRP(tb). The constant marginal tax rate specification assumes that the individual faces an exogenous linear budget constraint at her chosen number of hours of work with a constant marginal tax rate, *t*, and virtual income, PX-(1-t)WH. This specification of the budget constraint is one of the two main approaches to incorporate taxes in parametric labor supply models (see discussions in Blundell and Macurdy (1999) and Heim and Myer (2003)). Revealed preference clarifies the problem that linearizing the budget constraint creates for tests of the static labor supply model. In Figure 1, the solid line represents individual *i*'s budget constraint and the small dots represent individual *j*'s budget constraint. Note that *i* and *j* do not violate GARP: *j*'s commodity-hours bundle is directly revealed preferred to *i*'s and, when we observe *i*'s commodity-hours bundle being consumed, it is

[5] Consider a tax schedule that generates a convex choice set (and so is concave to the origin). Suppose two different points on this budget set are chosen by two different individuals. No quasi-concave utility function can rationalize both choices, although they are consistent with GARP because GARP is not violated by multiple consumption choices along the same budget set. The Strong Axiom of Revealed Preference (SARP) would be violated by these choices because SARP does not allow for multiple optima. Tests of SARP are inappropriate because multiple optima along a linear portion of the budget set are consistent with a quasi-concave utility function.

because *j*'s commodity-hours bundle is not affordable at *i*'s prices. Linearization of *i*'s budget constraint (the dashed line) leads the researcher to incorrectly conclude that commodity-hours bundle *j* is affordable *i*'s prices. Thus linearization of the budget constraint leads to a violation of GARP.

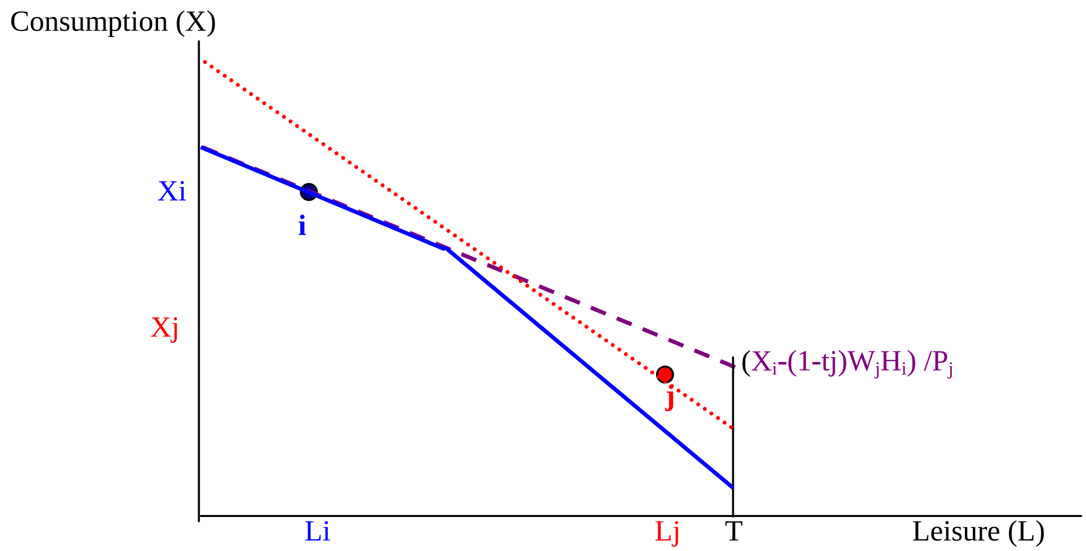

Figure 1: Linearization of the Budget Constraint Leads to a GARP Violation

Note that linearizing the budget constraint may be a reasonable approximation if the difference between the hours worked by individual *i* and individual *j* is relatively small. To examine how costly this simplification of the budget constraint is to tests of the static labor supply model, I modify the DRP relationship as follows:

(4) *i* DRP(mt) *j* if and only if $P_i(X_i - X_j) + (1-t_i)W_i(H_j - H_i) >= 0$.

where t_i is *i*'s marginal tax rate.

In the second modification of the DRP relationship to allow for taxes, I use all the parameters of the tax system to determine whether an individual *i* could afford the consumption-hours bundle chosen by individual *j*. Using *i*'s tax schedule and prices, I determine whether individual *i* can maintain the same or greater level of savings, work *j*'s hours, purchase *j*'s consumption bundle, *and* pay the tax bill implied by individual *j*'s hours choice. Thus:

(5) *i* DRP(tb) *j* if and only if $P_i(X_i - X_j) + W_i(H_j - H_i) + (TAXBILL_i - TAXBILL_{ij}) >= 0$.

where $TAXBILL_i$ is the total tax bill paid by person *i* and $TAXBILL_{ij}$ is what person *i*'s total tax bill would be if person *i* works person *j*'s hours.

Note that taxes solely operate through determining which commodity-hours bundles are affordable. Definitions of the revealed preference operators, WARP and GARP are analogous to those in the no tax case.

IV. REJECTING GARP

IV. A. Violation Rate

Strictly speaking, if even one pair of observations violates GARP, utility maximization is rejected. Given measurement or small optimization errors, it seems unreasonable to specify the acceptance region of the test as 100% consistency with GARP. In addition, increasing the sample size does not affect the rejection criterion even though the potential number of GARP violations rises approximately exponentially with the number of observations.

If there are *N* observations, then there are N*(N-1) pairwise comparisons of consumption-hours bundles that are either consistent with or violate GARP. Define an N x N matrix *A* and let $[a_{ij}]=0$ when *i* not RP *j* and *j* not RP *i*, $[a_{ij}]=1$ if *i* RP *j* and *j* DRP *i* (*i* and *j* violate GARP), and $[a_{ij}]=2$ otherwise. The number of violations, *V*, equals the number of terms in the *A* matrix where $[a_{ij}]=1$. The number of GARP consistent comparisons, *C*, is defined to be the number of terms in the *A* matrix where $[a_{ij}]=2$ and $i \neq j$. Define the violation rate to be V / (V + C).[6] The violation rate is a summary statistic that indicates the consistency of a set of data with the hypothesis of optimizing behavior and homogeneous preferences. Whether the hypothesis is rejected depends upon the researcher's tolerance for type I and type II errors.

IV.B. Increasing the Power of the Test: Differences in Incomes

Identifying the number of pairwise comparisons that violate GARP is straightforward using Varian's (1982) algorithm. The number of GARP consistent comparisons may be misleading because it is possible to be trivially consistent with GARP. When income differences are large, violations of GARP are unlikely, regardless of preferences, and so add little to the power of the test. The additional income required to consume *i*'s bundle at *j*'s prices is given by $[P_j X_i - (1-t_j)W_j H_i] - [P_j X_j - (1-t_j)W_j H_j]$. Figure 2, illustrates this Slutsky measure of the income effect.

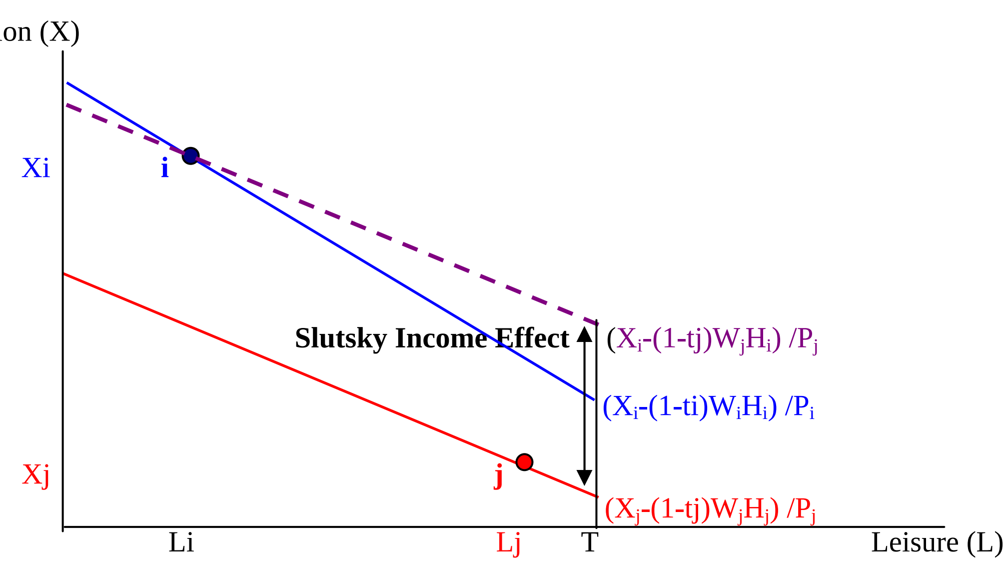

Figure 2: The Slutsky Income Effect

[6] Pairwise comparisons where $[a_{ij}]=0$ are not included in the violation rate because when the possibility of measurement error is introduced into the tests of GARP below, the number of comparisons where neither consumption-hours bundle is revealed preferred will increase.

Define the normalized income gap[7] to be:

$$(6)\quad \left| \frac{[P_j X_i - (1 - t_j)W_j H_i] - [P_j X_j - (1 - t_j)W_j H_j]}{0.5 P_j (X_i + X_j)} \right|$$

To increase the power of the test, I calculate the violation rate separately for those pairs of consumption-hours choices with similar incomes. I also use expression (6) to measure the normalized income gap for tests of GARP using the tax schedule budget constraint. Define an N x N matrix, *Z*, and let $[z_{ij}]=[a_{ij}]$ when *i* and *j* have similar incomes and $[z_{ij}]=0$ otherwise. Again the violation rate is calculated as V / (V + C) where *V* and *C* are now defined by elements in the *Z* matrix.

There are two parameters, chosen by the researcher, which determine the rejection region for the hypothesis test: (i) the critical value of the violation rate and (ii) the critical value of the normalized income gap, which determines the number of comparisons with similar incomes. A violation rate of 5%, a standard significance test critical value, is probably a reasonable criterion for the rejection of GARP. The critical value for the difference in income is less obvious. Higher critical values for income differences result in less powerful tests; we are more likely to make type II errors and accept the hypothesis of utility maximization regardless of the true model of consumer choice. Low critical values for income differences increase the likelihood of making type I errors and rejecting the null hypothesis due to measurement error. Famulari(1995) found that most violations of GARP occur among pairwise comparisons where the normalized income gap is less than .20. In the tests of GARP below, I present the violation rate across all pairwise

[7] Suppose, at *j*'s prices, the income gap is $5,000. Is this large or small? An income gap of $5,000 seems large if expenditures are $10,000 but small if expenditures are $50,000. Therefore, I normalize the size of the income gap by calculating it as a fraction of average total expenditures on bundles *i* and *j* (at *j*'s prices).

comparisons first and then among the pairwise comparisons where the normalized income gap is less than .20 to increase the power of the tests.

IV.C. Nonprice Determinants of Demand

Cross section tests of optimizing behavior jointly test the hypotheses of preference maximization and the existence of a single utility function for a group of consumers. The goal of this paper is to determine whether differences in similar households' consumption and hours of work choices can be explained solely by differences in prices, wages, and nonlabor income. Therefore, I test GARP on groups of households that are relatively homogeneous.

In parametric studies, differences in tastes across demographic groups can be introduced by allowing the parameters of the estimated utility function to vary across demographic groups either (i) by specifying that certain of the demand system parameters depend on demographic variables (see Pollak and Wales (1981)) or (ii) estimating a separate demand system for each demographic group. Option (ii) is also appropriate for tests of GARP. Note that the demographic groups can be as small as 20 households in tests of GARP whereas it would not be reasonable to estimate a demand system on 20 households. Even in a sample of size 20, there are 380 pairwise comparisons that could conceivably reject GARP. Therefore, it is possible to define more homogeneous demographic groups for tests of GARP than it is possible to use for estimating a parametric model.[8]

V. DATA

V. A. The Consumer Expenditure Interview Survey (CE)

The data used in this study were drawn from an internal Bureau of Labor Statistics demand analysis data set created by Julie Nelson in 1988. This data set combines annual household

---

[8] The approach in this paper to allow for preference heterogeneity is analogous to fully interacting all the parameters of a demand system with dummy variables for each of the 23 groups in a parametric test of demand theory.

expenditures and demographic characteristics from the 1982-85 Consumer Expenditure Interview Survey (CE) and intertemporal, interarea price indices for households in twenty-five major metropolitan areas.[9] Annual household expenditures were aggregated into nine categories: food at home, food away from home, household furnishings, medical care, personal care, transportation, apparel, housing, and "other". Intertemporal-interarea price indices for all goods except "other"[10] were created for good *k* in period *t* for metropolitan area *z* as:

$$P_{zt}^{k} = \left(\frac{CPI_{zt}^{k}}{CPI_{z0}^{k}}\right) SI_{z0}^{k}$$

For *k* = 1,..., 8 goods; *z* = 1,.., 25 areas; and *t* = the 12 month period the family was in the survey. CPI^{k}_{zt} is the average annual consumer price index for the 12 month survey period (reference year = autumn 1973) for good *k* in city *z* in time period *t*. The CPI prices a common basket of goods over time for area z and so CPI^{k}_{zt} / CPI^{k}_{z0} measures the intertemporal price change within area z. However, CPI^{k}_{zt} /CPI^{k}_{z0} does not measure price differences across areas because there is not a common basket of goods across areas. Sherwood(1975) developed an interarea price index, SI^{k}_{z0}, for good *k* in city *z* at period *0* (base year = 1973) by pricing a common basket of goods across areas at a point in time. By multiplying the two components of price variation together, I obtain intertemporal-interarea price indices. The sample was limited to families in the 25 metropolitan areas for which the interarea-intertemporal price indices could be computed. The household reports nominal expenditures. Quantity indices are the ratio of expenditures to the intertemporal-interarea price indices.

---

[9] The CE collects data for “consumer units” defined as a group of individuals who share expenditures on two out of the following three categories: food, housing, and all other. I use the term “household” instead of consumer unit.

[10] The price of "other" was computed as the weighted average sum of the city CPI changes, rebased to 1973, for the four CPI indices which essentially comprise the "other" category in the Nelson data set (entertainment, tobacco, alcohol, and other rental costs) and multiplied by the Sherwood index.

In the fifth interview, the household reports annual earnings, weeks worked per year, and hours worked per week contemporaneous with the expenditures data. Annual hours worked is the product of weeks worked per year and hours worked per week. The gross hourly wage for both the reference person and spouse (where relevant) is the ratio of annual earnings to annual hours worked. For a family to be in the sample: (1) the reference person must be employed (or the spouse in married households), (2) earnings, weeks, and hours must be reported (3) no more than 5% of labor earnings can be derived from self-employment, and (4) hourly wages must be between $1.50 and $65.00 per hour. This resulted in a total sample of 1790 households.

Total taxes paid are computed for each household and an effective marginal tax rate is developed for each employed person. Gross pre-tax income is based on the labor earnings of the taxpayer and spouse (where relevant). Married families were assumed to file jointly and to take the standard deductions. Family size determined the deductions for dependents. The tax program incorporated the parameters of the Federal, State, and City tax codes as well as payroll taxes for each of the four years, 18 states, and 25 cities in the analysis for married and single persons.[11] The Stata programs I use to construct total taxes paid and the effective marginal tax rate are available in Technical Appendix I at http://www.ucsd.edu/%7Emfamular/.

The sample is divided into mutually exclusive demographic groups on the basis of the following eight characteristics; (1) whether the household is comprised of one-person or the household is a married couple (with or without children) (2) whether the reference person is white or nonwhite, (3) whether the reference person is young (under 30 years old), middle-aged (ages

---

[11] Although CE households are asked to report taxes paid, I use the tax program to calculate taxes paid because of missing data. Missing data also lead to the decision to exclude household reported nonlabor income from the tax calculations and from the GARP tests. In the tax program, I did not model the Earned Income Tax Credit (EITC). Since the EITC was a relatively small program prior to 1987, the omission of the EITC parameters is not likely to greatly affect the tax calculations. Finally, given the static nature of these tests of GARP, I do not take into account the present discounted value of the benefits of any taxes paid.

30 to 64), or old (65 years old and older), (4) whether the family rents or owns its home[12], (5) for one-person households, the sex of the individual, (6) for married households, the number of household members, (7) for married households, whether there is a child less than six years old, and (8) for married households, whether the husband is the sole earner or the wife works too (a dual earner household). I use a demographic type if there are at least 20 households of that type. There were 23 demographic types with at least 20 households. These 23 groups account for 70% of the households and are described in Table 1.

Also reported in Table 1 are average expenditures, earnings, estimated taxes paid, annual hours worked, constructed hourly wages, and marginal tax rates for each of the 23 groups separately and averaging across the seven groups of one-person households, the six groups of married households where the husband is employed and the wife is not, and the 10 groups of married households where both the husband and the wife are employed.

Table 2 characterizes the variation in commodity prices and quantities, hours worked, and after-tax hourly wages both within and across demographic groups. The after-tax hourly wage varies more than commodity prices. Except for housing, the standard deviation of most prices is around 10% of the mean price while the standard deviation in hourly wages is more than 50% of the mean wage. Price variation within a demographic group is almost as great as price variation across all households. The bottom panel of Table 2 presents the mean and standard deviation of each quantity index and hours of work. Except for housing (and to a lesser extent food at home), there is substantially more variation in consumption quantities than in hours worked. Columns three through five present within-group standard deviations in consumption quantities and hours. Column four also controls for consumption prices while column 5 also controls for consumption prices, wages, taxes, and net savings (expenditures minus after tax earnings).[13] There is sizable

---

[12] Since the housing price and expenditure data for homeowners is based on rental equivalence, these data may be subject to more measurement error than the housing data for renters. As a result, homeowners and renters are compared separately in the analysis.

[13] The reported within-group standard deviation is the root mean squared error from a regression that includes 23 demographic group dummy variables and other explanatory variables as noted in the text.

variation in commodity choices and hours worked even controlling for demographic group, prices, wages, taxes, and net savings.

V. B. Hours and Earnings in the Consumer Expenditure and Current Population Surveys

The primary focus of the Current Expenditure Survey (CE) is to collect household consumption expenditures to provide the items and weights for the basket of goods in the Consumer Price Index. To obtain contemporaneous annual expenditures and earnings data, CE households had to respond to all five quarterly surveys to be in this study.[14] How does information on annual earnings, weeks worked per year, and hours worked per week reported by CE households compare to data from the Current Population Survey (CPS), a frequently analyzed source of labor market data? I obtained a comparable sample of individuals from the March CPS for 1982-85. There are 78,913 employed men and women between the ages of 18 and 64, in the same cities as in the CE sample, who report annual earnings, weeks worked per year, hours worked per week and occupation.

As a fraction of the CPS sample means, workers in the CE earn 18.6% more per year, work 4.7% more hours per year, and are paid 22.9% more per hour than workers in the CPS. These differences are due in part to significant differences in demographic characteristics across samples: CE workers have 0.4 year more education, are 3.37 years older, are 3.4 percentage points less likely to be black, and are more likely to be married (15.8 percentage points). In the CE data, the number of households per month is similar. However, the CE data used in this study start in September, 1982 and end in November, 1985. Therefore, only 9.5% of the CE sample is from 1982, when the economy was in a recession (the NBER dates November 1982 as the trough), whereas 24.9% of the CPS sample is from 1982. Using a standard decomposition, I find that differences in observed demographic, location and job characteristics across the two samples

[14] Reyes-Morales finds that complete responders are older (50.6 versus 40.9), have higher quarterly expenditures ($8,981 versus $7,504) and are more likely to be homeowners (73.2% versus 41.0%) and married (57.2% versus 39.8%) than incompletely responding households

explain most of these mean differences. A detailed comparison of the CE and CPS samples is available in Technical Appendix 2 at http://www.ucsd.edu/%7Emfamular/.

VI. RESULTS

VI. A. Tests of GARP: Nine Commodities and Leisure

Table 3 presents the total number of pairwise comparisons that violate and are consistent with GARP under the two parameterizations of the budget constraint. For tests of GARP using DRP(mt), the minimum violation rate is 1.5%, the maximum is 41.3% and the median is 14.1% across the 23 household groups. As a fraction of all pairwise comparisons, i.e., making no attempt to increase the power of these tests of GARP, these violation rates are high. For only 6 of the 23 demographic groups is the violation rate less than 5% across all pairwise comparison

The number of violations falls for all demographic groups one (where there is a substantial increase in the number of violations) when using the tax schedule. In the median household group, 21.6% of GARP violations are removed when using the tax schedule budget constraint. The minimum violation rate is 0.8%, the maximum is 32.3% and the median is 13.2% across the 23 groups. Even using the tax schedule budget constraint does not lead to demographic groups whose choices are consistent with maximization of a common utility function: 15 of the 23 household groups have violation rates across all pairwise comparisons that exceed 5%.

Proceeding to estimating a parametric, static utility model with these data would not be reasonable given the high fraction of the time that household choices and budget constraints violate the fundamental predictions of consumer theory. This conclusion is strengthened when I calculate the violation rate among pairwise comparisons with similar incomes. Table 4 presents violation rates among those pairwise comparisons where the normalized income gap, measured by expression (6), is less than .20.[15] Comparing column 1 in Table 3 with column 1 in Table 4,

---

[15] In these tests of GARP there are a total of 84,016 pairwise comparisons. The median comparison involves households whose normalized income gap is .365. Almost 29% of all household comparisons have a normalized income gap is less than 0.20.

29.4% of all pairwise comparisons compared household with similar incomes. Across the 23 demographic groups, the median violation rate among households with similar incomes is 31.0% using DRP(mt) and is 23.3% using DRP(tb). These GARP results suggest that any parametric static labor supply model is likely to be rejected by the data.

VI.B. Are the Violations of GARP Trivial? Afriat's *e*

We may be less concerned about small optimization errors on the part of the consumer because such errors may not have a substantial impact on the ability of the simple static model to predict behavior. In addition, measurement error can confound nonparametric analysis when a small difference in one of the measured quantities, prices, wages, or hours worked in a pair of consumption-hours bundles changes the preference orderings. Following an approach proposed by Afriat (1967), it is straightforward to modify the DRP relationship so that the comparison of "similar" choices provides no information about preferences. Let *i* be DRP(tb) to *j* if and only if:

$$P_i(X_i\text{-}X_j)+W_i(H_j\text{-}H_i)+(TAXBILL_i - TAXBILL_{ij}) >= 0 \textit{ and } i \text{ and } j \text{ are not similar.}$$

Although a DRP relationship will not be established for pairwise comparisons which are similar, an RP relationship may still be established. Following Varian (1993), *i* and *j* are considered too similar to use the *direct* revealed preference information obtained from their comparison if the normalized income gap is less than .05.[16] Doing so eliminates the DRP relationships for 7.5% of the pairwise comparisons in these data. As seen in Table 5, removing these DRP(tb) relationships

---

[16] Consider two bundles with similar incomes but where the quantities chosen are quite different. Are these bundles similar? If one is allowing for small optimization errors and if commodities are substitutable, then the consumer may actually consider these two bundles to be similar. As a result, I follow Varian and determine similarity on incomes alone. A problem with this approach is that it is precisely those commodity bundle comparisons with similar incomes that provide the most powerful tests of GARP and so Afriat's *e* will reduce the power of these tests of GARP.

reduces the number of violations of GARP. The overall violation rate for 11 of the 23 demographic groups is now less than 5%.

This low violation rate is misleading because many of the GARP-consistent pairwise comparisons are between individuals with large income differences where violations are unlikely to occur. When I calculate the violation rate among pairwise comparisons where the normalized income gap is less than 20%, the median violation rate across the 23 household groups is 12.7% and only 6 groups have violation rates less than 5%. Thus, these results do not support the hypothesis that the violations of GARP in these data are trivial.

VI.C. Are the Violations of GARP Due to a Few Highly Inconsistent Households?

The violation rate may be high even though there are only a few households with heterogeneous preferences who contribute to most of the violations.[17] If the removal of a few households with different preferences results in a set of data largely consistent with GARP, then it may be reasonable to remove the "bad violators" from the data and proceed to estimate a parametric labor supply model.

Of the 1246 household in the analysis, there are 48 households involved in five or more violations of WARP. These 48 households were removed from the sample and then tests of GARP using DRP(tb) were performed. Table 6 presents these results. Violation rates among household comparisons with similar incomes remain high across the 23 household groups ranging from a low of 3.0% to a high of 29.5%, with a median violation rate of 12.1%. Though 3.9% of the households in these data contribute to a substantial fraction of the violations of GARP, removal of these households does not lead to demographic groups whose choices are consistent with preference maximization and a common utility function.

---

[17] This is the main point of the maximal consistent subset approach advocated by Banker and Maindiratta (1988).

VI.E. Comparison with Tests of GARP on Commodity Choices Only

The inconsistency of households' budget constraints and commodity-hours of work choices in these data is in stark contrast to the results in Famulari (1995). In the previous paper, I tested for the consistency of commodity choices with GARP across 43 demographic groups, including the 23 used in this paper, and found consumption choices and budget constraints were generally consistent with the predictions of demand theory. Table 7 reports the violation rates for tests of GARP on consumption choices only for the 23 household groups used in this analysis. The violation rate in the median group is 0.7% and the maximum violation rate is 2.0%. There are three household groups where there are no violations of GARP.

To increase the power of the tests, I present the violation rate among the household comparisons where the normalized income gap is less than .20. Among pairwise comparisons with similar incomes, the violation rate for all 23 demographic groups is below 5%: the violation rate in the median group is 1.2% and the maximum violation rate is 4.2%. There are four household groups where there are no violations of GARP. Thus, the hypothesis of common preferences and utility maximization works quite well for commodity choices alone: the source of the GARP violations is the labor supply decision.

VI.D. The Effect of Measurement Error in Hours Worked on Tests of GARP

It is well known that measurement error in hours worked will lead to a spurious negative correlation between hours worked and wages when the wage is measured as earnings divided by hours worked.[18] This division bias also leads to violations of GARP. Figure 1 illustrates the budget constraints consumption-hours choices of two individuals, *i* and *j*. Note that the consumption-hours choices of these two individuals violate GARP and that no common utility function could explain these choices.

[18] See, for example, Borjas (1980) and Eklof and Sacklen (1999)

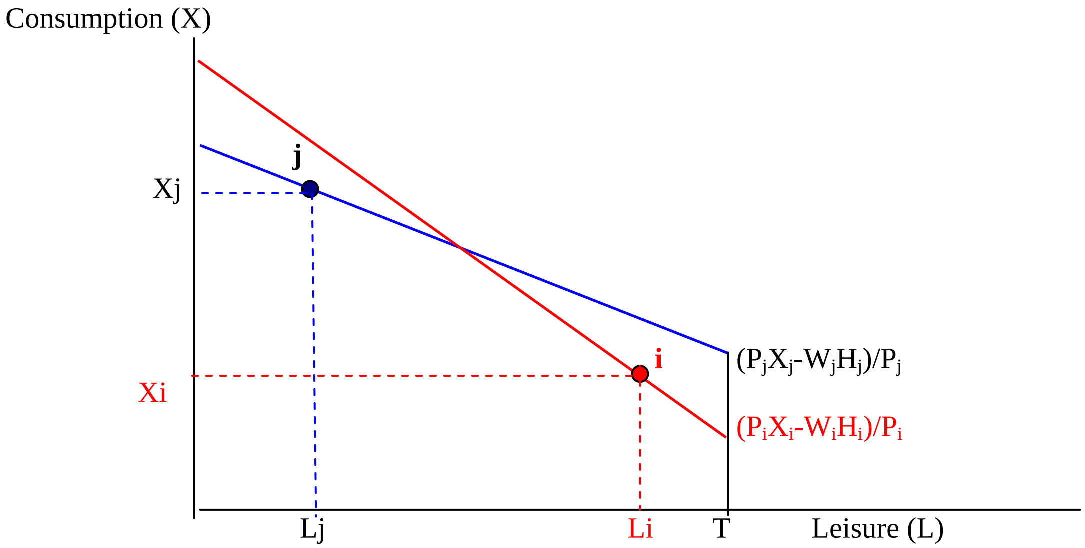

Figure 1: A Violation of GARP

What type of pairwise comparisons lead to GARP violations? Individual *i* has a high real wage, consumes a lot of leisure (works few hours), and consumes few other goods while individual *j* has a low wage but consumes little leisure (works many hours) and consumes many other goods. Note that when the hourly wage is constructed as earnings divided by hours worked, Figure 1 may simply result from measurement error in hours. If worker *j* overstates her actual hours worked, then her constructed hourly wage will be low, and low wages will be associated with few hours of leisure. If worker *i* understates her actual hours worked, then her constructed hourly wage will be high and high wages will be associated with many hours of leisure. Measurement error in hours, and the resulting bias in the hourly wage measured as earnings divided by hours, leads to a fundamental rejection of the static labor supply model.

In this section I simulate the effect of measurement error in hours worked on tests of GARP in two ways: when *only* hours worked are affected and when *both* hours worked and the hourly wage are affected due to constructing the wage as earnings divided by the mis-measured hours worked. Bound, Brown, Duncan and Rogers (1994) compare worker reports of annual hours with employer records of annual hours for the hourly wage earners at one large manufacturing firm.

The difference between log worker-reported hours and log establishment-reported hours is approximately normally distributed with mean .015 and standard deviation .102. How would measurement error of this magnitude affect tests of GARP? I use the actual budget constraints from the92 individuals in the single, white, male, renter demographic group. Given the wage, commodity prices, and nonlabor income (expenditures minus earnings) for each individual, I use a Stone-Geary utility function to determine utility maximizing commodity and hours of work bundles.[19] These commodity and hours choices and budget constraints, of course, pass GARP.

First, I randomly add measurement error to the Stone-Geary hours of work quantity according to the distribution of measurement error found in Bound et. al. (1994) and retest GARP. In 2000 simulations, I find that the median violation is 0.4% overall pairwise comparisons and rises to 2.0% among pairwise comparisons with similar incomes. Next I introduce division bias by calculating the hourly wage as the ratio of earnings to the hours measured with error. I find that the median violation rate across 2,000 simulations is 1.1% overall pairwise comparisons and rises to 4.9% among pairwise comparisons with similar incomes. This violation rate is larger, but far less than the violation rate found for the CE households.

Recall that Bound et al. (1994) report the measurement error in a worker's annual hours for the hourly wage earners employed in one firm. In their paper, the mean and standard deviation of log 1982 worker reported annual hours is 7.434 (.224) for the hourly workers in their sample. The mean and standard deviation of log 1982 annual hours is 7.402 (.665) in the CE sample and is 7.268 (.789) in a comparable CPS sample. The much larger variation in the log of annual hours worked is not surprising since the CE and CPS data are for workers who are both paid a salary and paid by the hour, both union and non-union, and from many different firms. Does this mean

---

[19] I chose the parameters of the Stone-Geary (the subsistence quantities and the beta coefficients) so that the average utility maximizing quantities were the same as the average observed quantities for the demographic group. For one individual these parameters led to negative 127 hours worked. I dropped this observation and conducted the simulations are based on the remaining 91 individuals.

that measurement error in annual hours is going to be two or three times as great? I do not know of a validation study which tells us the answer. Nevertheless, I doubled the standard deviation of the log measurement error and again simulated the effect of measurement error in hours alone and then allowed measurement error to also impact wages.[20] In 2000 replications, when there is just measurement error in annual hours the median violation rate is 1.3% overall pairwise comparisons and rises to 6.1% among pairwise comparisons with similar incomes. When I introduce division bias, by calculating the hourly wage as earnings divided by mismeasured hours, the median violation rate across 2000 replications, is 4.6% overall pairwise comparisons and rises to 18.9% among pairwise comparisons with similar incomes.

Therefore, if I allow enough measurement error in hours worked *and* determine the hourly wage by dividing earnings by these mismeasured hours, I obtain GARP violation rates that are similar to what I find in the 23 household groups. This suggests that errors in reported hours, combined with wages estimated as the ratio of earnings and hours, may explain the GARP violations found in the CE data.[21]

If the failure of these GARP tests is error in reported hours, combined with wages estimated as the ratio of earnings and hours, will using an instrumental variable for the wage lead to data that is consistent with revealed preference axioms? I estimate a log wage regression on a comparable sample of workers from the CPS.[22] I use the log wage coefficients from the CPS

---

[20] In terms of hours, how large is double the Bound et. al (1994) estimate? Averaging across 2000 simulations, the standard deviation of the error in annual hours was 458 hours.

[21] Division bias may also explain why GARP tests on commodity choices alone, using both aggregate time series data and cross section household data, are generally *consistent* with GARP. Researchers have long been concerned that the consistency with GARP has simply been the result of low power of the tests of GARP. Previous research has focused on the lack of budget set intersections as the source of the low power. However, division bias may be the source of the low power in tests of commodity demands. Researchers often have commodity expenditures and prices. Quantities are calculated as the ratio of expenditures and prices. Measurement error in prices will induce a spurious negative correlation between prices and quantities which reduces the likelihood of violating GARP.

[22] The independent variables in the log wage regressions are age, age squared, education, education squared, an education and age interaction, married, black and black interactions with all the demographic variables, “other race” and “other” interactions with all the demographic variables, 3 year, 13 industry, 8

comparison sample, to predict log hourly wages for the individuals in the CE used in this simulation. I again randomly add measurement error (with twice the standard deviation estimated by Bound et. al.) to the Stone-Geary hours of work quantity, use the CPS predicted wage and retest GARP. In 2,000 simulations, I find the median violation rate is 1.6% across all pairwise comparisons and 8.1% among pairwise comparisons with similar incomes. Therefore, using predicted wages greatly improves consistency with revealed preference axioms.[23]

There is one final issue of tests of GARP in the presence of measurement error that I must address: how likely am I to make a type II error? Using the Bound et. al (1994) estimate of measurement error and the CPS predicted wage, I find that 0.2% of the 2000 simulations have an overall violation rate that exceeds 5%. However, if the violation rate is calculated among pairwise comparisons with similar incomes, then 47.1% of the simulations have violation rates that exceed 5%. If measurement error is twice the Bound et. al (1994) estimate, I find that 4.5% of the 2000 simulations have an overall violation rate that exceeds 5%. However, if the violation rate is calculated among pairwise comparisons with similar incomes, then 82.8% of the simulations have violation rates that exceed 5%. While pairwise comparisons with similar incomes provide the most powerful tests of GARP, these comparisons are also the most affected by measurement error. As a result, in the remaining tests of GARP, I focus on GARP violation rates across *all* pairwise comparisons.

occupation, and 24 city dichotomous variables. Since I have the same demographic variables for the CE workers, I predict log wages for 92 individuals from the single, white, male, renter demographic group. The predicted wage equals $e^{(\log \text{wage} + .5(\text{MSE}^2))}$ where MSE is the mean squared error from the log wage regression.

[23] The compensated wage effect is often estimated to be larger for women than for men. These tests of GARP using predicted wages suggest that the different findings for men and women may result in part from the different wage measure used. To address the fact that a large fraction of women do not work, predicted wages are more often used when estimating labor supply regressions for women than for men.

VI.E. Tests of GARP for the 23 demographic groups using predicted CPS hourly wages

I now return to the 23 demographic groups from the Consumer Expenditure Survey and test whether using a wage instrument improves consistency of these household's choices with GARP. I obtain predicted hourly wages for each worker in the sample using the same approach as I used in the simulations: I use the worker's demographic characteristics and the log wage regression coefficients estimated for a comparable sample of CPS workers. In the tests of GARP I continue to use the household's reported annual hours worked.[24] Table 8 presents the tests of GARP using the CPS predicted wage. The costs of using the linearized budget constraint are higher when division bias is no longer present in the data. For the median household group, 50% of GARP violations found when using the linearized budget constraint, DRP(mt),are removed when using the tax schedule budget constraint, DRP(tb). When using the CPS predicted wage, 10 of the 23 household groups have DRP(mt) violation rates less than 5% and the median violation rate is 5.1% across all pairwise comparisons. When using the CPS predicted wage, 17 of the 23 household groups have DRP(tb) violation rates less than 5% and the median violation rate is 1.6% across all pairwise comparisons.

Comparison with Table 3, where the hourly wage was constructed as earrings divided by hours, shows the effects of using a wage instrument on tests of GARP. I focus on the tests of GARP using the tax schedule budget constraint, presented in the last three columns of Tables 3 and 8, to examine the effect of using the wage instrument. There is a substantial decline in the number of GARP violations when using the predicted CPS hourly wage: the number of GARP violations falls 69.1% for the median demographic group. Only 8 of the 23 demographic groups had DRP(tb) violation rates less than 5% and the median violation rate overall was 13.2% when

[24] A limitation of this approach is that the measure of household earnings used in the tests of GARP is no longer equal to the household's reported earnings and we know that households reported earnings fairly accurately (Bound and Krueger (1991) and Bound et.al. (1994)). Though perhaps tempting, I can not

GARP was tested using earnings divided by hours as the measure of hourly wages. When using the CPS predicted wage, When using the CPS wage and DRP(tb), the violate rate ranges from 0 (and so the demographic group passes GARP) to a high of 21.1%. The median violation rate is 1.6% across the 23 demographic groups and 17 of the 23 household groups have DRP(tb) violation rates less than 5%. Proceeding to estimate a parametric model with these data now seems reasonable.

VI. CONCLUSIONS

The piecewise-linear and nonconvex budget constraints implied by the U.S. tax system and transfer programs lead to endogenous wages which makes estimating labor supply models difficult. As a result, statistical inference in econometric models is conditional on numerous restrictive assumptions that may bias parametric tests of standard labor supply theory. The axioms of revealed preference, GARP and WARP, test whether a model of household optimization could rationalize the data without imposing restrictive assumptions on the functional forms for preferences, error terms, or the budget constraint.

The commodity-hours choices and budget constraints of similar households in this cross section frequently violate the axioms of revealed preference, even when the tax system is carefully modeled in the budget constraint. These results contrast sharply with these household's commodity choices and budget constraints which are generally consistent with homogeneous preferences and preference maximization. This suggests that parametric models of static labor supply are doomed to fail for these data.

One possibility is that the static labor supply model is an inadequate description of consumption-hours worked choices, and that dynamic models are needed to understand labor supply decisions. This paper demonstrates, however, that the failure of the static model may be

---

construct an annual hours worked measure by dividing household reported earnings by the CPS hourly earnings measure without re-introducing division bias into these tests of GARP.

due to measurement error in hours worked and calculating wages as earnings divided by hours worked. While measurement error in hours worked has long been known to introduce a spurious negative correlation between hours worked and wages, calculated as earnings divided by hours, this paper clarifies that it also leads to violations of GARP. When using a wage instrument and allowing for the complicated effects that taxes have on household budget constraints, I find that demographically similar household's consumption and leisure choices are reasonably consistent with the joint hypotheses of utility maximization and common preferences. Though not possible with these data, the GARP tests suggest that more accurately modeling the household's budget constraint by, for example, including time and money costs of work are likely to result in even greater consistency with the static utility model.

TABLE 1: Summary Characteristics of the 23 Demographic Groups

I. ONE-PERSON IN HOUSEHOLD:

Group	# Obs	Home	Age	Sex	White	Expenditures	Pre-tax Earnings	Tax bill	Annual Hours	Pre-tax Hourly Wage	Marginal Tax
1	71	Own	Midage	F	Yes	$18,947	$21,650	$5,922	1964	$11.96	.35
2	66	Rent	Young	F	Yes	$10,843	$14,814	$3,102	1963	$ 7.48	.29
3	86	Rent	Midage	F	Yes	$12,424	$19,322	$4,871	2043	$ 9.58	.36
4	76	Own	Midage	M	Yes	$18,927	$31,292	$9,833	2197	$14.23	.42
5	60	Rent	Young	M	Yes	$12,267	$15,326	$3,365	1815	$ 8.33	.29
6	92	Rent	Midage	M	Yes	$15,602	$24722	$7,252	2167	$11.49	.37
7	20	Rent	Midage	M	No	$ 9,104	$16,591	$3,905	1946	$ 8.54	.33
AVERAGE						$14,627	$21,403	$5,815	2036	$10.56	.35

II. MARRIED, WHITE, HOUSEHOLD WHERE ONLY HUSBAND IS EMPLOYED

Group	# Obs	Home	Age	Kids<6 yrs	Family Size	Expenditures	Pre-tax Earnings	Tax bill	Annual Hours	Pre-tax Hourly Wage	Marginal Tax
8	59	Own	Young	No	2	$22,717	$28,363	$7,203	1996	$15.10	.33
9	41	Own	Midage	No	3	$28,174	$33,456	$8,520	2168	$15.12	.35
10	55	Own	Midage	No	4	$30,622	$37,462	$9,882	2134	$17.31	.37
11	34	Own	Midage	Yes	4	$24,553	$32,523	$7,397	2307	$14.41	.34
12	33	Own	Midage	No	5	$28,569	$33,904	$7,967	2164	$15.51	.32
13	25	Own	Old	No	2	$23,065	$16,307	$3,471	1559	$ 9.89	.24
AVERAGE						$26,453	$31,327	$7,769	2076	$15.03	.33

III. MARRIED, WHITE, HOUSEHOLD, DUAL EARNERS

Group	# Obs	Home	Age	Kids<6 yrs	Family Size	Expenditures	Pre-tax Earnings	Tax bill	Annual Hours	Pre-tax Hourly Wage	Marginal Tax
14	38	Own	Young	No	2	$27,365	$44,542	$13,477	2186 (H) 2116 (W)	$11.46 (H) $ 9.03 (W)	.43 (H) .43 (W)
15	116	Own	Midage	No	2	$27,814	$46,931	$15,214	2104 (H) 1765 (W)	$13.35 (H) $ 9.92 (W)	.44 (H) .45 (W)
16	84	Own	Midage	No	3	$28,404	$40,730	$11,701	1942 (H) 1695 (W)	$12.72 (H) $ 8.64 (W)	.40 (H) .42 (W)
17	28	Own	Midage	Yes	3	$26,716	$41,471	$12,160	2083 (H) 1678 (W)	$11.61 (H) $10.45 (W)	.42 (H) .43 (W)
18	106	Own	Midage	No	4	$31,299	$43,523	$12,513	2139 (H) 1572 (W)	$14.04 (H) $ 7.95 (W)	.41 (H) .43 (W)
19	35	Own	Midage	Yes	4	$27,973	$44,097	$12,318	2259 (H) 1332 (W)	$14.33 (H) $11.36 (W)	.42 (H) .45 (W)
20	45	Own	Midage	No	5	$32,509	$41,120	$10,865	2107 (H) 1337 (W)	$14.36 (H) $ 8.83 (W)	.40 (H) .43 (W)
21	23	Own	Midage	Yes	5	$28,508	$37,138	$9,394	2273 (H) 1069 (W)	$12.16 (H) $ 7.56 (W)	.37 (H) .39 (W)
22	29	Rent	Young	No	2	$20,507	$29,552	$7,236	1884 (H) 1898 (W)	$ 8.59 (H) $ 6.95 (W)	.37 (H) .37 (W)
23	24	Rent	Midage	No	2	$20,913	$34,674	$9,432	2057 (H) 1526 (W)	$12.06 (H) $ 8.12 (W)	.36 (H) .36 (W)
AVERAGE						$28,243	$42,177	$12,309	2094 (H) 1631 (W)	$12.94 (H) $ 8.93 (W)	.41 (H) .43 (W)

TABLE 2: SUMMARY STATISTICS FOR PRICES, WAGES AND QUANTITIES

I. PRICES

Commodity Group	Mean	Standard Deviation. Overall	Standard Deviation Within-group
Medical Care	268.43	41.84	41.55
Transportation	244.59	17.77	17.73
Food Away	227.75	27.06	27.07
Housing	215.12	50.04	49.37
Personal Care	212.51	22.01	21.75
Other Goods	205.19	15.38	15.37
Household Furnish.	199.42	14.65	14.56
Food at Home	192.70	14.16	14.05
Apparel	154.42	14.58	14.51
Aftertax Wage 1	7.47	3.92	3.67
Aftertax Wage 2	4.94	2.96	2.95

II. QUANTITIES

			Within-group Standard Deviations		
Commodity Group	Mean	Overall Standard Deviation	Unconditional	Conditional on Prices	Conditional on Prices, Wages,Taxes, Net Saving
Medical Care	2.91	3.64	3.39	3.38	3.21
Transportation	21.32	21.56	20.47	20.39	15.26
Food Away	5.56	4.84	4.64	4.61	4.07
Housing	35.29	17.00	13.00	12.45	10.88
Personal Care	1.05	.93	.85	.83	.79
Other Goods	13.13	12.68	12.03	11.89	10.21
Household Furnish.	6.94	7.44	6.99	6.91	6.34
Food at Home	12.46	7.41	4.52	4.47	4.33
Apparel	9.50	9.08	8.63	8.49	7.39
Hours Worked 1	2068.45	611.15	600.93	600.54	498.87
Hours Worked 2	1631.03	745.10	714.47	714.52	599.00

TABLE 3: TESTS OF GARP FOR NINE COMMODITY GROUPS AND HOURS:
ALTERNATIVE TAX TREATMENTS

	TOTAL PAIRWISE COMPARISONS	**CONSTANT MARGINAL TAX: DRP(mt)**			**TAX SCHEDULE: DRP(tb)**		
		GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**
One-person Households							
1.	4970	4204	756	15.2%	4297	655	13.2%
2.	4290	4120	78	1.9%	4082	52	1.3%
3.	7310	6283	1013	13.9%	6310	956	13.2%
4.	5700	3342	2356	41.3%	4106	1588	27.9%
5.	3540	3151	347	9.9%	3355	119	3.4%
6.	8372	5617	2749	32.9%	5733	2633	31.5%
7.	380	368	6	1.6%	360	6	1.6%
Married Households, Husband Earner							
8.	3422	2693	729	21.3%	2747	659	19.3%
9.	1640	1608	26	1.6%	1603	13	0.8%
10.	2970	2495	467	15.8%	2722	240	8.1%
11.	1122	1080	16	1.5%	1086	12	1.1%
12.	1056	966	80	7.6%	966	78	7.5%
13.	600	562	18	3.1%	551	15	2.7%
Married Households, Dual Earner							
14.	1406	1090	304	21.8%	1125	271	19.4%
15.	13340	8324	5014	37.6%	9886	3448	25.9%
16.	6972	6246	720	10.3%	6514	416	6.0%
17.	756	641	105	14.1%	647	99	13.3%
18.	11130	8751	2367	21.3%	7513	3583	32.3%
19.	1190	955	235	19.7%	1029	161	13.5%
20.	1980	1453	505	25.8%	1558	396	20.3%
21.	812	772	36	4.5%	772	28	3.5%
22.	552	497	37	6.9%	509	21	4.0%
23.	506	399	107	21.1%	402	104	20.6%

Note: The number of pairwise comparisons where neither *i* RP *j* nor *j* RP *i* can be determined by subtracting the number of GARP consistent , C, and GARP violating, V, comparisons from the total number of pairwise comparisons.

TABLE 4: TESTS OF GARP FOR NINE COMMODITY GROUPS AND HOURS: FRACTION OF PAIRWISE COMPARISONS *WITH SIMILAR INCOMES* WHICH VIOLATE GARP: ALTERNATIVE TAX TREATMENTS:[1]

	PAIRWISE COMPARISONS With Similar Incomes	**CONSTANT MARGINAL TAX: DRP(mt)**			**TAX SCHEDULE: DRP(tb)**		
		GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**
One-person Households							
1.	1307	834	463	35.7%	890	400	31.0%
2.	1048	916	68	6.9%	900	44	4.7%
3.	2181	1460	708	32.7%	1486	653	30.5%
4.	1664	866	796	47.9%	980	678	40.9%
5.	766	498	232	31.8%	614	99	13.9%
6.	1965	1066	893	45.6%	1182	777	39.7%
7.	91	79	6	7.1%	72	6	7.7%
Married Households, Husband Earner							
8.	973	663	310	31.9%	655	311	32.2%
9.	383	352	26	6.9%	350	13	3.6%
10.	849	548	295	35.0%	638	205	24.3%
11.	360	320	14	4.2%	326	10	3.0%
12.	346	266	71	21.1%	267	69	20.5%
13.	178	145	16	9.9%	136	13	8.7%
Married Households, Dual Earner							
14.	529	375	143	27.6%	398	121	23.3%
15.	3976	2112	1862	46.9%	2407	1563	39.4%
16.	2005	1409	592	29.6%	1608	367	18.6%
17.	272	179	86	32.5%	185	80	30.2%
18.	3831	2537	1283	33.6%	2248	1554	40.9%
19.	374	270	104	27.8%	294	80	21.4%
20.	670	409	244	37.4%	447	202	31.1%
21.	235	199	32	13.9%	200	24	10.7%
22.	114	86	20	18.9%	94	9	8.7%
23.	210	145	65	31.0%	148	62	29.5%

[1] Household comparisons with similar incomes are defined to be those where normalized income gap, as defined in expression (6), is less than .2

TABLE 5: TESTS OF GARP FOR NINE CONSUMPTION GROUPS AND HOURS: TAX SCHEDULE BUDGET CONSTRAINT, DRP(tb), AND Afriat's E =.05

	All PAIRWISE COMPARISONS[2]			**COMARISONS WITH SIMILAR INCOMES**[3]		
	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**
One-person Households						
1.	4503	405	8.3%	999	247	24.7%
2.	3902	19	0.5%	732	11	1.5%
3.	6367	728	10.3%	1512	457	23.2%
4.	4275	1404	24.7%	1105	539	32.8%
5.	3337	63	1.9%	601	45	7.0%
6.	5874	2466	29.6%	1324	610	31.5%
7.	340	1	0.3%	53	1	1.9%
Married Households, Husband Earner						
8.	2909	448	13.3%	761	157	17.1%
9.	1536	3	0.2%	284	3	1.1%
10.	2779	66	2.3%	686	50	6.8%
11.	1032	6	0.6%	272	4	1.4%
12.	970	43	4.2%	270	35	11.5%
13.	547	5	0.9%	133	3	2.2%
Married Households, Dual Earner						
14.	1168	181	13.4%	399	74	18.5%
15.	10703	2533	19.3%	2812	1081	27.8%
16.	6572	204	3.0%	1657	177	9.7%
17.	666	68	9.3%	200	53	20.9%
18.	9074	1787	16.5%	2789	795	22.2%
19.	1045	125	10.7%	310	45	12.7%
20.	1636	270	14.2%	488	119	19.6%
21.	768	11	1.4%	195	8	3.9%
22.	490	19	3.7%	77	7	8.3%
23.	422	54	11.3%	151	29	16.1%

[2] The number of pairwise comparisons is the same as in Table 3.

[3] The number of pairwise comparisons is the same as in Table 4.

TABLE 6: REMOVING HOUSEHOLDS INVOLVED IN 5 OR MORE VIOLATIONS OF WARP. TESTS OF GARP ON REMAINING HOUSEHOLDS[4]

Demographic Group	Sample size	# Households Removed	Violation rate among Comparisons with similar incomes[5]
One-person Households			
1	71	2	13.5%
2	66	0	4.7%
3	86	5	6.6%
4	76	5	7.1%
5	60	0	13.9%
6	92	6	21.0%
7	20	0	7.7%
Married Households, Husband Earner			
8	59	2	17.1%
9	41	0	3.6%
10	55	2	15.3%
11	34	0	3.0%
12	33	1	6.0%
13	25	0	8.7%
Married Households, Dual Earner			
14	38	1	17.4%
15	116	11	9.9%
16	84	2	16.9%
17	28	1	13.1%
18	106	6	18.9%
19	35	1	13.1%
20	45	3	12.1%
21	29	0	10.7%
22	24	0	8.7%
23	23	0	29.5%

[4] Violation rates are calculated using the tax schedule budget constraint, DRP(tb) and among those household comparisons where normalized income gap, as defined in expression (5), is less than .2
[5] If no households are dropped, the violation rate is the same as in the final column of Table 4.

TABLE 7: TESTS OF GARP FOR 9 COMMODITY GROUPS ONLY

	All PAIRWISE COMPARISONS[1]			PAIRWISE COMARISONS WITH SIMILAR INCOMES[2]		
	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**
One-person Households						
1.	4910	16	0.3%	1271	10	0.8%
2.	4204	30	0.7%	1009	10	1.0%
3.	7095	145	2.0%	2054	85	4.0%
4.	5603	67	1.2%	1600	53	3.2%
5.	3466	36	1.0%	739	9	1.2%
6.	8214	106	1.3%	1892	62	3.2%
7.	364	2	0.5%	84	2	2.3%
Married Households, Husband Earner						
8.	3375	17	0.5%	946	7	0.7%
9.	1623	9	0.6%	374	4	1.1%
10.	2903	37	1.3%	792	35	4.2%
11.	1080	16	1.5%	340	7	2.0%
12.	1037	5	0.6%	329	4	1.2%
13.	590	0	0.0%	175	0	0.0%
Married Households, Dual Earner						
14.	1370	28	2.0%	504	18	3.4%
15.	13144	94	0.7%	3871	58	1.5%
16.	6853	81	1.2%	1941	37	1.9%
17.	744	0	0.0%	267	0	0.0%
18.	10932	162	1.5%	3670	132	3.5%
19.	1162	6	0.5%	357	2	0.6%
20.	1953	9	0.5%	653	7	1.1%
21.	804	0	0%	233	0	0.0%
22.	536	2	0.4%	103	0	0.0%
23.	500	4	0.8%	205	4	1.9%

1. The total number of pairwise comparisons is the same as in Table 3
2. The number of pairwise comparisons with similar incomes is the same as in Table 4.

TABLE 8: TESTS OF GARP USING PREDICTED CPS HOURLY WAGE AND HOUSEHOLD-REPORTED HOURS WORKED: ALTERNATIVE TAX TREATMENTS

	TOTAL PAIRWISE COMPARISONS	**CONSTANT MARGINAL TAX: DRP(mt)**			**TAX SCHEDULE: DRP(tb)**		
		GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**	GARP Consistent **(C)**	GARP Violation **(V)**	Violation Rate **V/(V+C)**
One-person Households							
1.	4970	4608	346	7.0%	4814	110	2.2%
2.	4290	3912	342	8.0%	4166	30	0.7%
3.	7310	6473	807	11.1%	6500	728	10.1%
4.	5700	4702	982	17.3%	4957	713	12.6%
5.	3540	3241	295	8.3%	3407	39	1.1%
6.	8372	5693	2641	31.7%	7464	814	9.8%
7.	380	346	12	3.2%	365	7	1.9%
Married Households, Husband Earner							
8.	3422	3235	175	5.1%	3330	50	1.5%
9.	1640	1618	4	0.2%	1610	0	0.0%
10.	2970	2777	157	5.4%	2837	95	3.2%
11.	1122	1106	4	0.4%	1104	6	0.5%
12.	1056	1025	13	1.3%	1029	11	1.1%
13.	600	562	16	2.8%	567	7	1.2%
Married Households, Dual Earner							
14.	1406	1334	56	4.0%	1358	18	1.3%
15.	13340	9993	3345	25.1%	10497	2807	21.1%
16.	6972	6455	509	7.3%	6825	111	1.6%
17.	756	716	20	2.7%	718	4	0.6%
18.	11130	9840	1288	11.6%	10180	922	8.3%
19.	1190	1139	39	3.3%	1157	15	1.3%
20.	1980	1865	101	5.1%	1871	89	4.5%
21.	812	711	99	12.2%	753	57	7.0%
22.	552	521	27	4.9%	518	18	3.4%
23.	506	486	8	1.6%	484	4	0.8%

REFERENCES

Afriat, S. N. (1967), "The Construction of Utility Functions for Expenditure Data," International Economic Review, 8, 67-77.

Banker, R. D. and A. Maindiratta (1988), "Nonparametric Analysis of Technical and Allocative Efficiencies in Production," Econometrica, 56, No. 6, 1315-1332.

Bloomquist, Soren (1996), "Estimation Methods for Male Labor Supply Functions: How to Take Account of Nonlinear Taxes," Journal of Econometrics, Vol 70, 383-405.

Blundell, Richard and Thomas MaCurdy (1999), "Labor Supply: A Review of Alternative Approaches." Handbook of Labor Economics, ed. by O. Ashenfelter and D. Card. New York: North-Holland.

Bound, John and Alan B. Krueger (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" Journal of Labor Economics, Vol. 9, No. 1, 1-24.

Bound, John, Charles Brown, Greg J. Duncan, Willard L. Rodgers (1994). "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data." Journal of Labor Economics, Vol. 12, No. 3, 345-368.

Burtless, Gary and Jerry A. Hausman (1978), "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment," Journal of Political Economy, Vol. 86, 1103-1130.

Columbino, Ugo and Daniela del Bocca (1990), "The Effect of Taxes on Labor Supply in Italy," Journal of Human Resources, 25, No. 3, 390-414.

Cox, James C. (1997), "On Testing Utility Hypothesis," The Economic Journal, 107, 1054-1078.

Eklof, Matias and Hans Sacklen (2000), "The Hausman-Macurdy Controversy: Why Do the Results Differ Across Studies?" Journal of Human Resources, Vol. 35, No. 1, 204-220.

Famulari, Melissa (1995), "A Household-Based, Nonparametric Test of Demand Theory," Review of Economics and Statistics, Vol. 2, No. 2, 371-382.

Heim, Bradley T. and Bruce D. Meyer (2003), "Structural Labor Supply Models when Budget Constraints are Nonlinear," Northwestern University Working Paper.

Heim, Bradley T. and Bruce D. Meyer (2004), "Work Costs and Nonconvex Preferences in the Estimation of Labor Supply Models," Journal of Public Economics, Vol. 88, 2323-2338.

Killingsworth, M. and J. Heckman (1986), "Female Labor Supply: A Survey," In Handbook of Labor Economics, ed. O. Ashenfelter and R. Layard. Amsterdam: North Holland, 103-204.

Macurdy, T., D. Green, and H. Paarsch (1990), "Assessing Empirical Approaches for Analyzing Taxes and Labor Supply," Journal of Human Resources, 25, No. 3, 415-489.

Moffitt, Robert (1990), "The Econometrics of Kinked Budget Constraints," Journal of Economic Perspectives, Vol. 4, No. 2, 119-39.

Pencavel, John (1986), "Labor Supply of Men: A Survey," In Handbook of Labor Economics, ed. O. Ashenfelter and R. Layard. Amsterdam: North Holland, 3-102.

Pollak, Robert A. and Terence J. Wales (1981) "Demographic Variables in Demand Analysis," Econometrica, 49, 1533-1551.

Reyes-Morales, Sally E. (2003), "Characteristics of Complete and Intermittent Responders in the Consumer Expenditure Quarterly Interview Survey," Consumer Expenditure Survey Anthology, Report 967, 25-29.

Sherwood, Mark K. (1975), "Family Budgets and Geographic Differences in Price Levels," Monthly Labor Review, 98, 8-15.

Triest, Robert K. (1990), "The Effect of Income Taxation on Labor Supply in the United States," Journal of Human Resources, 25, No. 3, 491-516.

Van Soest, Arthur, Isolde Woittiez and Ari Kapteyn (1990), "Labor Supply, Income Taxes, and Hours Restrictions in the Netherlands," Journal of Human Resources, 25, No. 3, 517-558.

Varian, Hal R. (1982) "The Non Parametric Approach to Demand Analysis," Econometrica, 50, 945-973.

Varian, Hal R. (1990), "Goodness-of-Fit in Optimizing Models," Journal of Econometrics, 46, 125-140.

Varian, Hal R. (1993), "Goodness-of-Fit for Revealed Preference Tests," Ann Arbor: University of Michigan mimeo.