Accepted Manuscript

Title: Big Data versus a Survey

Author: Stephan Whitaker

PII: S1062-9769(17)30017-0
DOI: http://dx.doi.org/doi:10.1016/j.qref.2017.07.011
Reference: QUAECO 1060

To appear in: *The Quarterly Review of Economics and Finance*

Received date: 12-1-2017
Revised date: 3-7-2017
Accepted date: 13-7-2017

Please cite this article as: Stephan Whitaker, Big Data versus a Survey, *<![CDATA[Quarterly Review of Economics and Finance]]>* (2017), http://dx.doi.org/10.1016/j.qref.2017.07.011

Highlights of "Big Data versus A Survey"

- I demonstrate the opportunities and challenges of substituting Big Data fora survey.
- I estimate models using credit bureau data and the Survey of Consumer Finances.
- Results are sensitive to adjustments made for population and variable definitions.
- Merging demographics into Big Data is effective in some instances, but not all.
- Findings could provide support or a caveat to many economists considering Big Data.

Big Data versus a Survey

Stephan Whitaker†
Research Economist, Research Department
Federal Reserve Bank of Cleveland
1455 East Sixth Street, Cleveland OH, 44114
216-579-2040 stephan.whitaker@clev.frb.org

July 3, 2017

Abstract

Economists are shifting resources from work on survey data to work involving "Big Data." This analysis is an empirical exploration of the trade-offs this substitution requires. Parallel models are estimated using Equifax credit bureau data and Survey of Consumer Finances data. After adjustments to account for different variable definitions and sampled populations, it is possible to arrive at similar models of total household debt. However, the estimates are sensitive to the adjustments. In this example, some external education and income measures are successfully integrated with the big data, but other external aggregates fail to adequately substitute for survey responses.

Keywords: Big Data; Survey Data; Household Debt

JEL Codes: C55, C81, D12

†The views expressed in this paper are those of the author and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Board of Governors of the Federal Reserve System.

Big Data versus a Survey

July 3, 2017

Abstract

Economists are shifting resources from work on survey data to work involving "Big Data." This analysis is an empirical exploration of the trade-offs this substitution requires. Parallel models are estimated using Equifax credit bureau data and Survey of Consumer Finances data. After adjustments to account for different variable definitions and sampled populations, it is possible to arrive at similar models of total household debt. However, the estimates are sensitive to the adjustments. In this example, some external education and income measures are successfully integrated with the big data, but other external aggregates fail to adequately substitute for survey responses.

Keywords: Big Data; Survey Data; Household Debt

JEL Codes: C55, C81, D12

1 Introduction

Economists appear to be rapidly shifting much of their research time and attention from work involving surveys to work involving "Big Data." In the process, there has been some discussion of the advantages and disadvantages of this transition, but little empirical exploration of the trade-offs (Einav and Levin, 2014; Cook, 2014; Sonka, 2014). This analysis will illuminate the discussion by estimating parallel models using data from the carefully designed, long-established Survey of Consumer Finances (SCF) and a sample from one of the oldest, most carefully maintained big-data data sets, the Equifax consumer credit records. The credit record sample is formally known as the Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP).

I estimate models of household debt using variables contained in both data sets as well as models with census-tract aggregate demographic data incorporated into the credit records. The SCF collects its own demographic measures. To maximize the chances of reaching comparable results, I take several steps to align the coverage and definitions in the two samples. Despite the adjustments, the corresponding coefficients in the models range from similar in magnitude and sign to starkly dissimilar. This example illustrates that while big data can offer frequencies and measures that surveys cannot match, we must be cautious about treating big data as a direct substitute for a carefully designed survey. If policy recommendations will hinge on the magnitude of parameters that econometricians estimate, then research based on big data could point in a different direction than research based on surveys.

Although the term "Big Data" has been applied in a variety of situations, some concepts are commonly associated with it. Big data is the byproduct of our daily activities. It has arisen with the automation of nearly all economic transactions and a major portion of our personal interactions. Every communication, payment, and trip is facilitated by computer systems and recorded. The resulting big data sets are updated frequently or continuously and their observation and variable counts are orders of magnitude larger than the surveys that researchers are accustomed to working with. The enormous size and complexity of big data sets in most cases requires that the data be stored in relational databases on multiple servers and accessed with structured query language (SQL) (Varian, 2014). In contrast, most survey and administrative data sets can be stored in a "flat" or "rectangular" format on a single personal computer and processed with any statistical software.

Big data is generally collected for nonresearch purposes but is often used for marketing research. Firms have an economic incentive to pay the fixed costs of data gathering and storage. Selling data to academic researchers could be a high-margin activity for big data

producers because the cost of data replication is relatively low. A disadvantage of big data is that the firms collecting it might not have the desire or ability to collect variables that would be of interest to academics and policymakers. For example, retailers might be unable to collect customers' demographic information if there is no reason to request it in the normal course of business. Surveys can pair questions about measures that researchers theorize are related. For example, a questionnaire could cover health and personal finances. In the world of big data, hospitals have detailed medical records and banks have records of balances and overdrafts. However, a researcher seeking to identify the impact of medical conditions on financial distress, for example, may be forbidden from linking the records. As with all human-subject data, privacy is a central concern. Even if consumers consent to their data being sold to researchers, firms face reputational risk. Privacy violations and identity theft could be extremely costly.

For certain types of questions, big data enables research that would be impossible with surveys. For example, in an hour of searching for a home online, an individual may conduct ten searches and click through fifty resulting links. Asking respondents to recall or log this activity manually would be burdensome and inaccurate. However, searches and clicks are recorded by search engines, internet service providers, and real estate listing sites, so the data are available for research. Before big data, we could observe home purchases, but not buyers' processes of searching, filtering, and evaluating.

In section 2, I will review the recent discussions of big data, research at the intersection of survey and administrative data, and articles about household debt. I will then give the relevant details about the data sets and adjustments made to increase comparability in section 3. Section 4 presents the results of the estimates and numerous alternate specifications. Section 5 concludes with a discussion of the implications of the analysis.

2 Literature

While there have been thousands of media reports and even a few books written about big data, there have only been a few articles in the academic economics literature that discuss the phenomenon directly (Eagle and Greene, 2014). The economist Francis Diebold maintains that he originated the phrase "Big Data" in a conference presentation in 2000 (2012). Einav and Levin discuss the use of big data for economic analysis (2014). They consider whether the predictive modeling tools that have been developed for use with big data can be applied to economic research. They note the opportunities as well as challenges related to obtaining the large data sets, which are usually proprietary. Hal Varian, chief economist for Google Inc., published an article introducing econometricians to the languages used to store and query big

data (2014). The same article includes suggestions for choosing variables for a useful model when hundreds or thousands of variables are available and enormous sample sizes make all coefficients significant by traditional criteria. Nickerson walks through the use of big data for contacting and mobilizing voters in recent national election cycles (2014). In the context of policy analysis, Cook argues that big data is useful for correlational exploration and prediction, but it may contribute less to identifying causal effects (2014). Similar discussions have been produced for other fields (Sonka, 2014; Pugh and Foster, 2014; Hazen et al., 2014).

The use of administrative data in social science research has a much longer history. Administrative data is also usually collected without considering how to make it useful for research. Unlike survey respondents, individuals represented by the data often do not know they are being studied. Administrative data observation counts are generally large relative to surveys, but demographic information is unavailable in many cases. The organizations that maintain administrative data and the individuals with records often have an economic incentive to maintain their accuracy. There is a substantial survey-validation literature in which researchers match survey respondents to administrative data. The authors then characterize the discrepancies between the records (for examples, see Warburton and Warburton (2004), Qiao (2005), Davies and Fisher (2009), Liegeois (2011), Dahl, DeLeire, and Schwabish (2011), Lynn, Jackle, Jenkins, and Sala (2012), Ansolabehere and Hersh (2012), and Czajka (2013)). Kapteyn and Ypma highlight the widely held assumption that administrative records represent true figures while survey responses contain all of the measurement error (2007). They test the sensitivity of published findings to this assumption and demonstrate the use of richer error structures. Abowd and Stinson propose a method that treats survey and administrative measures as noisy representations of the same true value (2013). They call for an explicit *a priori* assumption of a weighting to be used in a weighted-average estimate of the true value that combines the survey and administrative measures.

For the models of household debt that will be estimated in this analysis, I follow the literature on life-cycle patterns of household borrowing. The life-cycle hypotheses were first proposed in work by Modigliani and Brumberg (1954) and Friedman (1957). Since then, the relationships among current income, anticipated income, age, saving, and borrowing have been extensively studied. When researchers attempt to model household borrowing and test the impact of a novel factor, it is standard practice to include the borrowers' ages, marital status, children, income, and anticipated income. For recent examples, see Brown, Garino, and Taylor (2013), Schooley and Wordin (2010), and Tedula and Young (2005). In this analysis, age, income, and family structure will be incorporated to the extent that measures are available. Education serves as a rough proxy for anticipated income.

This analysis will build upon one publication in particular. Brown, Haughwout, Lee, and

van der Klaauw released a Federal Reserve Bank of New York Staff Report in which they compare household debts as measured in the CCP and the 2001-2010 SCFs (2015). They report the incidence of nonzero debt balances by type of debt as well as conditional means and medians. They compare implied national aggregate household debt levels and bankruptcy incidence. They plot the prevalence and conditional means of debts by the householder's age and the number of adults in the household. The analysis presented here takes steps beyond their descriptive statistics by merging the CCP with external data and estimating models using the CCP and SCF. Brown and her coauthors find that consumers and lenders report similar debt balances for secured debts including mortgages, home equity loans, and auto loans. However, substantial disparities appear in the unsecured debt categories of credit cards and student loans. The analysis in this paper will investigate whether the parallels between the prevalence and conditional means extend into the relationships between the debt balances and age, income, and family structure. In the cases of credit cards and student loans, I investigate whether similar estimated coefficients appear on right-hand-side variables despite the differences in the means and medians of the left-hand-side variables.

Merging variables from various sources and combining individual and aggregate measures are common practices in empirical research. Because these are central to the exercise presented here, I will briefly recall the issues related to merging in aggregate independent variables. In the simplest univariate regression, the use of aggregate variables can lead to unbiased estimates. Let i index individuals from 1 to n and let t index groups of these individuals, such as census tracts. y_i is the individual dependent variable observed in the big data. We want to estimate β as in

$$y_i = \beta x_i + \epsilon_i \tag{1}$$

If x_t is the mean of x_i in group t, and $x_i = x_t + u_i$, then u_i is uncorrelated with x_t. The estimated β is

$$\begin{aligned}
\hat{\beta} &= \frac{cov(y_i, x_t)}{var(x_t)} && (2)\\
&= \frac{cov(\beta x_i + \epsilon_i, x_t)}{var(x_t)} && (3)\\
&= \frac{cov(\beta(x_t + u_i) + \epsilon_i, x_t)}{var(x_t)} && (4)\\
&= \frac{cov(\beta x_t, x_t) + cov(\beta u_i, x_t) + cov(\epsilon_i, x_t)}{var(x_t)}. && (5)
\end{aligned}$$

If the aggregates (x_t) are also uncorrelated with the error term ϵ_i, only the variance of the

aggregates remains in the numerator and denominator. $\hat{\beta}$ is an unbiased estimate of β.

$$\text{plim } \hat{\beta} = \beta \frac{cov(x_t, x_t)}{var(x_t)} = \beta \tag{6}$$

The standard error of $\hat{\beta}$ will be larger than it would be if x_i observations were available because we are estimating the standard error with the mean squared residuals over the sum of the squared differences of the covariates from their mean. Σ represents $\sum_{i=1}^{n}$ below.

$$\hat{\sigma_\beta} = \frac{\frac{1}{n-2}\Sigma(y_i - \hat{y_i})^2}{\Sigma(x_t - \bar{x_i})^2} > \frac{\frac{1}{n-2}\Sigma(y_i - \hat{y_i})^2}{\Sigma(x_i - \bar{x_i})^2} \tag{7}$$

The variance of the aggregates will be less than the variance of the individual values. This could create a problem with precision, especially if the distributions within the groupings are similar to the distribution in the population. For example, gender measured at the individual level in a wage equation might be highly significant, but gender measured at a metropolitan or state level would have a very small variance and poor precision.

While merging in aggregate measures looks promising in the univariate case, there are several potential problems to keep in mind. First, using group aggregates other than the mean will not guarantee u_i is uncorrelated with x_t. Using the medians of skewed distributions, for example, will introduce a bias if the mean-median differences widen or narrow with x_t.[1] There is a literature following Openshaw and Taylor that demonstrates that regression estimates are sensitive to the units of aggregation (1979). For geographic groupings, this is referred to as the modifiable areal unit problem.

In most instances, we will be working in a multivariate regression context. Consider the case of two independent variables where w_i is measured individually and x_i is measured with the aggregates x_t. To simplify the notation, assume the variables are expressed as deviations from their mean. We are estimating the β_x and β_w in the equation:

$$y_i = \beta_w w_i + \beta_x x_i + \epsilon_i \tag{8}$$

If ϵ_i is not correlated with the other variables, we arrive at an estimate of β_x whose accuracy depends on the similarity of the variance and covariance properties of x_t and x_i.

$$\hat{\beta_x} = \frac{\Sigma w_i^2 \Sigma y_i x_t - \Sigma x_t w_i \Sigma y_i w_i}{\Sigma x_t^2 \Sigma w_i^2 - (\Sigma x_t w_i)^2} \tag{9}$$

[1] If the mean-median difference is the same at all levels of x_t, then the median is the mean plus a constant and is also uncorrelated with u_i.

$$= \frac{\Sigma w_i^2 \Sigma(\beta_w w_i + \beta_x x_i + \epsilon_i) x_t - \Sigma x_t w_i \Sigma(\beta_w w_i + \beta_x x_i + \epsilon_i) w_i}{\Sigma x_t^2 \Sigma w_i^2 - (\Sigma x_t w_i)^2} \tag{10}$$

$$\text{plim } \hat{\beta_x} = \beta_x \frac{\Sigma w_i^2 \Sigma x_i x_t - \Sigma x_t w_i \Sigma x_i w_i}{\Sigma x_t^2 \Sigma w_i^2 - (\Sigma x_t w_i)^2} \tag{11}$$

Likewise, the estimate of β_w includes a bias term that would only go to zero if $x_t = x_i$.

$$\hat{\beta_w} = \frac{\Sigma x_t^2 \Sigma y_i w_i - \Sigma w_i x_t \Sigma y_i x_t}{\Sigma w_i^2 \Sigma x_t^2 - (\Sigma w_i x_t)^2} \tag{12}$$

$$\text{plim } \hat{\beta_w} = \beta_w + \beta_x \frac{\Sigma x_t^2 \Sigma x_i w_i - \Sigma w_i x_t \Sigma x_i x_t}{\Sigma w_i^2 \Sigma x_t^2 - (\Sigma w_i x_t)^2} \tag{13}$$

To justify proceeding with the use of aggregate data to estimate β_x, one must argue that having an imperfect estimate of β_x is better than having no estimate at all. If β_w is the parameter of interest and x_t is a control, one must verify or assume that the bias remaining in β_w is smaller than the omitted variable bias that would exist if no measure of x was included. McCallum (1972) and Wickens (1972) demonstrate that using a predictive "proxy variable" always results in a smaller bias in β_w than omitting x entirely if u_i is uncorrelated with w_i. However, when using aggregates as proxies, u_i will inherit a correlation with w_i if x_i is correlated with w_i. Of course, even if the use of aggregate measures can be justified, all the standard challenges of econometric modeling remain, including the possibility of other omitted variables, measurement error, or misspecification.

3 Data and Descriptive Statistics

For complete descriptions of the data used in this analysis, readers should consult Lee and der Klaauw (2010), the SCF codebook, and the ACS documentation site.[2] I will discuss here the characteristics of the data most important to the modeling.

For decades, the SCF was the main source of information about Americans' household debts. The SCF has been conducted every three years since 1983. It is designed to be a nationally representative sample of households, and it records numerous measures related to the respondents' incomes, expenses, assets, and debts. The data is organized into "primary economic units," which exclude people living in the same household if they are financially independent from the primary respondent.

The analysis presented here is conducted entirely with the publicly available version of

[2]SCF Codebook: http://www.federalreserve.gov/econresdata/scf/files/codebk2013.txt. ACS Documentation: https://www.census.gov/programs-surveys/acs/technical-documentation.html. Accessed 28 June, 2017.

the SCF data. The publicly available data has several features intended to protect the privacy of the respondents. First, there is no geographic information in the records. This precludes merging in any aggregate measures, even at the metropolitan or state level. Second, the data are in a multiple imputation format. Each household is represented by five observations rather than a single observation. In instances where any variable had a missing value, the value has been replaced by five imputed values (Kennickell, 1998). Multiple imputation methods must be applied when estimating any descriptive statistics or models. These methods are well established, and corresponding routines are available in most statistics software packages (Rubin, 1987). In the course of the imputation, the Federal Reserve Board staff makes further changes to the data that are meant to prevent anyone from identifying the respondents. The exact details are not published to discourage attempts to reverse the anonymization. However, we might hypothesize that they would involve something like switching the income measures for two respondents. This would leave the mean, variance, and other moments of the income variable unchanged, but it would result in neither observation representing an identifiable person. For estimating correlations or regressions involving income and other variables, these swaps would also introduce measurement error proportional to the differences between the incomes.

The SCF asks respondents for the balance on their credit cards after their most recent payment. The question is designed to capture debt balances that are carried from month to month and accrue interest. The processed public data set reports zero balances for households that make "transactional" or "convenience" use of credit cards, meaning they pay the balance off in full each month. The dollar values involved in "transactional" uses may be small. However, whether or not these short-lived balances are included has a major impact on estimates of the percentage of households that use credit cards or use any credit product.

The CCP is a sample drawn from the Equifax credit bureau records. The Equifax database is "big" on several dimensions. It contains records on approximately 220 million individuals. It is updated continuously to reflect billions of payments to lenders and the consequent balance adjustments. The CCP sample is drawn quarterly, so twelve updates are available between SCF years. Because this analysis is attempting to mirror a cross-sectional survey, it does not delve into the never-before-possible research questions that are being explored with the CCP. Many of these opportunities arise from being able to observe the same borrowers immediately before and after an important event, or track individuals as they migrate within a metro area. This research is possible because the CCP contains an anonymous record ID that can create a quarterly panel of data for each borrower. Names, social security numbers, and street addresses are removed from the sample to guarantee

privacy and identity security.

The CCP sample contains outstanding balances by type of debt for approximately fifteen percent of all US residents with a credit history. The sample begins by randomly selecting pairs of digits and matching these to the last two digits of borrowers' social security numbers. When approximately five percent of the credit records have been selected, these records are designated "primary" observations, and every other individual with a credit record who is observed to be living at the same address is added to the sample. The sample for a specific quarter will contain approximately 13 million primary records and 29 million co-residents. Each observation has an indicator of the census 2000 block containing the current address of the credit record. The blocks map easily into Census 2010 tracts, which enables the merging of ACS estimates of income, education, and family demographics at the tract level. The data used here are as of the end of the second quarter of 2013.

While the CCP contains no information on the person's marital status, or even gender, we can infer something from co-residence. To prepare observations that are similar to what the SCF calls a "primary economic unit," I first drop co-resident observations that are over 15 years distant in age from the primary record. This should remove adult children and elderly parents who may have their own debts and income. If the children or parents happen to also be primary sample individuals, they are treated as a separate household. The records are then collapsed into households. Single households remain as they are, but married and cohabiting households become a single unit of observation with an indicator in their record that the household has two adults ($Couple = 1$). Roommates of similar ages are inadvertently treated as couples because there is no way to differentiate them. The SCF does exclude roommates because surveyers can ask respondents about their relationships. Approximately 3.2 percent of the CCP sample has to be discarded because anywhere from three to several hundred adults are reported as living at the same address. The larger head counts are probably apartments, condominiums, or other joint quarters where mail is delivered via individual names rather than unit numbers. In these cases, there is no way to identify which records represent single people or which adults are coupled.

To parallel the SCF, all CCP-reported balances secured by the primary residence are combined into the variable labeled *mortgage*. This includes closed-end home equity loans and the current balances on open-ended home equity lines of credit. Auto and student loans stand alone in both data sets. The variable *cards* includes balances on credit cards, retail cards, and miscellaneous other debts, such as rent-to-own durables contracts.

The CCP and SCF are drawn from different universes because the SCF seeks to be nationally representative while the CCP can only represent people with credit records. Unless models are conditional on debts being nonzero, the SCF representatives of nonusers of credit

will greatly influence the estimates. In the CCP, approximately 32 percent of singles' and eight percent of couples' records have zero total debt. We can assume that records that have no other debt except a credit card balance below $250 would appear as nonborrowers if they were sampled in the SCF. By this definition, 36 percent of singles and 11 percent of couples in the CCP are nonborrowers. Thirty-five percent of the singles in the SCF report no debts, and 23 percent of the couples report no debts. To balance the samples, I randomly drop 353 of the 853 nonborrower couples from the SCF sample. With this exclusion, the percentage of couples' households in the SCF declines from 58 to 55 percent. Only 51 percent of the households in the CCP are identified as couples. Some couples in the CCP sample might be misclassified as singles if they opt to keep all debts in one person's name.

The demographic controls used to augment the CCP are derived from the American Community Surveys conducted from 2008 to 2012. To reach a sufficient sample size for tract-level estimates, five years of observations must be aggregated. A number of different income values are available in the ACS estimates, and one set of models reported below (table 3) is estimated with five potential income measures.

The education measures are assigned to CCP households based on the household's tract and the age in the primary record. For example, if a 40-year-old is observed in a tract in which 33 percent of people aged 35-44 have a bachelor's degree, the variable $Bachelor's$ is assigned a value of 0.33 in her record. In the same tract, $Bachelor's$ may be set equal to 0.10 for a 70-year-old if that is the average undergraduate attainment for people in his age category in the tract. The percentage of households with children is assigned the same value for everyone in the same tract because no differentiation by age or number of adults in the household is available in the ACS aggregates. Obviously, there is a contrast between the continuous percentage values in the CCP-merged data and the binary values in the SCF data.

After the adjustments described above, table 1 illustrates how similar the samples become. In the CCP and adjusted SCF samples, 81 and 78 percent of households have some debt, respectively. However, the proportions having mortgage, card, and student debt are not equal. Thirty-eight percent of the households in the CCP have a mortgage or other home-secured debt, while 47 percent of SCF households have the equivalent debt. As mentioned above, the CCP does not distinguish between people who carry balances on their credit cards and those who pay off their entire balance each month. The SCF excludes the "transactional" use of credit cards. Thus the difference between the incidence of credit card debt in the CCP (72 percent) and the incidence in the SCF (49 percent) suggests approximately 23 percent of households make transactional use of cards. The incidence of student loans in the CCP is 71 percent of that in the SCF. This finding stands in contrast to the

finding reported in Brown et al. (2015). They were comparing household-level student loan reports in the 2010 SCF to individual-level reports in the CCP in 2010, and they estimated that SCF respondents underreported student loan debt by approximately 25 percent.

The second section of table 1 presents conditional means, standard deviations, and medians for each type of debt. On home-secured debt, the CCP and SCF are very closely aligned. The median home-secured debts differ by a trivial $16. The variances in the SCF data are higher for auto, credit card, and student debt, and they are driven by some extremely high observations in the SCF measures. Major disparities are evident in student loans, with the conditional median SCF-reported loans being over twice as high as the CCP-reported loans. This may be due to the SCF including dependent students in their parents' households.

Figure 1 (top left) represents the conditional (>0) distributions of total borrowing in the SCF and CCP. The CCP has more density below 9 log points, or approximately $8,100 of total debt. The distributions in both samples appear to be mixtures of two normal distributions. The higher distribution comprises households with mortgages, and the lower distribution is households without a mortgage. Looking at the distributions of debt for the subcategories in figure 1 reveals the source of the dissimilarity between the distributions of total borrowing. It appears to be concentrated in student loans. The CCP reflects a lower distribution of student loan debts, conditional on the debts being nonzero. The conditional distributions of mortgage debt, auto debt, and even credit card balances are very similar (the graph of auto debts can be found in the appendix). The congruent credit card distributions are surprising given that the SCF observations include only carried balances.

The descriptive statistics of the right-hand-side variables are presented in table 1. The challenge of identifying couples in the CCP can be seen in the difference between the shares of couples in the two samples. The SCF reports 55 percent of households having more than one adult, while the CCP reports 52 percent. The age distributions represented in the SCF and CCP are displayed in figure 2. The SCF sample includes adults from age 18 onward. Representation rises for ages between 19 and 30 as a larger fraction of each cohort has established its own households. In the CCP, young adults will be underrepresented because some fraction has not yet made its first reportable credit transaction. Regarding people 80 and older, they have a higher representation in the CCP. This could reflect that many families opt to continue paying debts owed by deceased family members if they do not see an advantage to changing the names on the accounts. In these cases, creditors would continue to report payments, and the deceased's records continue to be in the sample (Lee and der Klaauw, 2010). The SCF does not contain any records for the deceased. Also at very advanced ages people are more likely to become dependents in their adult children's households. In that case, the SCF would list the child's age as the primary respondent.

The various income measures from the ACS are summarized in table 1. The mean of the household incomes reported to the SCF is higher, at $85,214, than any of the means of the aggregate measures assigned to the CCP households. The tract median household income is approximately $7,000 higher than the median of the SCF household incomes. One Census Bureau-produced ACS table (Table B19215) provides median incomes by tract for subpopulations defined by combinations of male/female, living alone/not alone, and under 65/over 65 years of age. When household incomes are assigned to CCP households according to this table, the median ($48,630) is closer to the median of the SCF incomes ($46,668). The male and female figures are averaged before being assigned. The education measures linked to the CCP records reflect lower levels of education than those reported in the SCF. In particular, the SCF sample appears to contain more undergraduate degree holders and fewer people with some post-secondary education. Likewise, the tract-assigned percentage of households with children is centered around a value of 30 percent while 43 percent of the SCF households have children.

4 Results

4.1 Internal Data Models

As a first attempt to estimate the relationships between the household debt values and household demographics, I fit models with the two covariates available in both the SCF and CCP. The first models include only age and age squared. The coefficients on both are quite a bit larger in the SCF estimate, and the CCP coefficient would be outside a reasonable confidence interval around the former. When the *Couple* indicator is added to the models, the coefficients on this indicator appear quite similar, at 2.81 in the SCF estimate and 2.91 in the CCP estimate. The standard errors in the CCP results are much smaller, as we would expect given the sample sizes.

In the second pair of models in table 2, terms are introduced for age cubed, and all the age variables are interacted with *Couple*. Most of the newly introduced terms are not statistically significant in the SCF estimates, and they add little to either model. The R^2 terms increase only slightly. If the continuous age variables are replaced with categories and interacted with *Couple*, the models do not fit the data as well. Most of the coefficients in the categorical-age SCF estimate are very imprecisely measured and do not closely parallel those in the CCP estimates. The results of the categorical-age specification can be found in the appendix.

4.2 External Data Models

The models presented in table 3 incorporate the external income measures. In the SCF models, the observed household income is used while the CCP income measures are various tract-level aggregates merged in from ACS estimates. When income is introduced to the SCF model, the coefficient on *Couple* drops from 2.81 to 1.96. The first CCP model uses the log of the tract median household income. The coefficient on income in the SCF model is 0.90. The CCP coefficient on this tract median income measure is much higher at 1.77. Using tract per capita income rather than tract median household income returns similar results. While the overall income distribution is heavily skewed, within most census tracts it is approximately normal, especially after a log transformation. The correlation between the logged median and logged per capita values is above 0.95.

In the models with income assigned to the household by the age of the householder (table 3, column 5), we encounter a limitation to the usefulness of merging external data. In the CCP records we have a precise geography, an age, and a rough measure of the family structure. We might expect to increase the precision of our estimates by merging data using two or more of these characteristics. For example, the tract median income for households headed by people between 18 and 34 would have to be a more precise proxy for the income of a household headed by someone aged 25 than the overall tract median. However, if we are also including age in the model, merging age-subcategory measures within the geography leads to the merged measures' drawing explanatory power away from age. In the model with age-conditional income measures, the coefficients on Age and Age^2 fall. When a family-type conditional income measure replaces the age-conditional measure, the coefficient on age returns to its previous level and the coefficient on the family-type indicator ($Couple$) declines. None of these changes adds to the overall explanatory power of the model. Merging in the external data provides a point estimate for *income* and a corrected estimate for $Couple$. However, if our goal is prediction, the merged data did not provide new information in this case.

The final model in table 3 introduces the tracts' whole income distributions in the form of percentages of the households with incomes in ten categories. Despite the high collinearity among these groupings, due to neighborhood sorting on income, each measure is still highly significant. In the remainder of the analysis, the income measure used for the CCP models is the tract median income assigned by the household's type (Living alone/not alone, $<65/\geq 65$). Judging by the contrast with the SCF results, the other ACS income measures are not capturing the substantial income differences between one-earner and two-earner households.

Table 4 presents the SCF and CCP models when measures of education and children are introduced. The coefficients on the education measures in the SCF and CCP models

reflect parallel ordering and direction. Borrowing by an adult lacking a high school degree is estimated to be 1.27 log points lower than borrowing by high school graduates (the omitted category) in the SCF model. In the CCP model, the coefficient on the no-high-school-degree share in the tract is -1.60. People with some postsecondary education borrow more than high school graduates, and BA holders borrow more yet. Graduate degree holders' borrowing is higher than that of high school graduates, but lower than that of BA holders, conditional on other observables.

The coefficient on *Children* in the CCP estimate (1.44) is over twice as large as it is in the SCF estimate (0.61). The CCP model implies that living in a tract where a higher percentage of households have children is associated with more borrowing on the credit record. The SCF model suggests that having children in the home is associated with additional borrowing. The models agree on the direction, but the big data estimate seems to be capturing something fundamentally different with the geographic aggregates. Perhaps the strong tract-level relationship reflects the phenomenon of newly constructed subdivisions concentrating uniformly high-priced homes, recently originated high-balance mortgages, and families with young children. Tracts with older housing stock will include higher shares of empty-nest households (fewer children), lower-priced fixer-uppers, and paid-down mortgages.

4.3 Sample Sizes

Through the models discussed so far, we have seen that the standard errors estimated in the SCF models are generally two or more orders of magnitude larger than the standard errors in the CCP models. One could even argue that it is not critical to report standard errors or statistical significance in big data estimates like these because the standard errors are always minuscule. In the 11 CCP models discussed above, all but one coefficient was significant at the 0.001 level. Recall that the CCP is just a 5 percent sample of the data set maintained by Equifax. Researchers considering using big data may want to consider even smaller samples if the cost of the proprietary data is proportional to the sample size. The estimates in table 5 illustrate what difference we might expect if we could only afford a 1 percent, 0.1 percent or 0.01 percent sample. As the sample shrinks, the standard errors naturally grow. For the relationships internal to the data set (coefficients on *Age*, Age^2, and *Couple*) the errors remain small enough to easily identify statistical differences from zero. The standard errors rise more rapidly on the geographically merged education and children measures. Only when we reach the SCF-sized sample do the point estimates change appreciably. The models' R^2 values are also indistinguishable.

A survey like the SCF could cost several million dollars to conduct. If a researcher is only

interested in a few measures and some demographics, and a big data provider is available, purchasing 10,000 observations out of a 100,000,000-observation data set might be more feasible. If we draw a single sample out of the CCP that is the same size as the SCF, we can see that there is still sufficient power to identify most coefficients. However, it is concerning that some of the point estimates change. A 90 confidence interval on the coefficient for *Couple* would easily contain the coefficient estimated with the 5 percent sample. The same cannot be said for the coefficients on *Age* and Age^2. Was this just an idiosyncratic shift in this particular sample of N=5,634?

To explore this, I draw multiple small samples and repeat the estimation. I randomly group the 11,040,764 CCP observations into 1,840 samples of N=6,000. I estimate the model on each sample and plot the coefficients in figure 3. The coefficient of 0.20 on *Age* (table 5, column 6) appears to be a low draw from the distribution of coefficients (figure 3, top left). The coefficient on Age^2 (-0.22) is offsetting in that it is a high draw. The coefficient on *Income* of 0.43 also appears to have been a low draw. The 90 percent confidence intervals on the estimates from the SCF model contain the coefficients from the 5 percent sample CCP model for all variables except $Bachelor's$ and *Children*. This is a confirmation that the SCF is sufficiently large to estimate models of this type, even with the additional measurement error built into the publicly released data.

4.4 Sensitivity to Adjustments to Equate Coverage

As discussed in section 3, adjustments need to be made to account for the CCP's lack of observations representing nonborrowers. In table 6, there are four models that demonstrate how sensitive the results are to alternate adjustments to the data. The SCF estimates presented above were all estimated after dropping 353 nonborrower couples. If we estimate the model with the full SCF public data set, the coefficients on *Income* and *Children* increase somewhat. The coefficient on *Couple* declines from 2.03 to 1.27. This reflects that the distribution of couples' borrowing becomes more similar to that of singles when more nonborrowers are included because a higher share of singles are nonborrowers. The other coefficients are quite similar between the full sample and the adjusted sample.

If we are uncomfortable with dropping only some of the nonborrowers, we could estimate the models conditional on observing nonzero balances. The second and third models presented in table 6 are conditional on the total being nonzero. Very few of the coefficients are similar between the SCF and CCP models. This is due in part to the numerous low balances observed in the CCP for households whose only debt is transactional credit card balances. If we drop the 364,000 observations with no debts other than card balances below $250, the

coefficients of the model all shift toward those of the SCF conditional model. However, the coefficients estimated with this limited CCP sample still fall outside reasonable confidence intervals on the SCF conditional estimates. Overall, these alternate adjustments to the samples are not reassuring. They suggest that similarity in the models presented in table 1 is highly sensitive to the adjustments selected. Choosing other reasonable adjustments could leave us with divergent estimates.

4.5 Alternate Specifications

Using logs of the debt and income measures has the advantage of muting the influence of extreme observations, without having to determine which extreme observations to drop. However, the coefficients are then measured in log points or elasticities, which are not as intuitive or as easily conveyed to general audiences. Table 6 presents results if we return to dollar units. To avoid having a handful of observations exerting extreme leverage, I exclude all observations in either sample that have total debts or household income over $1,000,000. The coefficients on *Age* differ by approximately $1,000 in the levels models. They are not as similar as the coefficients in the log models. Near the mean of logged total debt outstanding (10.7-10.9), the difference between the coefficients in the log models (0.0066) corresponds to about $300. The CCP and SCF models agree that households with education beyond high school borrow more. However, the CCP level model returns a positive coefficient on the share of people in the records' tract and age category who do not have a high school degree. The measure of children, as in the log models, attributes more additional debt to the presence of children if the children are measured at the tract rather than household level.

Although the analysis thus far has combined all debts together, it should not be surprising that total household debt is dominated by mortgage debt. When separate models are estimated for each type of debt, as in table 7, the results for the mortgage debt model are similar to those of the total debt model. The coefficient on income in the SCF model is higher (1.20) in the mortgage model than in the total debt model (0.65). The coefficients on *Children* in the mortgage models are both much higher than in the total debt models, but they do not agree. In the models of nonmortgage debt, the estimates based on big data rarely agree with those based on the survey. The coefficients on *Age* and Age^2 are similar in the credit card and student loan models. Among the coefficients on the education and children's measures, huge differences in magnitude and differences in sign are more common than coefficients that agree.

For the analysis so far, we have opted to use the CCP data with individual observations and tract values assigned to all individuals living in a tract. How different would the results

be if we aggregated the CCP data to tracts before estimating the model? Estimates of this kind might be the only option if, for example, the big data provider is not willing to share individual observations but is willing to release tract-level aggregates. The seventh model presented in table 6 is estimated with the CCP data collapsed to the census tract. Only three of the coefficients in the tract-aggregate model are recognizable from the individual model. These three are coefficients on tract aggregate values for education levels at a bachelor's and below. The coefficients on Age, Age^2, $Couple$, $Income$, and $Children$ all make major shifts.

The results of two additional sets of specifications can be found in the appendix, and they are described here. Some of the literature on household debt prefers tobit specifications because zeroes are common in household debt data (Brown et al., 2013; Schooley and Worden, 2010). The zeroes could represent a preference for borrowing that is below zero. The household's true optimal borrowing bundle may not be available in incomplete markets. Alternately, the optimal bundle could be represented by savings data, but savings are not observed in the CCP. As mentioned above, multiple imputation routines are available for many types of econometric models, but the tobit model is not among them. Tobit models can be estimated on each of the five implicates separately. The results for the five SCF models are all very similar, and the coefficients are uniformly highly significant. Compared with the CCP tobit model, the SCF tobit models agree on most coefficients. The two points of disagreement are coefficients on $Bachelor's$ and $Children$.

The final alternate set of models explores the possibility of moving one of the geographically merged aggregate measures to the left-hand side. In the CCP models, the debts are divided by the tract median income assigned by household type (single/couple, <65/ >65), to create a debt-to-income (DTI) ratio. The log of the DTI is taken to reduce the influence of extreme values. Modeling the DTI ratios with the remaining explanatory variables results in a mixture of agreement and disagreement between the big-data-based and the survey-based estimates. The coefficients on Age and Age^2 appear similar, although not statistically indistinguishable. The other coefficients in the total debt models are comparable with the exception, again, of $Bachelor's$ and $Children$.

5 Conclusions

Through this example, we have learned that it is possible to arrive at similar model estimates using big data in place of a survey. However, this result is dependent on adjustments that must be made to one or the other data set to account for differences in the sampled universe and definitions of the variables. To arrive at similar model estimates using the CCP and SCF, one must first adjust for the CCP's lack of observation of people with no credit records.

Some nonborrowers need to be dropped from the SCF sample or added to the CCP sample. Also, the similarity in the models seems to be driven by the predictability of the largest category of debt, mortgages. Models of auto debt arrive at very different estimates using CCP data rather than SCF data even though the two sampled distributions of auto debt are very similar. Models of credit card and student loan debt show even more disparity.

While surveys usually collect demographic data and questions on multiple related topics, big data sets will only contain variables created for the data sets' original purposes. In the demonstration above, we see both the potential and limitations of merging in external data. The CCP data was augmented with ACS data by assigning tract-level measures according to location, age, and family structure. Estimates using the merged income and education data appear to do an adequate to good job of replicating individual observations. However, in the case of representing the influence of children on borrowing, the attempt is not successful. The prevalence of children in the borrower's tract seems to be representing something different than an indicator of children in the borrower's own household. The model coefficients are much higher on the tract-level measure.

In the coming years, we can anticipate repeated debates about the advisability of substituting purchased big data sets for survey data where possible. The analysis presented here gives arguments both for and against such a substitution. The massive sample sizes available in big data sets can provide levels of precision far beyond what is obtainable from a survey. We saw above that a single sample with several thousand observations can return model coefficient estimates that are substantially higher or lower than other samples of equal size. The survey researcher has no way to know where the current draw is relative to others. On the other hand, if big data research has to incorporate external data because key variables are not in the data set, then a parallel survey is essential as a benchmark.

References

Abowd, J. M. and M. H. Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics 95*(5), 1451 – 1467.

Ansolabehere, S. and E. Hersh (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis 20*(4), 437–459.

Brown, M., A. Haughwout, D. Lee, and W. v. der Klaauw (2015). Do we know what we owe? A comparison of borrower- and lender-reported consumer debt. *Federal Reserve Bank of New York Economic Policy Review 21*(1), 19–44.

Brown, S., G. Garino, and K. Taylor (2013). Household debt and attitudes toward risk. *Review of Income and Wealth 59*(2), 283 – 304.

Cook, T. D. (2014). 'Big Data' in research on social policy. *Journal of Policy Analysis and Management 33*(2), 544 – 547.

Czajka, J. L. (2013). Can administrative records be used to reduce nonresponse bias? *Annals of the American Academy of Political and Social Science 645*, 171 – 184.

Dahl, M., T. DeLeire, and J. A. Schwabish (2011). Estimates of year-to-year volatility in earnings and in household incomes from administrative, survey, and matched data. *Journal of Human Resources 46*(4), 750 – 774.

Davies, P. S. and T. L. Fisher (2009). Measurement issues associated with using survey data matched with administrative data from the Social Security Administration. *Social Security Bulletin 69*(2), 1 – 12.

Diebold, F. (2012). A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline, second version.

Eagle, N. and K. Greene (2014). *Reality Mining: Using Big Data to Engineer a Better World*. EBL-Schweitzer. MIT Press.

Einav, L. and J. Levin (2014). The data revolution and economic analysis. *Innovation Policy and the Economy 14*(1), 1–24.

Friedman, M. (1957). *A Theory of the Consumption Function*. National Bureau of Economic Research, Inc.

Hazen, B. T., C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics 154*, 72 – 80.

Kapteyn, A. and J. Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25*(3), 513 – 551.

Kennickell, A. B. (1998). Multiple imputation in the Survey of Consumer Finances. *Federal Reserve Board Working Paper*.

Lee, D. and W. v. der Klaauw (2010). An introduction to the FRBNY Consumer Credit Panel. *Federal Reserve Bank of New York Staff Reports* (479).

Liegeois, P., F. Berger, N. Islam, and R. Wagener (2011). Cross-validating administrative and survey datasets through microsimulation. *International Journal of Microsimulation 4*(1), 54 – 71.

Lynn, P., A. Jackle, S. P. Jenkins, and E. Sala (2012). The impact of questioning method on measurement error in panel survey measures of benefit receipt: Evidence from a validation study. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(1), 289 – 308.

McCallum, B. T. (1972). Relative asymptotic bias from errors of omission and measurement. *Econometrica 40*(4), 757 – 758.

Modigliani, F. and R. Brumberg (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In K. Kurihara (Ed.), *PostKeynesian Economics*. Rutgers University Press.

Nickerson, D. W. and T. Rogers (2014). Political campaigns and big data. *Journal of Economic Perspectives 28*(2), 51 – 74.

Openshaw, S. and P. J. Taylor (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences 21*, 127–144.

Pugh, K. and G. Foster (2014). Australia's national school data and the 'Big Data' revolution in education economics. *Australian Economic Review 47*(2), 258 – 268.

Qiao, C. (2005). Combining administrative and survey data to derive small-area estimates using loglinear modelling. *Labour 19*(4), 767 – 800.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schooley, D. K. and D. D. Worden (2010). Fueling the credit crisis: Who uses consumer credit and what drives debt burden? *Business Economics 45*(4), 266 – 276.

Sonka, S. (2014). Big data and the ag sector: More than lots of numbers. *International Food and Agribusiness Management Review 17*(1), 1 – 19.

Tudela, M. and G. Young (2005). The determinants of household debt and balance sheets in the United Kingdom. *Bank of England. Quarterly Bulletin 45*(3), 373.

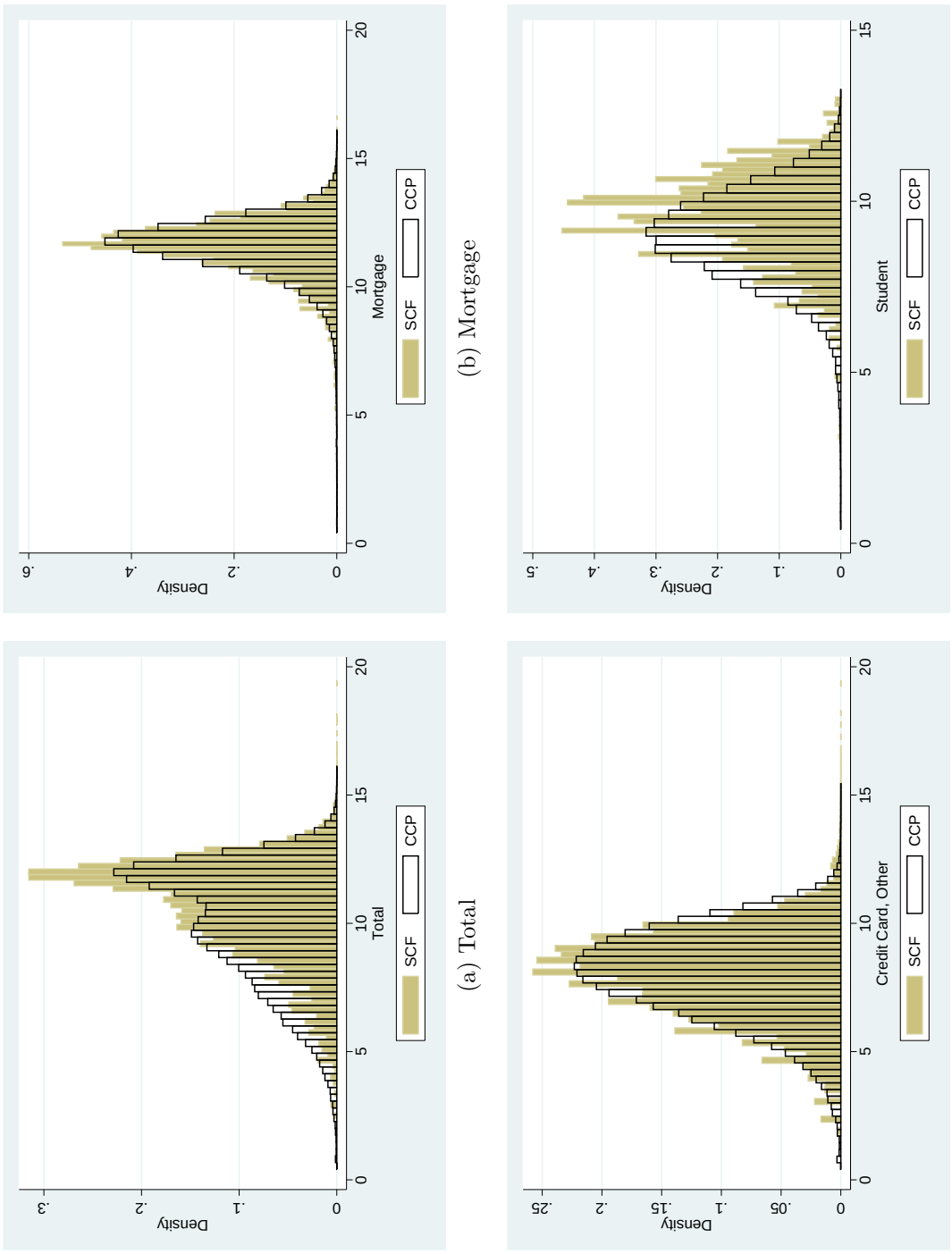
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3 – 28.

Warburton, R. N. and W. P. Warburton (2004). Canada needs better data for evidence-based policy: Inconsistencies between administrative and survey data on welfare dependence and education. *Canadian Public Policy 30*(3), 241 – 255.

Wickens, M. R. (1972). A note on the use of proxy variables. *Econometrica 40*(4), 759 – 761.

	Mean		SD		Median		Min		Max	
	CCP	SCF	CCP	SCF	CCP	SCF	CCP	SCF	CCP	SCF
Any Debt	0.81	0.78	0.40	0.41	1	1	0	0	1	1
Any Mortgage Debt	0.38	0.47	0.49	0.50	0	0	0	0	1	1
Any Auto Debt	0.34	0.32	0.47	0.47	0	0	0	0	1	1
Any Card Debt	0.72	0.49	0.45	0.50	1	1	0	0	1	1
Any Student Debt	0.15	0.21	0.35	0.41	0	0	0	0	1	1
Mortgage	$170,850	$167,956	$205,943	$199,815	$117,984	$118,000	$1.0	$188	$8,091,418	$16,615,000
Auto	$14,024	$14,567	$13,057	$21,954	$10,704	$12,000	$0.5	$72	$449,702	$6,237,500
Credit Cards, Other	$10,120	$11,596	$25,787	$498,580	$3,443	$13,900	$0.5	$8	$2,006,751	$277,512,496
Student	$16,713	$29,100	$27,949	$40,390	$8,297	$17,000	$0.5	$22	$390,428	$413,198
Total	$97,891	$122,135	$175,837	$436,196	$27,859	$60,400	$0.5	$10	$8,106,715	$277,512,500
Couple	0.52	0.55	0.50	0.50	1	1	0	0	1	1
Age	53	51	19	17	52	50	18	18	112	95
Household Income	.	$85,214	.	$364,944	.	$46,668	.	$0	.	$176,845,511
Tract Median Income	$59,720	.	$27,949	.	$53,554	.	$4,010	.	$246,500	.
Tract Per Capita Income	$74,667	.	$36,461	.	$65,843	.	$6,079	.	$514,211	.
Tract Income by Age	$59,896	.	$31,819	.	$53,182	.	$2,656	.	$249,554	.
Tract Income by Couple/Senior	$55,177	.	$32,981	.	$48,630	.	$2,639	.	$249,028	.
Log Income	10.90	10.73	0.45	1.14	10.89	10.75	8.30	0.00	12.42	19.01
No Degree	0.15	0.11	0.14	0.31	0.11	0	0	0	1	1
High School	0.29	0.31	0.14	0.46	0.29	0	0	0	1	1
Some College	0.29	0.19	0.12	0.39	0.28	0	0	0	1	1
Bachelor's	0.17	0.26	0.12	0.44	0.15	0	0	0	1	1
Graduate	0.10	0.13	0.10	0.34	0.07	0	0	0	0.85	1
Children	0.30	0.43	0.10	0.50	0.30	0	0	0	1	1

Table 1: Descriptive Statistics - Income and Demographics. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). The demographic observations for the SCF are included in the survey. The CCP measures age and whether the household contains a single adult or couple. All income, education, and child observations are merged into the CCP from the ACS. In the SCF sample, N=5,662. In the CCP sample, N=11,040,764.



(a) Total (b) Mortgage

(c) Credit Card (d) Student

Figure 1: Distributions of Logged Household Debts (Conditional (balance >0)) Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP) and American Community Surveys 2008-2012 (ACS).

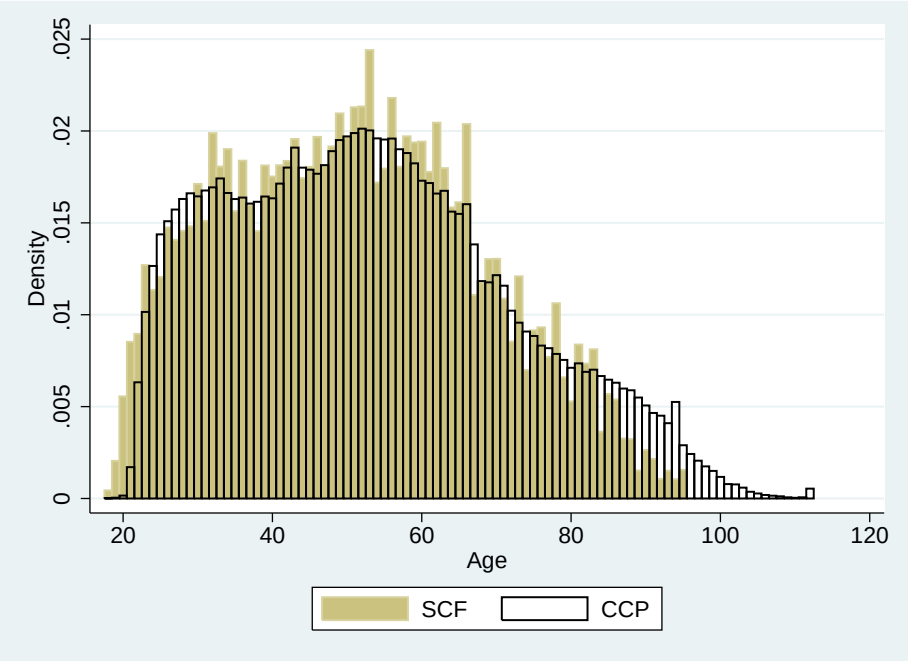


Figure 2: Age of Primary Respondent or Record. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF).

	SCF	CCP	SCF	CCP	SCF	CCP
Age	0.3822***	0.3083***	0.3066***	0.2522***	0.3288***	0.3053***
	(0.0214)	(0.0005)	(0.0213)	(0.0004)	(0.0906)	(0.0018)
Age^2	−0.0042***	−0.0032***	−0.0034***	−0.0027***	−0.0043*	−0.0042***
	(0.0002)	(0.0000)	(0.0002)	(0.0000)	(0.0018)	(0.0000)
Age^3					0.0000	0.0000***
					(0.0000)	(0.0000)
Couple			2.8122***	2.9051***		
			(0.1331)	(0.0029)		
Couple*Age					0.1212***	0.1347***
					(0.0276)	(0.0005)
Couple*Age^2					−0.0013	−0.0015***
					(0.0010)	(0.0000)
Couple*Age^3					0.0000	0.0000***
					(0.0000)	(0.0000)
Constant	0.7588	1.6396***	0.8313	1.6046***	0.8954	1.1019***
	(0.5369)	(0.0113)	(0.5171)	(0.0105)	(1.4191)	(0.0260)
R^2	0.1266	0.1256	0.2116	0.2305	0.2130	0.2338
N	5,662	11,040,764	5,662	11,040,764	5,662	11,040,764

Table 2: Models of Logged Total Household Debt with Internal Demographics. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). Significance key: * for p<.05, ** for p<.01, and *** for p<.001.

	SCF	CCP				
		Tract Median	Tract Per Capita	Age Merge	Couple/Senior Merge	Tract Share in Categories
Age	0.2458***	0.2460***	0.2467***	0.1950***	0.2451***	0.2459***
	(0.0223)	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0004)
Age^2	−0.0028***	−0.0027***	−0.0027***	−0.0022***	−0.0026***	−0.0027***
	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Couple	1.9599***	2.7329***	2.7508***	2.7815***	1.9018***	2.7232***
	(0.1608)	(0.0027)	(0.0027)	(0.0027)	(0.0049)	(0.0027)
Income	0.8964***	1.7605***	1.8256***	1.4287***	1.3454***	
	(0.1046)	(0.0056)	(0.0069)	(0.0046)	(0.0053)	
<$10k						−0.0393***
						(0.0006)
$10k-$15k						−0.0399***
						(0.0009)
$15k-$25k						−0.0321***
						(0.0007)
$25k-$35k						−0.0225***
						(0.0007)
$35k-$50k						−0.0101***
						(0.0007)
$75k-$100k						0.0108***
						(0.0007)
$100k-$150k						0.0145***
						(0.0006)
$150k-$200k						0.0131***
						(0.0008)
>$200k						0.0132***
						(0.0005)
Constant	−6.8243***	−17.3411***	−18.4812***	−12.6136***	−12.2604***	2.5718***
	(1.0336)	(0.0611)	(0.0753)	(0.0484)	(0.0559)	(0.0390)
R^2	0.2480	0.2613	0.2600	0.2576	0.2515	0.2628
N	5,662	11,030,492	11,033,464	10,785,638	10,972,930	11,034,525

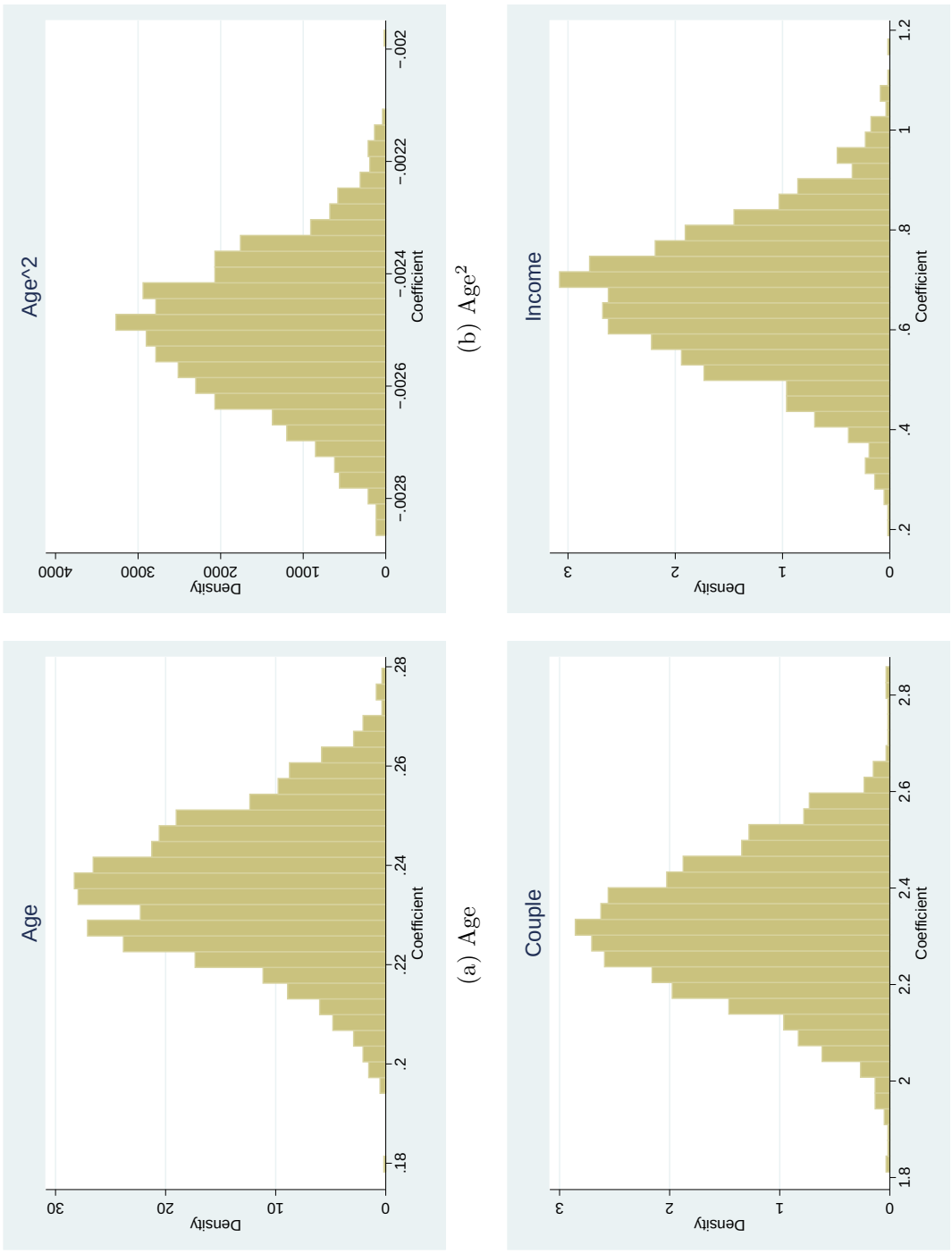
Table 3: Models of Logged Total Household Debt with Income Data. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). CCP model standard errors are clustered on the census tract. Significance key: * for p<.05, ** for p<.01, and *** for p<.001.

	SCF	CCP	SCF	CCP
Age	0.2458***	0.2450***	0.2420***	0.2354***
	(0.0223)	(0.0004)	(0.0216)	(0.0004)
Age^2	−0.0028***	−0.0026***	−0.0027***	−0.0025***
	(0.0002)	(0.0000)	(0.0002)	(0.0000)
Couple	1.9599***	1.9022***	2.0332***	2.3191***
	(0.1608)	(0.0049)	(0.1583)	(0.0047)
Income	0.8964***	1.3451***	0.6374***	0.6642***
	(0.1046)	(0.0053)	(0.0983)	(0.0057)
No Degree			−1.2656***	−1.6006***
			(0.2426)	(0.0235)
Some College			0.8679***	0.9485***
			(0.1750)	(0.0203)
Bachelor's			1.2994***	2.1667***
			(0.1656)	(0.0212)
Graduate			1.0962***	1.5634***
			(0.2067)	(0.0272)
Children			0.6118***	1.4387***
			(0.1294)	(0.0249)
Constant	−6.8243***	−12.2545***	−5.1233***	−6.0770***
	(1.0336)	(0.0559)	(0.9815)	(0.0589)
R^2	0.2481	0.2515	0.2750	0.2606
N	5,662	10,973,062	5,662	10,950,025

Table 4: Models of Logged Total Household Debt with Demographic Data. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). CCP model standard errors are clustered on the census tract. Significance key: * for p<.05, ** for p<.01, and *** for p<.001.

	CCP					SCF
	5 %	1 %	.1 %	.01 %	SCF Size	
Age	0.2354***	0.2365***	0.2367***	0.2418***	0.2032***	0.2420***
	(0.0004)	(0.0008)	(0.0024)	(0.0076)	(0.0157)	(0.0216)
Age^2	−0.0025***	−0.0025***	−0.0025***	−0.0025***	−0.0022***	−0.0027***
	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0002)
Couple	2.3191***	2.3278***	2.3160***	2.2999***	2.4664***	2.0332***
	(0.0047)	(0.0082)	(0.0239)	(0.0747)	(0.1426)	(0.1583)
Income	0.6642***	0.6494***	0.6623***	0.7661***	0.4328**	0.6374***
	(0.0057)	(0.0087)	(0.0236)	(0.0738)	(0.1412)	(0.0983)
No Degree	−1.6006***	−1.6173***	−1.6879***	−1.5997***	−2.3619***	−1.2656***
	(0.0235)	(0.0377)	(0.1057)	(0.3273)	(0.6314)	(0.2426)
Some College	0.9485***	0.9822***	1.0186***	1.2702***	0.3898	0.8679***
	(0.0203)	(0.0339)	(0.0975)	(0.3029)	(0.5872)	(0.1750)
Bachelor's	2.1667***	2.1578***	2.0593***	1.8728***	3.3225***	1.2994***
	(0.0212)	(0.0348)	(0.0976)	(0.2977)	(0.5772)	(0.1656)
Graduate	1.5634***	1.5866***	1.5465***	1.6561***	1.2324	1.0962***
	(0.0272)	(0.0420)	(0.1159)	(0.3417)	(0.7074)	(0.2067)
Children	1.4387***	1.4255***	1.4143***	1.3854***	2.1084***	0.6118***
	(0.0249)	(0.0343)	(0.0854)	(0.2589)	(0.5123)	(0.1294)
Constant	−6.0770***	−5.9540***	−6.0654***	−7.4511***	−2.9855*	−5.1233***
	(0.0589)	(0.0914)	(0.2498)	(0.7880)	(1.5086)	(0.9815)
R^2	0.2606	0.2601	0.2616	0.2656	0.2668	0.2750
N	10,950,025	2,190,047	219,005	21,922	5,634	5,662

Table 5: Model of Logged Household Debt using Various Sample Sizes. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). CCP model standard errors are clustered on the census tract. Significance key: * for p<.05, ** for p<.01, and *** for p<.001.



(a) Age (b) Age^2 (c) Couple (d) Income

Figure 3: Regression Coefficients based on 1,840 samples, N=6,000, from the CCP data. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP) and American Community Surveys 2008-2012 (ACS).

	SCF All Observations	CCP Debt >$0	CCP Debt >$250	SCF Levels	CCP Levels	CCP Tract Aggregates
Age	0.2476***	0.1734***	0.1503***	4078.3154***	5075.8555***	0.0574***
	(0.0220)	(0.0003)	(0.0003)	(403.7912)	(16.5588)	(0.0046)
Age^2	−0.0027***	−0.0017***	−0.0015***	−40.7422***	−48.3402***	−0.0009***
	(0.0002)	(0.0000)	(0.0000)	(3.7065)	(0.1472)	(0.0000)
Couple	1.2718***	0.8319***	0.7092***	30847.1071***	33277.4611***	4.8359***
	(0.1627)	(0.0029)	(0.0026)	(3374.7107)	(208.6129)	(0.0318)
Income	0.7005***	0.2250***	0.2275***	0.5132***	0.6379***	0.9834***
	(0.1025)	(0.0035)	(0.0032)	(0.0412)	(0.0052)	(0.0088)
No Degree	−1.4210***	−0.6326***	−0.5646***	−16698.7977***	4056.5604***	−1.5117***
	(0.2458)	(0.0145)	(0.0132)	(3491.7209)	(668.3976)	(0.0415)
Some College	0.9135***	0.7664***	0.6788***	7775.8987*	39981.7862***	0.9120***
	(0.182)	(0.0127)	(0.0115)	(3621.1836)	(739.1379)	(0.0428)
Bachelor's	1.4297***	1.3795***	1.3425***	38696.8369***	89690.9086***	1.9929***
	(0.1724)	(0.0141)	(0.0130)	(4390.0115)	(973.8093)	(0.0456)
Graduate	0.8868***	1.2999***	1.2154***	55186.0059***	121583.1351***	0.2227***
	(0.2248)	(0.0185)	(0.0167)	(6632.8609)	(1375.7355)	(0.0474)
Children	0.7956***	1.3751***	1.3043***	28020.6138***	74420.5474***	0.7956***
	(0.1364)	(0.0156)	(0.0143)	(3409.2542)	(1128.2653)	(0.0271)
Constant	−5.8784***	2.1972***	2.8944***	−86112.7718***	−163007.4144***	−6.3487***
	(1.0176)	(0.0363)	(0.0330)	(10124.2750)	(777.3295)	(0.1417)
R^2	0.2318	0.1957	0.1856	0.3112	0.1977	0.7667
N	6,015	8,737,664	8,373,889	5,159	10,920,300	71,737

Table 6: Alternate Specifications for Models of Total Household Debt. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). CCP model standard errors are clustered on the census tract. Significance key: * for p<.05, ** for p<.01, and *** for p<.001.

	Mortgage		Auto		Cards		Student	
	SCF	CCP	SCF	CCP	SCF	CCP	SCF	CCP
Age	0.3694***	0.3633***	0.0462*	0.1125***	0.1906***	0.2253***	−0.1703***	−0.1662***
	(0.0245)	(0.0006)	(0.0186)	(0.0004)	(0.0192)	(0.0004)	(0.0177)	(0.0004)
Age^2	−0.0034***	−0.0032***	−0.0007***	−0.0013***	−0.0019***	−0.0021***	0.0010***	0.0010***
	(0.0002)	(0.0000)	(0.0002)	(0.0000)	(0.0002)	(0.0000)	(0.0001)	(0.0000)
Couple	1.8197***	2.1308***	1.5606***	1.7980***	0.7969***	2.0560***	0.5661***	0.9489***
	(0.2106)	(0.0073)	(0.1432)	(0.0054)	(0.1394)	(0.0042)	(0.1175)	(0.0039)
Income	1.2035***	0.5704***	0.4866***	−0.0215**	0.1096	0.4844***	−0.2819***	−0.1917***
	(0.1503)	(0.0083)	(0.0663)	(0.0065)	(0.0648)	(0.0050)	(0.0581)	(0.0044)
No Degree	−1.1816***	−2.8316***	−0.1354	−2.0943***	−0.6494**	−1.0374***	−0.7596***	−0.5600***
	(0.2623)	(0.0337)	(0.2037)	(0.0262)	(0.2079)	(0.0209)	(0.1357)	(0.0170)
Some College	0.4329	1.2925***	0.1509	0.2064***	0.5525**	0.3654***	0.8890***	0.4821***
	(0.2212)	(0.0306)	(0.1761)	(0.0246)	(0.1753)	(0.0182)	(0.1503)	(0.0179)
Bachelor's	1.3572***	2.5422***	0.0050	0.9271***	0.1839	2.0372***	1.4195***	1.4489***
	(0.2210)	(0.0377)	(0.1704)	(0.0281)	(0.1667)	(0.0181)	(0.1497)	(0.0201)
Graduate	1.3520***	2.3240***	−0.3592	−0.8025***	−0.5420**	1.5229***	1.9684***	0.2077***
	(0.2770)	(0.0470)	(0.2173)	(0.0359)	(0.2076)	(0.0212)	(0.2037)	(0.0229)
Children	1.0536***	4.0920***	0.3539*	2.0201***	0.2834*	0.4911***	0.6478***	−0.4584***
	(0.1681)	(0.0448)	(0.1395)	(0.0344)	(0.1359)	(0.0221)	(0.1240)	(0.0210)
Constant	−18.5324***	−14.0517***	−3.5443***	0.2855***	−2.0348**	−6.4342***	9.5317***	8.5479***
	(1.3587)	(0.0839)	(0.7312)	(0.0677)	(0.7133)	(0.0524)	(0.7103)	(0.0477)
R^2	0.2576	0.1869	0.1043	0.1005	0.0613	0.2124	0.1701	0.1303
N	5,662	10,950,025	5,662	10,950,025	5,662	10,950,025	5,662	10,950,025

Table 7: Models of Logged Total Household Debt by Type of Debt. Data Sources: Federal Reserve Bank of New York Consumer Credit Panel/Equifax (CCP), Survey of Consumer Finances 2013 (SCF), and American Community Surveys 2008-2012 (ACS). CCP model standard errors are clustered on the census tract. Significance key: * for p<.05, ** for p<.01, and *** for p<.001.