Adjusting for selection bias in Web surveys using propensity scores: the case of the Health and Retirement Study

Matthias Schonlau[1], Arthur van Soest[1], Arie Kapteyn[1], Mick Couper[2], Joachim Winter[3]

[1]RAND

[2]University of Michigan

[3]University of Mannheim

Corresponding author: Matthias Schonlau, RAND 1700 Main Street, Santa Monica, P.O.Box 2138, CA. matt@rand.org

Paper prepared for presentation at the Joint Statistical Meetings, Toronto, August 2004.

Abstract

Many web surveys allow respondents to self select into the survey. Making inference about the population from a self-selected survey is very difficult. We analyzed data from the Health and Retirement Study (HRS) respondents of the 2002 wave as well as supplementary information about which subset of HRS respondents also responded to an additional web survey (web responders). The HRS is a longitudinal study of health, retirement and aging. The target population of the HRS includes all adults in the contiguous United States, aged 51 and over, who reside in households. We investigated whether it is possible to adjust for selection bias using propensity scores. We found that it is possible to make inferences for financial assets based on data from web responders only. However, making inferences about home values was not possible based on data from the web responders only.

Introduction

Web surveys are often the survey mode of choice because they are perceived to be expedient and inexpensive relative to phone or mail surveys. The desired speed can readily be achieved by using a convenience sample rather than a random[1] sample. However, it is randomization – with equal or unequal sampling probabilities - that makes it possible to draw inferences beyond the sample at hand (e.g. Kish, 1965, Cochran, 1977). Drawing inferences from convenience samples, including estimates of population frequencies and percentages, is a very hard problem. Unfortunately, this problem is often ignored (also see Schonlau et al., 2002).

Biostatisticians have long been accustomed to drawing inferences from observational studies because the randomization required for experiments can be unethical when dealing with human subjects or difficult to achieve in practice. Propensity scoring (Rosenbaum and Rubin, 1983, Rosenbaum, 2002) is commonly used to draw inferences from observational data.

Harris Interactive, a commercial web survey company, has adopted this approach for the use of web surveys. The Harris approach involves partitioning the propensity score into a categorical variable. This and other variables are then used for post-stratification. The Harris Interactive approach is described in more detail in Schonlau et al. (2004). Propensity scores in the context of web surveys is also described by Danielsson (2004).

The central issue is whether and under what circumstances propensity adjusted estimates are comparable to those based on random surveys. An integral component of the issue is what questions should be asked that capture the difference between the web respondents and the general population. Harris Interactive calls these elusive questions "webographic" questions, analogous to demographic questions. Other

[1] In this paper we use "random sample" and "probability sample" interchangeably.

researchers call them "lifestyle" or "attitudinal" questions.

Forsman and Varedian (2004) investigate this issue in the context of a marketing survey about the use of hygiene products and attitudes toward local banks. A phone survey (N=347) and a web survey (N=4724) were conducted in a northern European country. Their survey included lifestyle questions that were trying to capture a respondent's "modernity". They use logistic regression on lifestyle questions and demographic questions to capture the selection effect. They conclude that the estimates obtained from web and RDD phone surveys were different. Further, various different weighting schemes did not change the results very much.

Schonlau et al (2004) compared estimates from an RDD phone survey with propensity- adjusted estimates from a web survey conducted by Harris Interactive. They found that 8 out of 37 estimates investigated were not significantly different. Estimates from the web survey were significantly more likely to agree with estimates from the RDD phone survey for factual questions, when the question concerned the respondent's personal health, and when the question contained two as opposed to multiple categories.

In this paper we explore whether it is possible to make inferences for financial assets and house value based on data from web responders only for a future a wave of the HRS.

Method

The Health and Retirement Study is a panel study of persons 50 years and older, and their spouses (Juster and Suzman, 1995). The HRS is mainly a phone survey, but respondents over 80 years old are interviewed in person. The overall goals of the project are to explore the feasibility of using the Internet to supplement interviewer administered data collection in the HRS and to explore a variety of methodological issues related to Web –based measurement. The focus of this paper is to explore whether it is possible to adjust for selection bias to estimate the distribution of financial assets and the value of the primary residence given home ownership. Financial assets include shares of stock and stock mutual funds, checking accounts, saving accounts, bonds, money market funds and other financial assets.

The 2002 wave of the HRS contained a question asking about willingness to participate in a future web survey:

"We may want to try out a procedure for asking questions of some of the participants of this study, using the Internet. Would you be willing to consider answering questions on the Internet, if it took about 15 minutes of your time?"

This question was only asked to respondents who had indicated that they had access to the Internet in an earlier question. A companion paper (Couper et al, 2004) looks at the individual stages of selection: Internet access, willingness to participate given access, and nonresponse rates given willingness to participate. The 2002 wave of the HRS contained 16,698 respondents (respondents with person level weight equal to zero were excluded). 29.7% of the sample were identified as having Internet access. Of these, 73.3% indicated willingness to participate in a Web survey. A random sample of those who indicated willingness was sent a mailed invitation to participate in a Web survey. 78.4% completed the Web survey. For details see Couper et al. (2004). Overall, among the 16,698 HRS respondents 1,893 respondents completed the web survey.

We call respondents who volunteered to participate and completed the web survey the web responders or the web sample. Nowhere in this paper do we use the data of the actual web survey; we only use the information of who responded to the survey. This setup is different from the usual setup where estimates from a random sample are compared with those derived from a convenience samples. We compare estimates based on a random sample with estimates derived from the subsample of web responders. This setup isolates

the selection bias from the mode effect. Any differences found are due to selection bias only.

The 2002 wave of the HRS did not have any webographic questions that specifically targeted the difference between n the online and the general population. We use demographic questions, health related questions, and others that were available in the 2002 wave of the HRS.

Propensity scoring and weights

Let Y be a variable of interest (e.g. assets), X a set of x-variables (e.g. demographic and webographic variables) and I the indicator variable of whether the respondent belongs to the web sample or not. Denote the HRS sampling weights as w^{HRS}_i.

We wish to determine whether the distribution of Y can be estimated using only the web sample. First, we test whether Y and I are independent conditional on X, i.e. P(I=1|X)=P(I=1|X,Y). We perform a logistic regression of the indicator variables on (X,Y). If the regression coefficient for Y is statistically significant there is selectivity and it is not possible to estimate the distribution of Y from the web sample. We call this test the selection test. X may be a potentially large set of X variables. Using a backward stepwise regression (significance level for removal of a variable p=0.05) we reduce the number of x variables to a smaller set. In the following computations we use this reduced set of variables.

Second, we compute propensity scores $p_i = P(I_i=1|X_i)$ for i=1,…n where n is the number of observations in the HRS sample. The propensity scores are estimated from an unweighted logistic regression. Missing values were imputed prior to propensity scoring using a hotdeck procedure.

Because the web sample is a subsample of the HRS sample the two samples are not independent. To achieve independent samples we derive all HRS estimates only from the subset of respondents who did not respond to the web survey (I=0). We call this sample the non-web sample. About 90% of the respondents of the HRS sample are part of the non-web sample. The propensity weights for the two samples (I=0 and I=1) are as follows:

$$w_i^{ps} = \begin{cases} 1/p_i & \text{if } I=1 \\ 1/(1-p_i) & \text{if } I=0 \end{cases} \qquad (1.1)$$

The weights are the product of the sampling weights and the propensity weights.

$$w_i = w_i^{HRS} w_i^{ps} \qquad (1.2)$$

where w^{HRS}_i are the HRS sampling weights and i=1,…n. The unequal selection probabilities give rise to a probability design effect (Kish, 1965). The probability design effect, DE_p, can be computed as

$$DE_p = \frac{n\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} \qquad (1.3)$$

where n refers to the sample size of the sample or subsample of interest.

Throughout, we call estimates using the combined propensity and HRS sampling weights "adjusted web survey" estimates. We call estimates using the HRS sampling weight only "unadjusted web survey" estimates.

Analysis Plan

We analyzed the results both graphically (using parallel box plots and histograms) and using a formal test of significance. We test the hypothesis that the two empirical distributions arise from the same theoretical distribution. Both empirical distributions are constructed from weighted data. The commonly used Kolmogorov –Smirnov statistic for equality of distributions is based on the maximal difference between two cumulative distribution functions (CDF). To our knowledge this test has not yet been developed for weighted data. The Cramer-van-Mises test is based on the integrated squared difference between two CDF's. It could be adapted for use with weights; however,

the implementation is awkward as it would require numerical integration.

Instead, we test the hypothesis that the two distributions are equal as follows: For the non-web sample, we divide the empirical distribution of the variable of interest into 10 deciles. For the web sample, we use the same cutoff values as for the non-web sample and also divide the distribution of the variable of interest into 10 categories. If the two empirical probability distribution functions stem from the same distribution then the categories for the web sample also should each contain approximately 10% of the observations. We use Pearson's chi-squared test adjusted for use with weights (Rao and Scott, 1981 and 1984) to test independence in the two-way table.

Throughout we use Stata for the analysis. Pearson's chi squared test for weighted data is implemented in Stata's procedure for survey data "svytab".

Results

For the selection test we used all variables listed in Table 1 as well as log financial assets[2], indicator variable of home ownership, log house value, an indicator for smoking, indicator variables for religion, presence of various diseases (lung disease, heart attack, arthritis, stroke, cancer, diabetes, depression, high blood pressure), various measures of problems with of activities of daily living (climbing stairs, etc.), indicator variables for loneliness and happiness, and whether the respondent had any doctor visit in the previous 2 years or not.

Most of the variables were not significant. The significant variables were the variables listed in Table 1 as well as house ownership and house value, the indicator variables for smoking, having seen a doctor, lung disease and an indicator of that taking a bath/showering is difficult for the respondent.

Table 1 displays the results of the logistic regression that generates the propensity scores. The variables contained in Table 1 are the variables used to adjust for selection bias for both assets and house value. As expected, participation in the web survey (which requires web access) is greater among younger, White, educated respondents with greater income and who are in better health. Unexpected is that females are significantly more likely to respond and those with zero income. The latter may be a marker for retired respondents.

		Odds Ratio	p
Race /Ethnicity	White	1.00	
	African American	0.29	0.00
	Other Race	0.45	0.00
	Hispanic	0.36	0.00
Education	<high school	0.25	0.00
	high school	1.00	
	some college	1.86	0.00
	>= college	2.74	0.00
Age	<55	1.17	0.11
	55-65	1.00	
	65-75	0.70	0.00
	>75	0.24	0.00
Gender	Male	0.82	0.00
Marital Status	Married	1.00	
	Partnered, unmarried	1.24	0.17
	separated, divorced	0.67	0.00
	widowed, never married	0.62	0.00
Income	log10 income	1.67	0.00
	Indicator (Income==0)	6.36	0.00
Self Assessed	Excellent	1.06	0.40
Health	Very Good	1.00	
	Good	0.75	0.00
	Fair	0.54	0.00
	Poor	0.40	0.00

Table 1: Logistic regression of participation in the web survey on various covariates.

[2] The log transformations for assets and house values are computed as $\log 10(y + \sqrt{1 + y^2})$ where y is the variable to be transformed. This transformation ensures that the argument of the log is never negative.

Descriptive and Graphical analysis

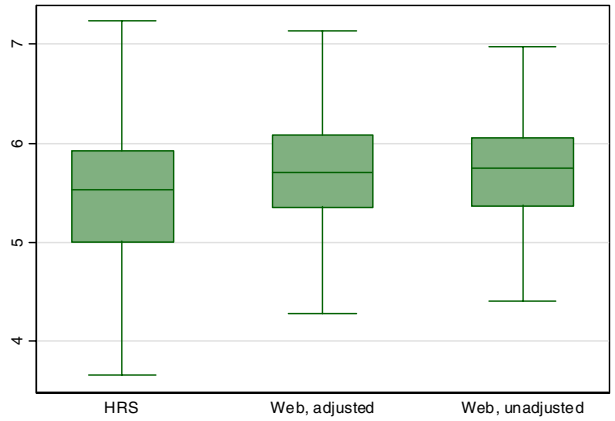

Figure 1: Parallel Box plots of assets based on the non-web sample (HRS), the adjusted web sample and the unadjusted web sample.

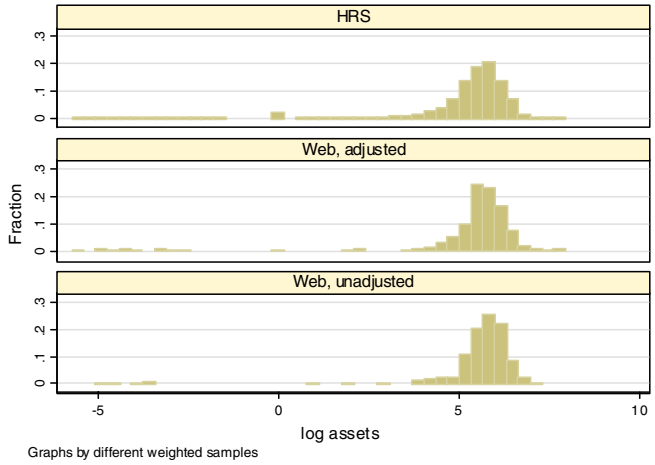

Figure 2: Histograms of log assets based on the non-web sample (HRS), the adjusted web sample and the unadjusted web sample.

The mean log asset values are: 5.03 (non-web sample), 5.20 (adjusted web sample), 5.58 (unadjusted web sample). Figure 1 and Figure 2 contain box plots and histograms of the assets respectively. The shaded area in the box plots corresponds to the inter-quartile range; the horizontal line in the center of the shaded area is the median. The adjustment seems to work in the right direction. The box plots show that the weighting adjustment to the web sample lowers the estimate for median assets and increases the variation.

Among those respondents who own a house, the mean log house values are: 5.38 (HRS), 5.46 (adjusted web), 5.53 (unadjusted web). Figure 3 and Figure 4 contain box plots and histograms of the house value respectively. Again, the propensity scoring adjustment reduces the differences. It is unclear from the graphical analysis whether the differences are significant or not.

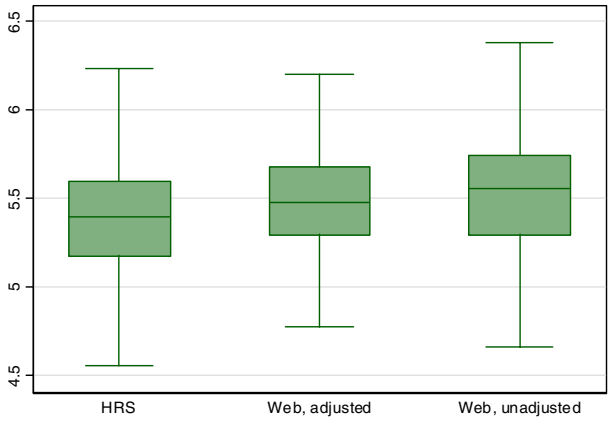

Figure 3: Parallel Box plots of house values based on the non-web sample (HRS), the adjusted web sample and the unadjusted web sample.

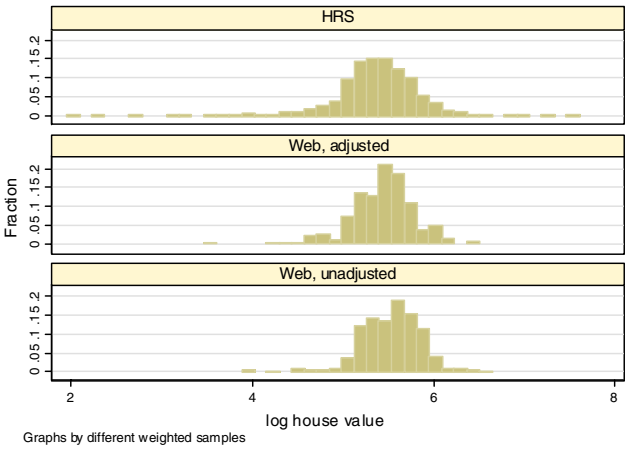

Figure 4: Histograms of log house value based on the non-web sample (HRS), the adjusted web sample and the unadjusted web sample.

Formal tests for differences in distribution

The non-web sample was divided into 10 groups such that each group would have 10% of the observations. The web sample was divided into 10 groups using the same cutoff points as used for the non-web sample. Under the null hypothesis that the two distributions are equal the proportions should also be evenly divided among the 10 groups. Without the adjustment the distribution of assets based on web and non-web samples respectively were significantly different ($p<0.0001$). With the propensity adjustment, there was no evidence to reject the null hypothesis ($p=0.42$). Table 2 gives the 2 by 10 table on which the weighted chi-squared test is based. By design each decile in the non-web sample contains 10% of the data. Decile 1 corresponds to the respondents with the least assets; decile 10 to those with the most assets.

Decile	non-web sample	adjusted Web Sample	Difference
1	10.0%	7.7%	-2.4%
2	10.0%	9.3%	-0.7%
3	10.0%	7.2%	-2.8%
4	10.0%	9.2%	-0.8%
5	10.0%	11.1%	1.1%
6	10.0%	8.1%	-1.9%
7	10.0%	10.7%	0.7%
8	10.0%	11.8%	1.8%
9	10.0%	13.3%	3.3%
10	10.0%	11.6%	1.6%
sum	100.0%	100.0%	0.0%

Table 2: Weighted percentages of observations of non-web (HRS) and web samples in each of 10 deciles for log assets.

Without the adjustment the distribution of the value of the house (given home ownership) based on web and non-web (or HRS) samples respectively were different ($p<0.0001$). With the propensity adjustment, there the distributions were still significantly different ($p<0.0001$). Table 3 gives the table underlying on which the weighted chi-squared test is based. The column corresponding to the non-web sample is not as evenly divided as the one in Table 2. This is due to duplicates in the empirical distribution of house values. There are only 415 unique values for house value, whereas there are 4613 unique values for assets in the combined sample. (The house value was derived from a single question). Generally, the house values estimated from the web sample are higher than those from the non-web sample. Therefore, there are fewer respondents estimated to be in the lowest deciles.

Decile	non-web sample	adjusted Web sample	Difference
1	10.3%	4.6%	-5.7%
2	9.8%	6.6%	-3.2%
3	10.6%	8.4%	-2.2%
4	12.7%	13.7%	1.0%
5	7.9%	8.4%	0.4%
6	11.1%	12.2%	1.1%
7	7.9%	9.1%	1.3%
8	12.6%	16.4%	3.8%
9	8.6%	11.5%	2.9%
10	8.6%	9.2%	0.6%
sum	100.0%	100.0%	0.0%

Table 3: Weighted percentages of observations of the non-web (HRS) and web samples in each of 10 deciles for log house value given house ownership

The test is robust w.r.t. the number of categories. We used 50, 10 and 20 categories and achieved similar results. Instead of the person level HRS sampling weight we applied the household level weight and achieved the same results. However, when dropping the sampling weights altogether the test for log assets became marginally significant at 2.3% mostly due to differences in the lowest decile of the distribution of log assets.

Probability design effects and the effect of trimming

When some weights are large, sampling weights are often trimmed (Kish, 1965) to avoid that a few observations which large weights have a large influence on the estimates. Trimming is one version of the classic bias-variance tradeoff, in which one is willing to accept a small amount of bias in exchange for a large reduction of variation. The effect on variation is most easily explored through the probability design effect. A larger probability design effect reduces the effective sample size, which in turn increases the variance.

Table 4 displays various probability design effects corresponding to the following weights: sampling weights only (Full HRS sample), sampling weights of the non-web responders with corresponding propensity adjustment (non-web sample with propensity), sampling weights of the web responders only (web sample), sampling weights and propensity adjustment of the web responders only (web sample with propensity).

	Design Effect	DE - Weights trimmed at 10
Full HRS sample	1.41	1.41
Non-web sample with propensity	1.53	1.53
Web sample	3.98	2.36
Web sample with propensity	5.94	2.68

Table 4: Probability design effects with and without trimming

The probability design effect for the HRS and the non-web sample with propensity do not change because the trimming cutoff, 10, is larger than the largest weight.

It is interesting that the design effect for the web sample is more than twice as large than for the HRS sample even before propensity scoring. It implies sampling probabilities of web survey respondents were more variable than those of HRS respondents. With propensity scoring the design effect, 5.94, for web surveys is very large. It implies that the 1893 respondents in the web sample really only represent an effective sample size of 318 respondents. Trimming the weights at 10 reduces the design effect very substantially.

We ran the formal test for differences in distribution with weights trimmed at 10. With trimmed weights the distribution of asset as estimated by the two samples were (highly) significantly different.

In conclusion, trimming is a bad idea when trying to adjust for selection bias. This is unusual because for most probability surveys with large weights trimming is a sensible strategy. While large weights are usually undesirable, the propensity score methodology uses weights to adjust for selection bias. By trimming the weights the propensity score adjustment process is partially disabled.

Discussion

Our results suggest that assets can be estimated based on data from web survey respondents only (assuming one can adjust for potential mode effects). If a future HRS wave were to be conducted partially on the web, it would be sufficient to ask asset questions only in the web survey. For the adjustment to work the following 7 variables should be asked in both survey modes (based on Table 1): education, age, race/ethnicity, gender, income, marital status, and self-assessed health. Income more than other variables is prone to missing values. We do not consider this an obstacle – missing values can be imputed prior to propensity scoring.

In the propensity score adjustment we used 7 variables that were motivated by the selection test. This attempted to reduce this number of variables even more. The minimal set of variables required to adjust for selection bias for the distribution of assets (i.e. not rejecting the hypothesis that the distributions are equal) consisted of only three variables: race / ethnicity, education, and age. We used the other variables (marital status, gender,

income, and self-assessed health) in addition in part for face validity and in part because they did bring the distributions (insignificantly) closer together.

It is not possible to adjust for the selection bias to estimate house values with the current set of variables. There is large regional variation in house prices in the U.S., and the model inadequately captures regional variation. The model contains no information of whether the house is in a city or in a rural area where one might expect lower prices. It should be noted though that the adjustment for the selection bias goes part way in the right direction.

We have only estimated the distribution of two variables, financial assets and house values. Our goal is to look at a series of variables, and to explore whether it is possible to find a single set of x variables that can be used to adjust for selection bias. A single set of x variables would imply a single set of propensity scores as well as a single set of weights.

Acknowledgement

Support for this research comes from grant R01AG20717 from the National Institute of Aging of the U.S. National Institutes to RAND (Arie Kapteyn, P.I.) and from the University of Michigan's Survey Research Center (Robert J Willis, P.I.).

References

Cochran WG. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons; 1977.

Couper MP, Kapteyn A, Schonlau M, Winter J. Noncoverage and Nonresponse in an Internet Survey. In *Proceedings of the International Conference on Social Science Methodology* (RC33), Amsterdam, August 2004.

Danielsson, S. The propensity score and estimation in nonrandom surveys: an overview. Department of Statistics, University of Linkoeping. Report no. 18 from the project " Modern statistical survey methods". http://www.statistics.su.se/modernsurveys/publ/11.pdf (accessed in August 2004).

Juster FT, Suzman R. (1995) An Overview of the Health and Retirement Study. *The Journal of Human Resources*, 30, Special Issue on the Health and Retirement Study: Data Quality and Early Results. pp. S7-S56.

Kish L. *Survey Sampling*. New York: John Wiley & Sons; 1965.

Rao JNK, Scott AJ. (1981) The Analysis of Categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, no 374, 221-230.

Rao JNK, Scott AJ. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*. 12:46-60.

Rosenbaum PR. *Observational Studies*. 2nd ed. New York: Springer-Verlag; 2002.

Rosenbaum, PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41-55.

Schonlau, M., Fricker R., Elliott, M. *Conducting Research Surveys via Email and the Web*, Santa Monica, CA: RAND. 2002.

Schonlau M, Zapert K, Payne Simon L, Sanstad K, Marcus S, Adams J, Spranca M, Kan H-J, Turner R, Berry S. A comparison between a propensity weighted web survey and an identical RDD survey. *Social Science Computer Review.* 2004; 22(1).

Varedian M and Forsman G (2003). "Comparing Propensity Score Weighting with Other Weighting Methods: A Case Study on Web Data" In *Proceedings of the Section on Survey Statistics*, American Statistical Association; 2003, CD-ROM.