Do the Differences Matter?

In the previous section, we documented various errors in the Homescan data. The Homescan data are an input into statistical analysis. In principle, even though the data are recorded with error, the analysis could still be mostly unaffected. In this section we ask if the recording errors matter for the conclusions drawn from the analysis. Obviously, the answer depends on the use of the data. We focus on one particular use. Recently, researchers have used Homescan data to study how the prices paid vary with household demographics (e.g., Aguiar and Hurst, 2007). We perform a simple version of such a study in order to evaluate the impact of the errors. Our goal is not to replicate any particular study, but just to investigate whether the errors could have important implications for certain bottom lines.

We present results from a least-squares regression of price paid on household characteristic and UPC fixed effects (table 7). In each set of columns, the first column displays the estimates and the second displays the t-statistics. An observation is a product (UPC) in a matched large trip, i.e., in a large trip with r_1 greater than 0.7. The first two columns use as the dependent variable the price, in cents, as recorded in Homescan, and the next two columns use

Table 7

Illustrative analysis of how errors could affect bottom line

Dependent variable	Price (Homescan)		Price (Retailer)		Same	Same statistical	Coefficient
	Coefficient	t-stat	Coefficient	t-stats	sign	significance	ratio
Constant	286.15	27.39	295.10	9.95			0.97
HH size	-1.32	-2.26	-3.11	0.56	yes	yes	0.42
HH income	0.01	0.86	0.09	0.01	yes	no	0.14
No female head of HH	-41.12	-4.36	-32.85	8.99	yes	yes	1.25
Age female	-1.25	-3.45	-1.71	0.34	yes	yes	0.73
Age female ^ 2	0.01	2.94	0.02	0.00	yes	yes	0.51
No male head of HH	11.51	1.18	-33.06	9.27	no	no	NA
Age male	-0.40	-1.04	-1.34	0.36	yes	no	0.29
Age male ^ 2	0.01	1.37	0.01	0.00	yes	no	0.41
No. of children < 18 yrs	3.42	2.43	1.84	1.34	yes	no	1.87
No. of young children < 6 yrs	-0.81	-0.39	3.61	1.96	no	yes	NA
Male employed	-0.58	-0.27	-11.02	2.08	yes	no	0.05
Male fully employed	5.48	2.64	17.66	1.98	yes	yes	0.31
Female employed	5.26	4.28	1.01	1.17	yes	no	5.18
Female fully employed	-4.08	-3.37	-3.29	1.16	yes	yes	1.24
Male education	1.19	2.69	-1.32	0.42	no	yes	NA
Female education	-1.33	-2.74	1.25	0.46	no	yes	NA
Married	4.79	3.96	1.90	1.15	yes	no	2.52
Non-white	-3.63	-2.35	1.30	1.47	no	no	NA
Hispanic	-3.45	-1.88	-2.99	1.75	yes	yes	1.16
“15K” HH	-1.14	-0.83	-2.47	1.31	yes	yes	0.46
UPC fixed effects	yes		yes				
R^2	0.912		0.910				
Observations	41,158		41,158				

HH = household

An observation in this table is a distinct item (UPC) in a given trip.

The sample used in both regressions is all matched items in the matched large trips.

Regressions include UPC fixed effects, so coefficients indicate the effect of demographics on price paid for an identical item.

Source: Authors’ calculations using Homescan and retailer data.

the price in the retailer’s data. The last three columns report whether the sign on the coefficients is the same in the two specifications, whether they agree in terms of statistical significance of the coefficient (at a 5-percent confidence level), and the ratio between the coefficient (when the signs agree). Since the regressions include UPC fixed effects, the results tell how the demographics correlate with the price a particular household paid relative to the average (in the sample) price paid for the same item.

The different data give different results. Out of the 20 slope parameters, 5 have different signs, 9 do not agree on their statistical significance, and 13 are statistically different. It is interesting to note that in almost all the cases of statistically significance disagreement, the retailer’s data generate significant estimates, while the Homescan data do not. In many cases the difference is also economically meaningful. For example, in the Homescan data the coefficient on race dummy variable is negative and significant, which implies that non-White consumers pay a lower price. On the other hand, in the retailer’s data the coefficient is positive but not significant. A researcher using the Homescan data to study discrimination would probably reach different conclusions than one using the retailer’s data to study the same question, using the very same set of shopping trips. Another example is in the impact of age on price paid. The Homescan data suggest a flatter impact of age, especially for males, than the retailer’s data. Once again researchers using the data to study life cycle consumption might reach wrong conclusion using the Homescan data.

There are two factors that cause the difference in the results. First, Nielsen imputes store level prices for many of the observations. Suppose that all the price information in the Homescan data were imputed and consider, for example, the race dummy variable. In this case, the regression using the Homescan data shows that non-White households tend to buy at cheaper stores, i.e., stores where the average consumer in the store pays less for the same item. The regression using the retailer’s data tells us that despite going to cheaper stores non-White panelists do not pay less on average.

A second reason for the difference in the results is due to recording errors. Suppose that none of the prices are imputed and the only difference is due to recording mistakes made by the panelist. Once again, we use the race dummy variables as an example. The regression using the Homescan data tells us that nonwhite consumers report a lower price. On the other hand, the regression using the retailer’s data suggests that those consumers do not actually pay less, maybe even slightly more. Together these suggest that White consumers tend to over-report prices relative to non-White consumers, not that the White consumers are likely to pay more.