Not-So-Classical Measurement Error: Evidence from *HomeScan*

Liran Einav[1] Ephraim Leibtag[2] Aviv Nevo[3]

[1]Stanford University and NBER

[2]USDA ERS

[3]Northwestern University and NBER

2008 World Congress on National Accounts and Economic Performance Measures for Nations

Big Picture

- Surveys and self-reported data are at the heart of many economic data sets, e.g. PSID, CPS, CEX, ...
- The data quality ("making of the sausage") is important to trust findings based on these data;
- In general, survey data raise 2 concerns: sample selection and recording error;
- Sample selection is one of the most studied areas of econometrics;
- Recording/measurement error is somewhat less studied:
    - classical error in the linear model;
    - general results are hard to get;
- This paper fits into a literature that uses cross-validation samples to study recording errors and their implications;

Overview

- *Nielsen Homescan* is a large and increasingly used data set in which panelists scan (at home) all their grocery purchases;
- The Homescan data set has been used for:
  - Marketing/IO purposes
  - Study consumption
  - Generate price indices
- In an ongoing project we plan to use these data (together with retailer price data) to analyze the store choice.
- Along the way, we realized that this also provides us with a rare opportunity to run a "validation study":
  assess the extent and nature of measurement errors in *Homescan* using external data (from a retailer) about the "truth."

Who and why should care?

- Possible general interest relevance: studies of measurement errors:
    - classical errors often assumed. evidence?
    - validation studies (we are aware of) are in the context of labor (e.g. PSID). we look at IO/Marketing type of data.
    - Some of the results seem to be relevant elsewhere: “smarter” and less busy individuals less likely to be an issue.
- *Homescan* specific relevance:
    - Could provide specific guidance to the use of Homescan
        - Huge amount of non-academic use (suppliers, retailers, gov. agencies)
        - Smaller academic use but increasing
    - May impact results in the literature
- Nielsen: May guide ways to better collect/report the data

Goals

- Cross validate the Homescan data
    - Is there mis-recording in the data?
    - If so, what is the magnitude? what are the patterns?
- The impact of mis-recording on the bottom line
    - Is mis-recording correlated with household attributes?
    - Can mis-recording bias results?
    - Can a correlation between a price “paid” and demographics be driven by mis-recording?
- Suggest ways to either select the more reliable data or make adjustments to improve the quality of the data
- More broadly, a rare opportunity to learn about the reliability of self-reported data

General Strategy

- Start with 2004 Homescan data, and construct matched data from a large retailer (R) in two steps:
  1. Select sample from Homescan trips to R's stores and request entire transaction record for these store-days
  2. Find matched transactions, and use it to match with loyalty card, then request entire transactions of that household in R's data
- Describe quality of matched trips
- Used matched transactions to document mis-recording of product and price/quantity
- Correlate mis-recording with demographics and compare regressions using HS and R's data

Terminology

- "Truth" $=$ R's data (even though not always true)
- "Mis-Recording," "Reporting Errors," etc. refers (interchangeably) to panelists and/or Nielsen's data construction.

Summary of Results

- For roughly 20% of the Homescan trips we can say (with high probability) that no match exists in R's records;
- Product, for matched trips:
  - On average, approximately 10-14% of the items in R's records are not reported in Homescan;
- Price/quantity information, for matched items:
  - Quantity: 93% match
  - Deal indicator: matches in 80% of cases
  - Price: match in less than 70% of cases
- Heterogeneity across households; correlated errors within households;
- Errors are correlated with demographics.

Comment on Price Imputation

- The price (and expenditure) variable(s) had the lowest match rate
- This should not be surprising given the way the price variable is generated
- Indeed, conditional on no deal the match rate increased significantly (in principle, imputation should be less problematic)
- For some purposes the imputed price is very useful (indeed, maybe better than the actual price)
    - It is also easier to collect
- However, in many cases having only the imputed price might be a problem

Outline

- The data sets
- Data construction and matching algorithm
- Documenting the accuracy of the data
- Using the validation sample to correct the reporting error
- Implications

Homescan Data

- We use all food purchases in the Homescan data during 2004.
- 61,000 panelists, mainly in big markets. (15,000 of those also record produce and other fresh food). "Static" sample: approx 40K (8K).
- In principle ... all grocery shopping trips should be recorded, including a gum they buy at the movies.
- Overall, quite unique data. Main advantages over alternatives (e.g., POS data, loyalty-card panels, competitors data):
  - multiple stores and mass merchants (e.g., Wal-mart)
  - many households with variation in location and demographics
  - many product categories including random weight and fresh food
- Two commonly raised concerns
  - Is the sample of households representative of the population of interest?
  - Do the panelists record their purchases properly?
- Our primary focus is on the latter.

Possible mis-recording

- Trip
  - Miss a trip to a store
  - Mis-record trip details (store/date)
- Product
  - Not record or mis-record product (UPC) information
- Price/quantity Information
  - Mis-record price/quantity/expenditure/deal information

Retailer Data

- We obtained a rich data set from one large retailer
- For each day-store record of all the transactions
- For each transaction
    - list of all UPC's bought
    - cashier id
- For each UPC
    - expenditure (gross and net) and quantity
    - exact time and sequence in purchase

Data Construction Step 1

- Select a sample of trips in the Homescan data to the R's stores
  - Focused on 189 stores in 2 markets
  - looked at HS households that:
    - had at least one trip of at least 5 items after Feb 15
    - household expense in R more than 20% and less than 80%
  - Gave us 342 households
  - For 240 we choose a single (random) trip
  - Other 102 (with at least 10 trips but no more than 20, to R) all trips
- Obtain data from R for all the transaction for these store-days
  - Got 1,603 store days,
- **Bottom line**: 2,579 potential trips to match.

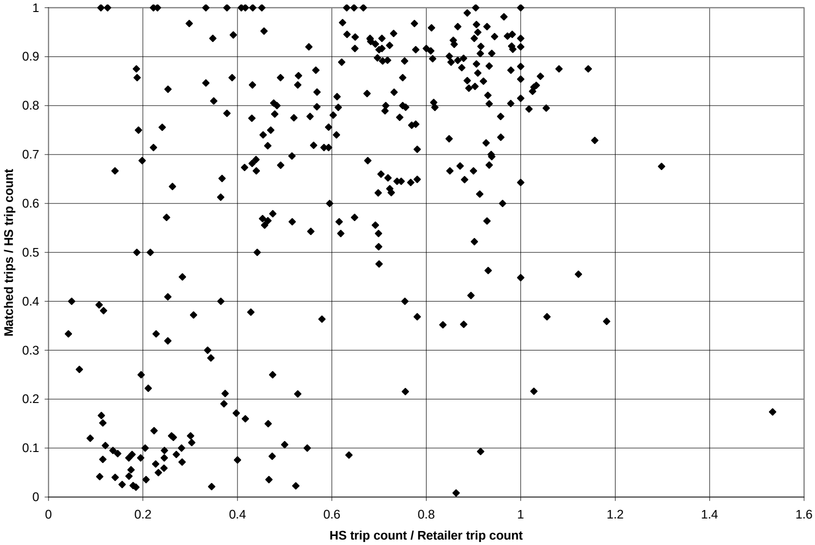
Data Construction Step 2

- Used a simple matching process: found 1,372 likely matched trips (293 households) from Step 1
- Asked R for all transactions of these involving loyalty cards of these 293 households (R tries to link different cards used by a household)
  - Got 40,036 transactions, with 27,746 unique store-date-HH combinations
  - 3,884 of these were already in the Step 1 R data

Record-matching Overview

- **The Goal:** Classify each Homescan record as either
  - matched with a unique R transaction
  - no match (i.e., with high probability does not have a match)
  - uncertain (i.e., none of the above)
- The information is different for Step 1 and Step 2 data
- Step 1 data: match Homescan record to one of many R transactions
- Step 2 data: ask if Homescan trip matches the single transaction obtained for the loyalty card

Figure 5: Households' Missed and Mis-recorded Trips



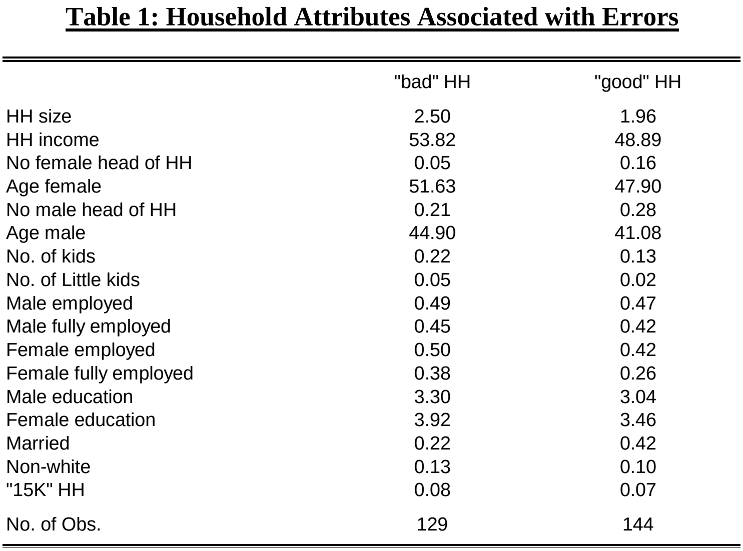


Table 1: Household Attributes Associated with Errors

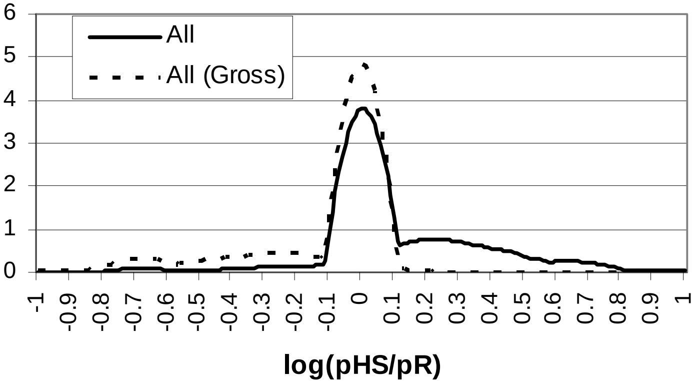
	"bad" HH	"good" HH
HH size	2.50	1.96
HH income	53.82	48.89
No female head of HH	0.05	0.16
Age female	51.63	47.90
No male head of HH	0.21	0.28
Age male	44.90	41.08
No. of kids	0.22	0.13
No. of Little kids	0.05	0.02
Male employed	0.49	0.47
Male fully employed	0.45	0.42
Female employed	0.50	0.42
Female fully employed	0.38	0.26
Male education	3.30	3.04
Female education	3.92	3.46
Married	0.22	0.42
Non-white	0.13	0.10
"15K" HH	0.08	0.07
No. of Obs.	129	144

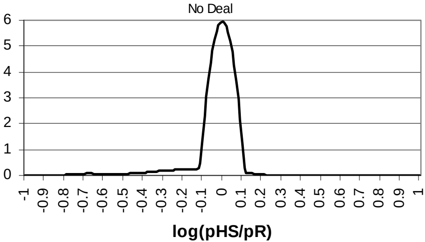
Price/Quantity Information

What fraction of observations have incorrect price/quantity

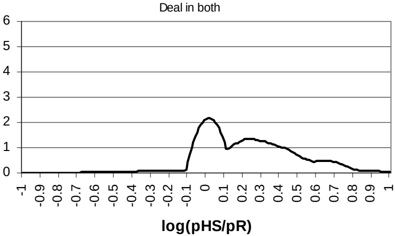
- We focus on matched items in matched trips
- We find
  - Quantity matched 93%
  - Expenditure matched less than 60%
  - Price matched less than 70%
  - Deal indicator matched 80%
- Expenditure and price are impacted by price imputation

Price matching quality





No Deal



Deal in both

Using the validation sample to control for recording error

- Basic intuition:
  - Use the validation sample to learn the distribution of the error (conditional on variables observed in the primary data)
  - Use the recovered distribution to "integrate over" the distribution of the error in the primary data;
- Key assumption: the (conditional) distribution of the error is the same in both data sets.
- For example, in our application this implies that the recording errors are the same in all stores.

Using the validation sample to control for recording error

- Moment condition: $E[m(X^*, \beta_0)] = 0$
- Primary data set: $\{X_{pi} : i = 1...N_p\}$
- Validation data set $\{(X^*_{vj}, X_{pi}) : j = 1...N_v\}$
- Key Assumption: $f_{X_v^*|X_v=x} = f_{X_p^*|X_p=x}$
- One possible way to proceed is to compute

$$\widehat{f_{X_p^*}(x^*)} = \int f_{X_v^*|X_v=x}(x)\widehat{f_{X_p}(x)}dx$$

$$\widehat{\beta} = \arg\min(\int m(X^*, \beta)\widehat{f_{X_p^*}}(x^*)dx^*)'\widehat{W}(\int m(X^*, \beta)\widehat{f_{X_p^*}}(x^*)dx^*)$$

Using the validation sample to control for recording error

- This is computationally intense so we follow Chen, Hong and Tamer (REStud, 2005)
- Define

$$g(X,\beta) \equiv E[m(X^*,\beta)|X_p = x] = \int m(X^*,\beta) f_{X_p^*|X_p=x}(x^*)dx^*$$

- given our moment condition

$$E_p[g(X,\beta)] = \int g(X,\beta) f_{X_p}(x)dx = 0$$

- The key condition implies

$$g(X,\beta) \equiv E[m(X_v^*,\beta)|X_v = x] = \int m(X^*,\beta) f_{X_v^*|X_v=x}(x^*)dx^*$$

Using the validation sample to control for recording error

- Chen, Hong and Tamer propose

$$\widehat{\beta} = \arg\min\left(\frac{1}{N_p}\sum_{i=1}^{N_p}\widehat{g}(X_{pi},\beta)\right)'\widehat{W}\left(\frac{1}{N_p}\sum_{i=1}^{N_p}\widehat{g}(X_{pi},\beta)\right)$$

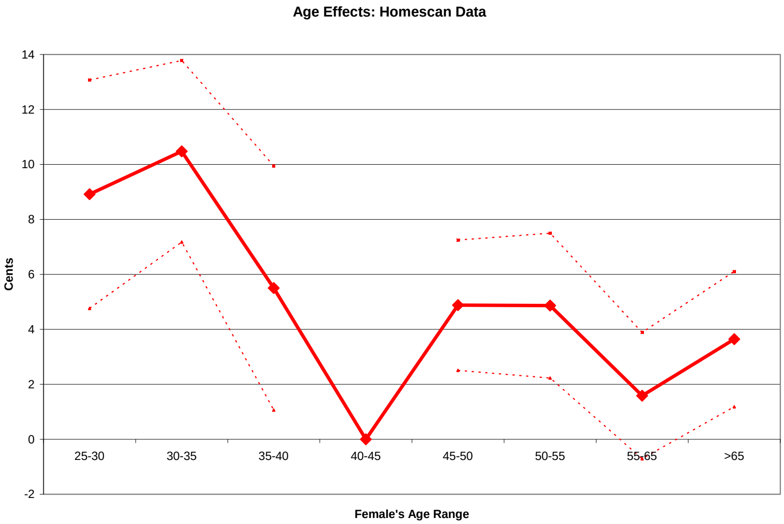
- where $\widehat{g}(X_{pi},\beta)$ is a non-parametric estimate of $g(X_{pi},\beta)$ and $\widehat{W}$ is a weight matrix.

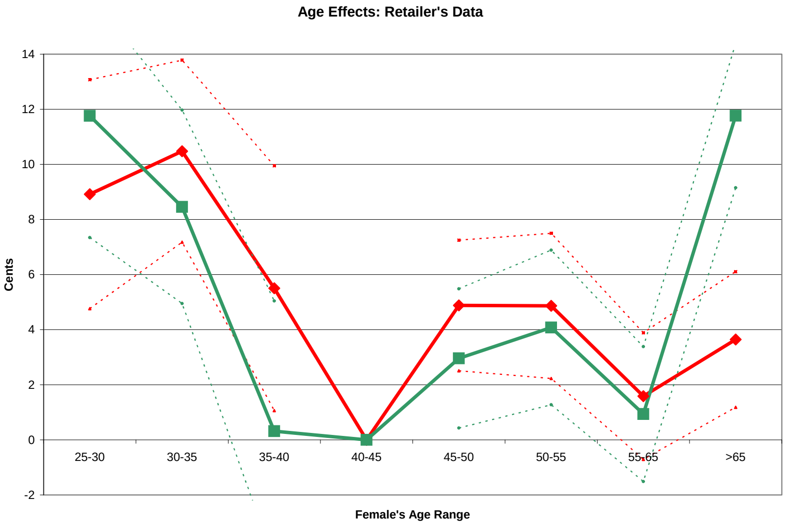
Using the validation sample to control for recording error

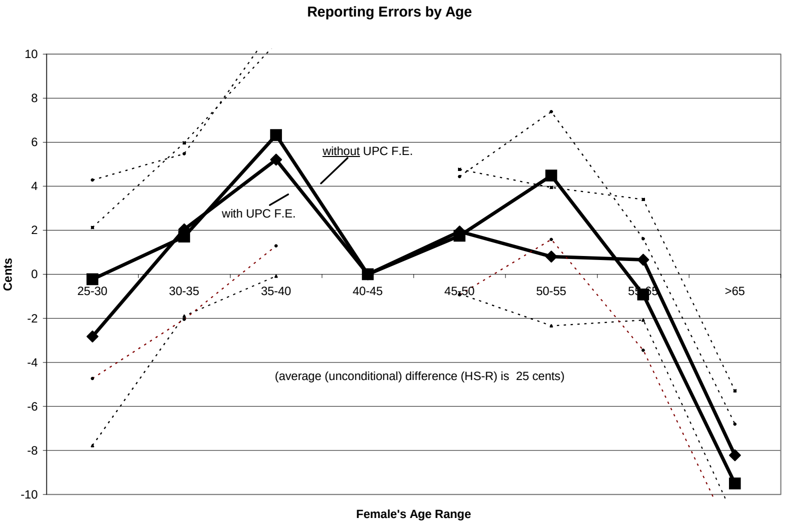
- In a linear model this simplifies to a fairly simple procedure.
- Suppose we want to correlate price paid to demographics
  - In validation sample - regress R price (the "true" price) on HS price and demographics;
  - In primary sample - compute the predicted price;
  - In primary sample - regress the predicted price on demographics
- Why not just use the validation sample?
  - Efficiency
  - Some variables might only be observed in the primary sample (e.g., store choice)
  - Same data but different coverage (markets, years)

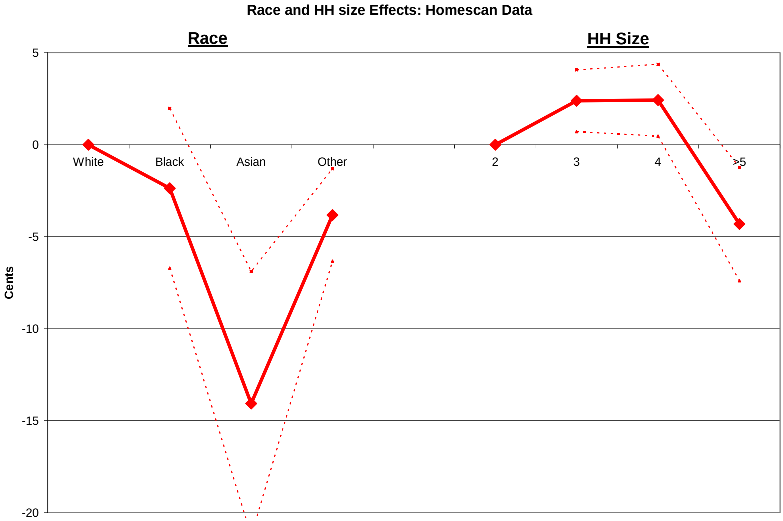
Implications: Do the discrepancies matter?

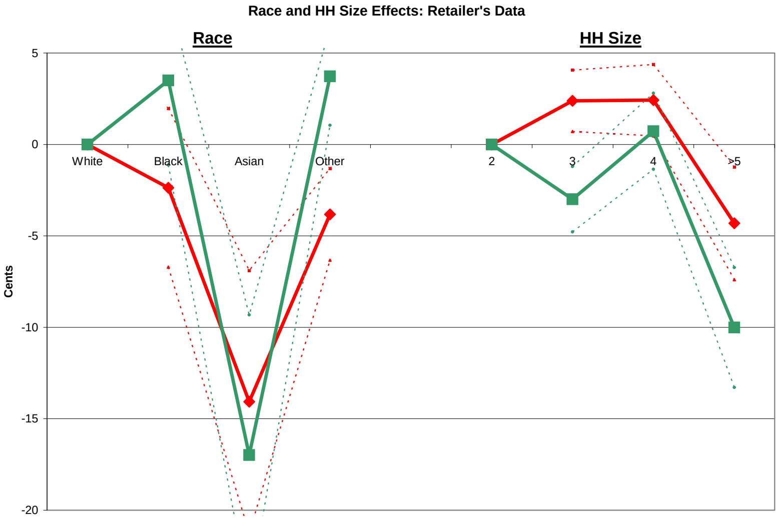
Dep. Var	Price HS		Price R		Same sign	Coef. ratio	Same stat. sig.
	Coef.	t-stat	Coef.	t-stat			
_cons	291.04	35.94	294.54	34.35	yes	0.99	yes
HH size	-1.67	-3.70	-3.55	-7.45	yes	0.47	yes
HH income	0.04	3.17	0.10	7.56	yes	0.40	yes
No female head of HH	-29.92	-4.15	-36.40	-4.77	yes	0.82	yes
Age female	-0.64	-2.31	-1.73	-5.87	yes	0.37	yes
Age female ^ 2	0.00	1.72	0.02	7.03	yes	0.23	no
No male head of HH	-0.27	-0.04	-30.45	-3.74	yes	0.01	no
Age male	-0.27	-0.89	-1.13	-3.57	yes	0.24	no
Age male ^ 2	0.00	0.75	0.01	3.42	yes	0.21	no
No. of kids	1.19	1.08	3.26	2.78	yes	0.37	no
No. of Little kids	-0.24	-0.15	4.24	2.55	no	NA	no
Male employed	-0.14	-0.08	-8.56	-4.76	yes	0.02	no
Male fully employed	-0.15	-0.09	14.63	8.66	no	NA	no
Female employed	1.20	1.26	0.96	0.95	yes	1.25	yes
Female fully employed	-3.58	-3.78	-3.49	-3.48	yes	1.03	yes
Male education	0.36	1.03	-1.76	-4.81	no	NA	no
Female education	-1.95	-5.20	1.02	2.57	no	NA	yes
Married	-3.91	-4.11	-2.07	-2.06	yes	1.89	yes
Non-white	-3.01	-2.43	1.45	1.10	no	NA	no
Hispanic	-1.28	-0.87	-1.64	-1.05	yes	0.78	yes
"15K" HH	-1.28	-1.14	-2.21	-1.85	yes	0.58	yes
UPC fixed effects	yes		yes				
Obs.	50,600		50,600				

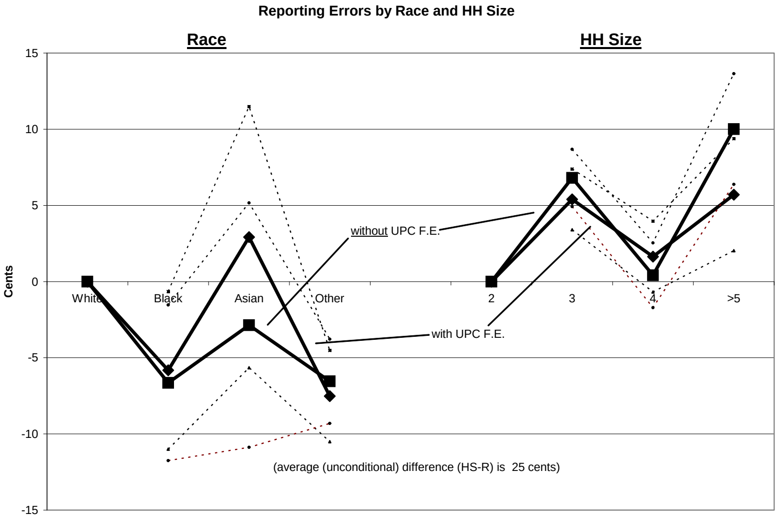

Age Effects: Homescan Data
Cents
14
12
10
8
6
4
2
0
-2
25-30
30-35
35-40
40-45
45-50
50-55
55-65
>65
Female's Age Range


Age Effects: Retailer's Data
Cents
14
12
10
8
6
4
2
0
-2
25-30
30-35
35-40
40-45
45-50
50-55
55-65
>65
Female's Age Range


Reporting Errors by Age
10
8
6
4
2
0
-2
-4
-6
-8
-10
Cents
25-30
30-35
35-40
40-45
45-50
50-55
55-65
>65
without UPC F.E.
with UPC F.E.
(average (unconditional) difference (HS-R) is 25 cents)
Female's Age Range


Race and HH size Effects: Homescan Data
Race
HH Size
Cents
5
0
-5
-10
-15
-20
White
Black
Asian
Other
2
3
4
>5


Race and HH Size Effects: Retailer's Data
Race
HH Size
Cents
5
0
-5
-10
-15
-20
White
Black
Asian
Other
2
3
4
>5


Reporting Errors by Race and HH Size
Race
HH Size
15
10
5
0
-5
-10
-15
Cents
White
Black
Asian
Other
2
3
4
>5
without UPC F.E.
with UPC F.E.
(average (unconditional) difference (HS-R) is 25 cents)

Why should the error vary by demographic group?

- Recall 2 types of error in price
  - price imputation
  - recording error
- As we saw recording error varies with demographics;
- However, the data suggests that for the price variable imputation is a key source of error;
- A simple story
  - 2 types of consumers O and Y;
  - O all use loyalty cards and shop in store A;
  - Only 50% of O use loyalty cards and shop in store B;
  - Selection into HS: HS panelist always use card;
- Price imputation creates a wedge between price paid by consumer and average price in the store;
- ⇒ Imputed price and actual price the same for O, different for Y;
- (according to this story) *Within-group* selection into becoming an HS panelist is key;

Implications: choice models

- Suppose we want to estimate store/product choice;
- The impact of the recording error is more complicated:
  - non-linear model;
  - error in both choice and price data;
- In principle, can use the same procedure
- Application ... coming

Summary

- Homescan data have recording errors, which correlate with other variables
    - Unclear that the Homescan data is more prone to error than other economic data sets
- Errors are important and can impact findings
- Robustness: use the validation data, “correct” the estimates, and assess differences.
- What next?
    - Further implications
    - Use the additional information for estimation