Data Imputation Models for Non-trended Price Data

Kennon Copeland[1], Boris Brodsky[2]
[1]NORC at the University of Chicago, 1350 Connecticut Ave, NW, Washington, DC, 20036
[2]IMS Government Solutions

Abstract

Missing data is a common problem for sample surveys that, if ignored, results in increased variance and likely bias for survey estimates. Data values based upon small samples or that have large error may also adversely affect the variance for survey estimates. Imputation, in which values are assigned for missing and "weak" data, is one approach commonly taken to reduce the variance and bias for survey estimates. Imputation for price data is typically carried out through the use of models reflecting change in prices over time. However, if no historical data are available (e.g., initial collection, for rare items), alternative models are required. Approaches for modeling prescription drug prices using related prices from the sample data will be presented, performance of the models will be reviewed, and implications for future direction will be discussed.

Keywords: Imputation, Proportional Model, Price Data, Composite Estimator

1. Overview

Missing data is a problem that affects virtually all surveys. Missing data results in increased variance for survey estimates due to reduced sample sizes, and yields the potential for bias in survey estimates. In addition, small sample sizes for selected items of interest can also adversely affect the variance and accuracy of survey estimates.

Price data for retail prescriptions provide an extensive array of missing and small sample data issues. When attempting to estimate prescription prices at the detailed product/form/strength/package size level, missing data are not uncommon, and small sample sizes are prevalent. Developing best estimates in these situations is an important methodological question for understating prescription price.

Imputation is a common approach to dealing with missing survey data (Kalton and Kasprzyk, 1982). For price data, imputation models commonly make use of change over time. For example, a simple model employed is to assume the price rate of change for the item with the missing price at time t is the same as that for some other item or set of items for which prices are available for time t (Armknecht and Maitland-Smith, 1999).

These change over time models assume historical data for the item with missing price at time t are available, and that the change model provides a reasonable fit. For new items and for items with small volumes, however, historical data may not be available. In addition, the change model may not provide adequate fit when looking across items.

This research was carried out in support of development of appropriate prescription price estimates for all pharmaceutical products. We provide a description of the data source and the missing data/small sample size problem in Section 2, explore imputation models in Section 3, describe the imputation approach in Section 4, present results in Section 5, and discuss implications in Section 6.

2. Description of Data Source

IMS obtains prescription information on a weekly basis from over 35,000 retail pharmacies nationwide. This sample represents approximately 67% of retail pharmacies and 73% of retail prescription volume, and is geographically spread throughout the U.S. The reporting week is Saturday through Friday. Prescription information provided to IMS is that recorded within pharmacy software systems as part of regular prescription management conducted by pharmacies. Thus, there is an incentive for complete and accurate reporting by pharmacies.

All types of prescriptions (Cash, Medicaid, 3rd Party) are reported, and various data elements are reported for each prescription (e.g., date, NDC11, quantity dispensed, price, method of payment). Among the uses for the prescription data are estimating measures related to the dispensed price (e.g., total price for dispensed prescriptions, average price per quantity).

Issues to be dealt with in developing appropriate estimates including missing price data for an NDC11 (which identifies manufacturer, product, form, strength, and package size for a pharmaceutical product) as well as small number of prescriptions for an NDC11.

Prescription data for November 2006 through May 2007 were extracted from the prescription database, along with relevant data elements, for the purpose of developing an approach for imputing price for missing and small sample NDC11's.

Estimation cells were created based upon prior analysis of price differences and variability, resulting in six cells: Method of Payment (Cash, Medicaid, 3rd Party) and Channel (Retail, Mail Order). Roughly 60% of dispensed prescriptions have a 3rd Party method of payment. (See Table 1.) The Mail Order channel accounts for only about 13% of dispensed prescriptions, with the vast majority (~95%) of Mail Order prescriptions having a 3rd Party method of payment.

Missing prices are common for the Cash Mail Order estimation cell (47%). For the remaining cells, missing prices occur for 5% to 15% of the NDC11's. (See Table 2.) Small sample sizes are extensive across estimation cells. The median number of sample observations for an NDC11 is less than 20 for all estimation cells. Both the missing prices and the small sample sizes are due primarily to small total dispensed prescriptions when looking at the estimation cell by NDC11 level.

3. Model Exploration

3.1 Model Development

In order to develop a model, we sought to leverage relationships between estimation cells. The underlying hypothesis was that the ratio of NDC11 prices for a pair of estimation cells was relatively stable across NDC11's. In terms of setting up an imputation model, this can be viewed in terms of a target cell (the estimation cell with a missing or small sample price for the NDC11) and donor cell (the estimation cell containing a usable price for the NDC11), as represented by

$$X_{NDC11,ij} = R_{T,ij}^{i'j'} * X_{NDC11,i'j'} + e_{NDC11,ij}^{i'j'}$$

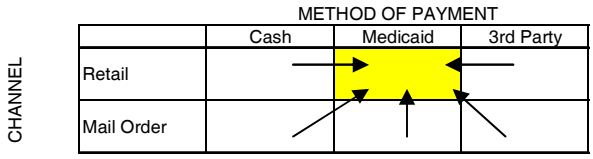
where

$X_{NDC11,ij}$ = cell ij average price for an NDC11

$R_{T,ij}^{i'j'}$ = RxD ratio of average prices of cell ij to cell $i'j'$ $\left(i,i' = 1,2,3;\ j,j' = 1,2; i'j' \neq ij\right)$ for NDCs of Type T (Brand, Generic)

$e_{NDC11,ij}^{i'j'}$ = error in the model for paired cells ij and $i'j'$ for an NDC

As an illustration, when attempting to develop imputed values for target cell Medicaid/Retail, ratio estimates could be developed using the remaining estimation cells as shown below.



In particular, the ratio estimate for target cell Medicaid/Retail based on donor cell Cash/Retail would be

$R_{Medicaid,Retail}^{Cash,Retail}$ = ratio of Medicaid/Retail price to Cash/Medicaid price

3.2 Findings

3.2.1 Ratio Estimates by Product Type

Prescription price data from the November, 2006 data month were examined to determine if additional variables needed to be accounted for in the model. As alluded to in section 3.1, ratios differed by product type (Brand, Generic). (See Table 3.)

In addition, it can be seen in Table 3 that the ratio estimates derived for Brand NDC11's are more stable than those for Generic NDC11's, another justification for deriving estimated ratios separately by product type.

3.2.2 Stability Across Sample Size

Prescription price data from the November, 2006 data month were examined to assess the assumption that ratio estimates derived from NDC11's with large sample sizes would be appropriate for use in imputing prices for NDC'11's with small sample sizes. As illustrated in Figure 1 (which is for the Cash/Retail target cell and 3rd Party/Retail donor cell), the ratio of prices in the target cell to those in the donor cell appears stable across different sample sizes, based upon the median ratio.

Based upon this analysis, it was determined that it would be appropriate to estimate ratios based upon NDC11's with larger sample sizes and apply the ratios to NDC11's with small or missing sample.

3.2.3 Stability Across Time

A concern with use of the proportional model was that estimated ratios for a given pair of cells could fluctuate across time, leading to the potential for variability across time for the imputed prices. Data from November, 2006, to May, 2007, data months were examined.

As seen in Table 4, estimated ratios for Brand NDC11's appear very stable across time. The least stable ratios were those involving Cash/Mail as either a target cell or a donor cell, although the standard deviation of the estimated ratios across months was 0.002 or less.

The estimated ratios for Generic NDC11's, presented in Table 5, show less stability than for Brand; however the ratios are still very stable with the exception of ratios involving Cash/Mail as either the target cell or the donor cell.

4. Imputation Approach

As described in section 3.1, each target cell could have up to 5 ratio estimates for price. The number of available ratio estimates for price will depend upon the sample sizes available within the donor cells. A minimum required sample size was established for a donor cell to be used, with the estimated price and standard deviation of price for the target cell defined as

$$\hat{X}_{NDC11,ij}^{i'j} = \hat{R}_{T,ij}^{i'j} \hat{X}_{RxD,NDC11,i'j}$$

$$sd\left(\hat{X}^{i'j'}_{NDC11,ij}\right) = \sqrt{\left[\hat{X}_{NDC11,i'j'} sd\left(\hat{R}^{i'j'}_{T,ij}\right)\right]^2 + \left[\hat{R}^{i'j'}_{T,ij} sd\left(\hat{X}_{NDC11,i'j'}\right)\right]^2}$$

In addition, the target cell will have an empirical price based upon available sample, unless the price is missing. The empirical price was used in developing the imputed price.

A composite estimator can be defined using all available ratio estimates for the target cell along with the empirical estimate, if available. The composite estimator and standard deviation of the composite estimator can be shown to be

$$\hat{X}^{C}_{NDC11,ij} = \frac{\sum_k \left(\hat{X}^{k}_{NDC1,ij} \prod_{k' \neq k} \left[sd\left(\hat{X}^{k'}_{NDC1,ij}\right)\right]^2 \right)}{\sum_k \left(\prod_{k' \neq k} \left[sd\left(\hat{X}^{k'}_{NDC1,ij}\right)\right]^2 \right)}$$

$$sd\left(\hat{X}^{C}_{NDC11,ij}\right) = \sqrt{\frac{\prod_k \left[sd\left(\hat{X}^{k}_{NDC1,ij}\right)\right]^2}{\sum_k \left(\prod_{k' \neq k} \left[sd\left(\hat{X}^{k'}_{NDC1,ij}\right)\right]^2 \right)}}$$

where the sum is across all available ratio estimates and empirical estimate.

This composite estimator is a straightforward extension of the commonly used composite estimator based upon two independent estimates (Schaible, 1978). This estimator provides greater weights to ratio estimates/empirical estimates with smaller variability.

5. Results

Data from the November, 2006, data month were used to assess the performance of the imputation model and approach.

Figure 2 presents the relative deviation between the empirical and composite estimated prices for Brand NDC11's in the Retail channel plotted against sample size. There appears to be good correspondence between the empirical and composite estimates, especially for 3rd Party method of payment and larger sample sizes. For over 95% of all NDC11's there was less than a 10% relative deviation between the empirical and composite estimated prices.

Generic NDC11's showed greater variability, as expected from the analysis of the imputation model. (See Figure 3.) Even so, for over 70% of all NDC11's there was less than a 10% relative deviation between the empirical and composite estimated prices, which suggests the model would perform well for NDC11's with small and missing data.

Figures 4 and 5 show the extent of reduction in standard deviation of estimated price achieved by the composite estimated price relative to the empirical price. The reduction is greater the smaller the sample size, which is where the imputation would be applied. Reductions for Brand NDC11's tended to greatest for Cash method of payment (roughly 50% on average) and less for Medicaid (~20% on average) and 3rd Party (~5% on average), whereas reductions for Generic NDC11's tended to be equivalent regardless of method of payment (roughly 20% on average)..

6. Summary

This research has resulted in the development of a usable imputation approach for prescription price data through a proportional model, providing price estimates for missing and small sample NDC11's. The application of a composite estimator allows use of all available data of sufficient sample size, and reduces the standard deviation of the resultant price estimates.

References

Armknecht, P.A. and Maitland-Smith, F. (1999). "Price Imputation and Other Techniques for Dealing with Missing Observations, Seasonality and Quality Change in Price Indices," *Working Paper of the International Monetary Fund*.

Kalton, G. and Kasprzyk, D. (1986). "Imputing for Missing Survey Response," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 194-199.

Schaible, W.L. (1978). "Choosing Weights for Composite Estimators for Small Area Statistics," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

.

Table 1

Rx Distribution
February, 2007

Channel	Method of Payment		
	Cash	Medicaid	3rd Party
Retail	26.5%	13.5%	47.2%
Mail Order	0.5%	0.1%	12.2%

Table 2

Profile of Rx Transaction Sizes by Cell
February, 2007

MOP	Channel	# of NDC11's	% Missing	Percentiles for Rx Transactions						
				5th	10th	25th	Median	75th	90th	95th
Cash	Retail	41,096	14.8%	1	1	2	7	57	381	1,081
	Mail Order	6,321	47.4%	1	1	1	2	4	12	27
Medicaid	Retail	26,198	9.6%	1	1	3	15	120	618	1,443
	Mail Order	4,414	6.3%	1	1	1	3	7	20	39
3rd Party	Retail	40,196	11.6%	1	1	2	17	306	3,395	10,254
	Mail Order	14,114	13.2%	1	1	3	16	119	647	1,647

Table 3

Comparison of Across-Cell Ratio Estimates
Brand vs. Generic NDC11's
Nov '06

Target Cell		Donor Cell		Ratio		Stdev	
MOP	Channel	MOP	Channel	Brand	Generic	Brand	Generic
Cash	Retail	Medicaid	Retail	1.19	1.25	0.093	0.548
		3rd Party	Retail	1.25	1.46	0.109	0.597
		Cash	Mail	1.22	1.62	0.184	1.017
		Medicaid	Mail	1.19	1.16	0.178	1.279
		3rd Party	Mail	1.31	1.48	0.191	0.882
Medicaid	Retail	Cash	Retail	0.84	0.80	0.060	0.308
		3rd Party	Retail	1.04	1.19	0.055	0.325
		Cash	Mail	1.07	1.26	0.193	0.769
		Medicaid	Mail	1.01	0.92	0.209	1.194
		3rd Party	Mail	1.09	1.17	0.155	0.618
3rd Party	Retail	Cash	Retail	0.80	0.69	0.082	0.303
		Medicaid	Retail	0.96	0.84	0.052	0.206
		Cash	Mail	1.04	1.01	0.151	0.496
		Medicaid	Mail	0.98	0.75	0.202	1.007
		3rd Party	Mail	1.05	1.04	0.123	0.465
Cash	Mail	Cash	Retail	0.82	0.62	0.114	0.431
		Medicaid	Retail	0.93	0.79	0.458	0.593
		3rd Party	Retail	0.96	0.99	0.126	0.657
		Medicaid	Mail	0.92	0.91	0.150	0.540
		3rd Party	Mail	1.04	1.03	0.424	0.639
Medicaid	Mail	Cash	Retail	0.84	0.86	0.089	0.478
		Medicaid	Retail	0.99	1.09	0.109	0.445
		3rd Party	Retail	1.02	1.33	0.106	0.750
		Cash	Mail	1.09	1.10	0.129	1.020
		3rd Party	Mail	1.05	1.26	0.123	1.638
3rd Party	Mail	Cash	Retail	0.76	0.67	0.081	0.438
		Medicaid	Retail	0.91	0.85	0.087	0.871
		3rd Party	Retail	0.95	0.96	0.075	1.006
		Cash	Mail	0.96	0.97	0.172	0.907
		Medicaid	Mail	0.95	0.79	0.202	1.132

Table 4

Stability of Across-Cell Ratio Estimates
Brand NDC11's
Nov '06 - May '07

Target Cell		Donor Cell		Month								
MOP	Channel	MOP	Channel	Nov	Dec	Jan	Feb	Mar	Apr	May	Ave	sd
Cash	Retail	Medicaid	Retail	1.19	1.20	1.20	1.20	1.20	1.19	1.19	1.20	0.004
		3rd Party	Retail	1.25	1.25	1.25	1.25	1.25	1.25	1.25	1.25	0.001
		Cash	Mail	1.22	1.20	1.19	1.21	1.21	1.23	1.22	1.21	0.014
		Medicaid	Mail	1.19	1.19	1.19	1.19	1.19	1.18	1.18	1.19	0.004
		3rd Party	Mail	1.31	1.31	1.32	1.33	1.32	1.33	1.33	1.32	0.008
Medicaid	Retail	Cash	Retail	0.84	0.84	0.83	0.83	0.84	0.84	0.84	0.84	0.002
		3rd Party	Retail	1.04	1.04	1.03	1.03	1.04	1.04	1.04	1.04	0.003
		Cash	Mail	1.07	1.07	1.09	1.08	1.08	1.11	1.12	1.09	0.019
		Medicaid	Mail	1.01	1.00	1.00	1.01	1.01	1.01	1.01	1.01	0.001
		3rd Party	Mail	1.09	1.09	1.09	1.09	1.10	1.10	1.10	1.09	0.006
3rd Party	Retail	Cash	Retail	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.001
		Medicaid	Retail	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.97	0.002
		Cash	Mail	1.04	1.04	1.06	1.06	1.06	1.06	1.07	1.06	0.013
		Medicaid	Mail	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.001
		3rd Party	Mail	1.05	1.05	1.05	1.06	1.06	1.06	1.06	1.06	0.004
Cash	Mail	Cash	Retail	0.82	0.83	0.82	0.82	0.83	0.82	0.82	0.82	0.007
		Medicaid	Retail	0.93	0.93	0.92	0.92	0.92	0.90	0.89	0.92	0.015
		3rd Party	Retail	0.96	0.96	0.94	0.94	0.92	0.93	0.92	0.94	0.017
		Medicaid	Mail	0.91	0.95	0.93	0.92	0.92	0.90	0.90	0.92	0.019
		3rd Party	Mail	1.03	1.03	1.01	1.03	1.07	1.05	1.06	1.04	0.022
Medicaid	Mail	Cash	Retail	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.003
		Medicaid	Retail	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.001
		3rd Party	Retail	1.02	1.01	1.02	1.02	1.02	1.02	1.02	1.02	0.001
		Cash	Mail	1.08	1.05	1.08	1.04	1.07	1.11	1.05	1.07	0.023
		3rd Party	Mail	1.05	1.04	1.05	1.05	1.05	1.05	1.04	1.05	0.005
3rd Party	Mail	Cash	Retail	0.76	0.76	0.76	0.75	0.76	0.75	0.75	0.76	0.005
		Medicaid	Retail	0.91	0.92	0.92	0.91	0.91	0.91	0.91	0.91	0.005
		3rd Party	Retail	0.95	0.95	0.95	0.94	0.95	0.94	0.95	0.95	0.003
		Cash	Mail	0.97	0.97	0.99	0.97	0.92	0.92	0.91	0.95	0.032
		Medicaid	Mail	0.95	0.96	0.96	0.95	0.95	0.95	0.96	0.95	0.005

Table 5

Stability of Across-Cell Ratio Estimates
Generic NDC11's
Nov '06 - Feb '07

Target Cell		Donor Cell		Month								
MOP	Channel	MOP	Channel	Nov	Dec	Jan	Feb	Mar	Apr	May	Ave	sd
Cash	Retail	Medicaid	Retail	1.25	1.24	1.24	1.25	1.25	1.24	1.25	1.25	0.003
		3rd Party	Retail	1.45	1.46	1.43	1.46	1.45	1.44	1.45	1.45	0.011
		Cash	Mail	1.62	1.70	2.00	1.97	1.79	1.76	1.78	1.80	0.137
		Medicaid	Mail	1.19	1.23	1.21	1.19	1.16	1.25	1.22	1.21	0.030
		3rd Party	Mail	1.45	1.45	1.43	1.46	1.47	1.46	1.46	1.45	0.012
Medicaid	Retail	Cash	Retail	0.80	0.80	0.81	0.80	0.80	0.80	0.80	0.80	0.002
		3rd Party	Retail	1.18	1.19	1.17	1.18	1.18	1.17	1.18	1.18	0.006
		Cash	Mail	1.26	1.42	1.72	1.68	1.55	1.53	1.55	1.53	0.154
		Medicaid	Mail	0.92	0.93	0.99	0.91	0.89	0.94	0.96	0.93	0.031
		3rd Party	Mail	1.15	1.15	1.15	1.16	1.15	1.15	1.15	1.15	0.004
3rd Party	Retail	Cash	Retail	0.69	0.69	0.70	0.69	0.69	0.69	0.69	0.69	0.005
		Medicaid	Retail	0.84	0.84	0.85	0.84	0.85	0.85	0.85	0.85	0.004
		Cash	Mail	1.02	1.09	1.30	1.31	1.23	1.27	1.25	1.21	0.112
		Medicaid	Mail	0.76	0.77	0.82	0.74	0.75	0.81	0.85	0.78	0.044
		3rd Party	Mail	1.02	1.02	1.04	1.02	1.02	1.02	1.02	1.02	0.008
Cash	Mail	Cash	Retail	0.62	0.59	0.50	0.51	0.56	0.57	0.56	0.56	0.042
		Medicaid	Retail	0.79	0.70	0.58	0.60	0.64	0.66	0.65	0.66	0.071
		3rd Party	Retail	0.99	0.91	0.77	0.77	0.81	0.79	0.80	0.83	0.083
		Medicaid	Mail	0.91	0.94	0.90	0.90	0.92	0.82	0.78	0.88	0.057
		3rd Party	Mail	1.01	0.98	0.88	0.85	0.94	0.85	0.87	0.91	0.064
Medicaid	Mail	Cash	Retail	0.84	0.81	0.83	0.84	0.85	0.80	0.82	0.83	0.018
		Medicaid	Retail	1.09	1.08	1.01	1.09	1.11	1.07	1.04	1.07	0.034
		3rd Party	Retail	1.31	1.31	1.22	1.34	1.34	1.24	1.17	1.28	0.066
		Cash	Mail	1.08	1.03	1.09	1.07	1.09	1.21	1.28	1.12	0.089
		3rd Party	Mail	1.23	1.21	1.19	1.26	1.29	1.17	1.16	1.22	0.046
3rd Party	Mail	Cash	Retail	0.69	0.69	0.70	0.68	0.68	0.69	0.69	0.69	0.006
		Medicaid	Retail	0.87	0.87	0.87	0.86	0.87	0.87	0.87	0.87	0.003
		3rd Party	Retail	0.98	0.98	0.96	0.98	0.98	0.98	0.98	0.98	0.008
		Cash	Mail	0.99	1.02	1.14	1.18	1.07	1.17	1.15	1.10	0.075
		Medicaid	Mail	0.81	0.82	0.84	0.78	0.77	0.85	0.85	0.82	0.031

Figure 1

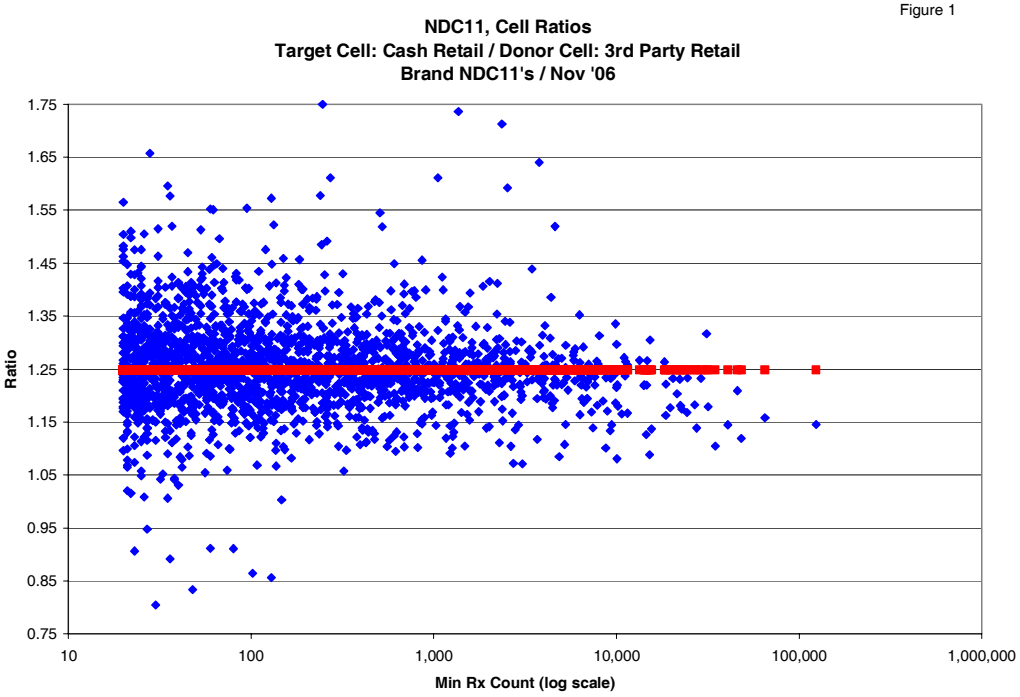


Figure 2

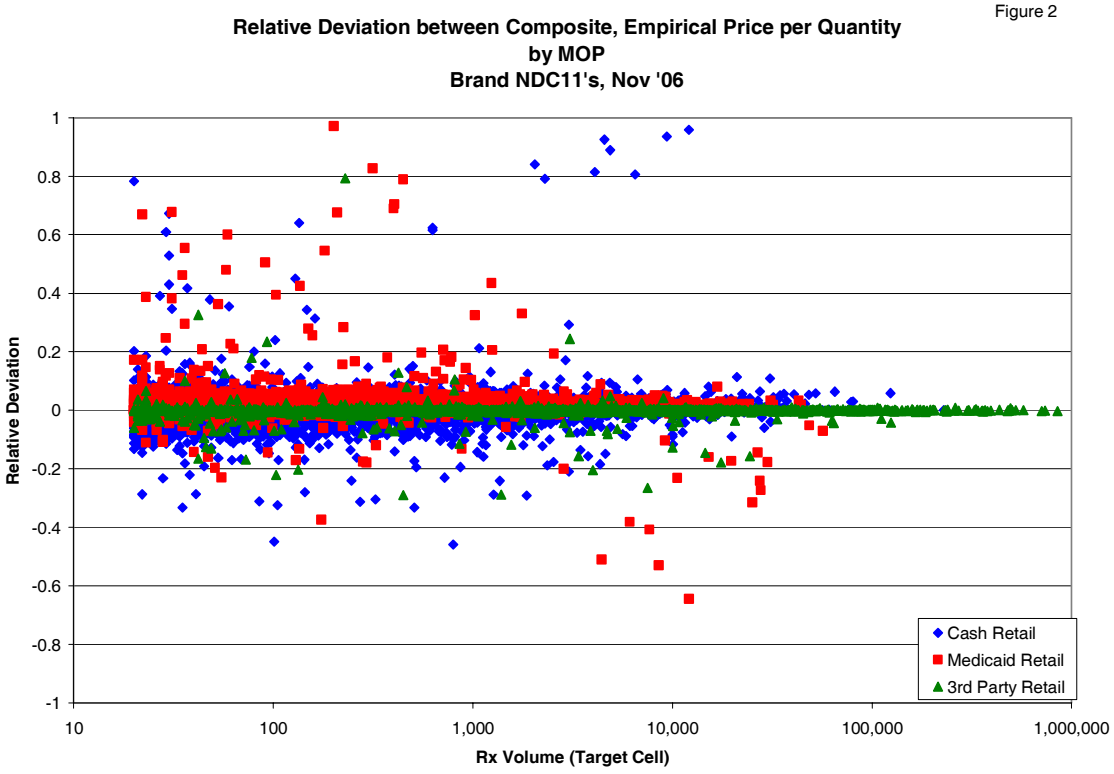


Figure 3

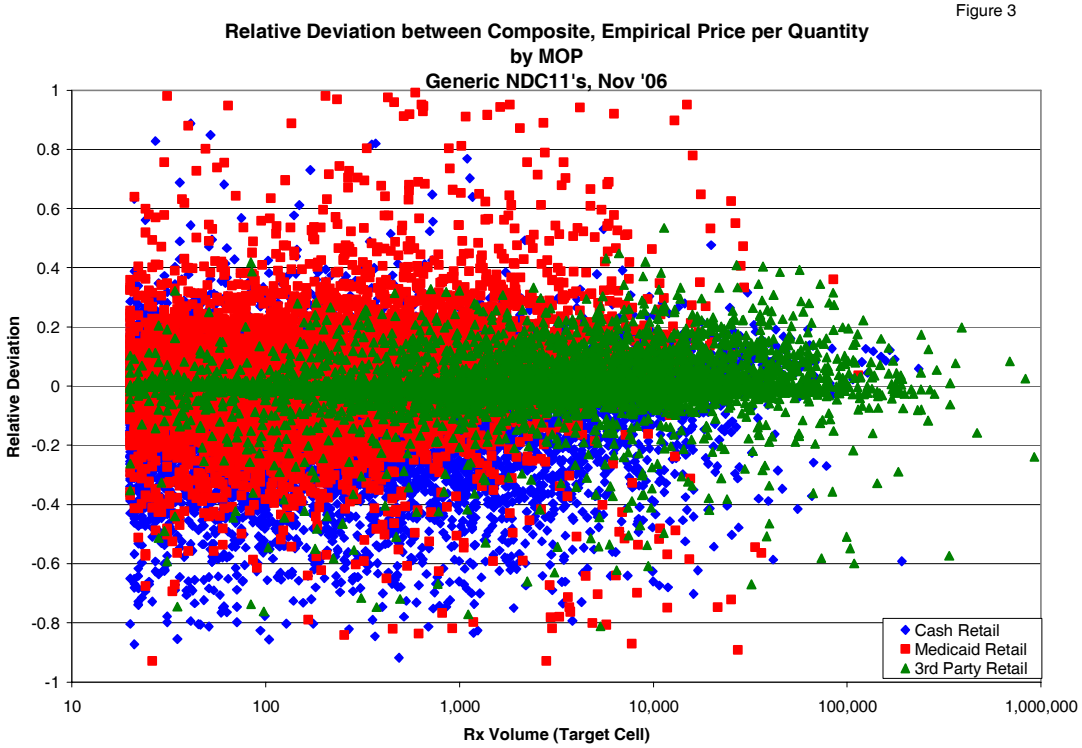


Figure 4

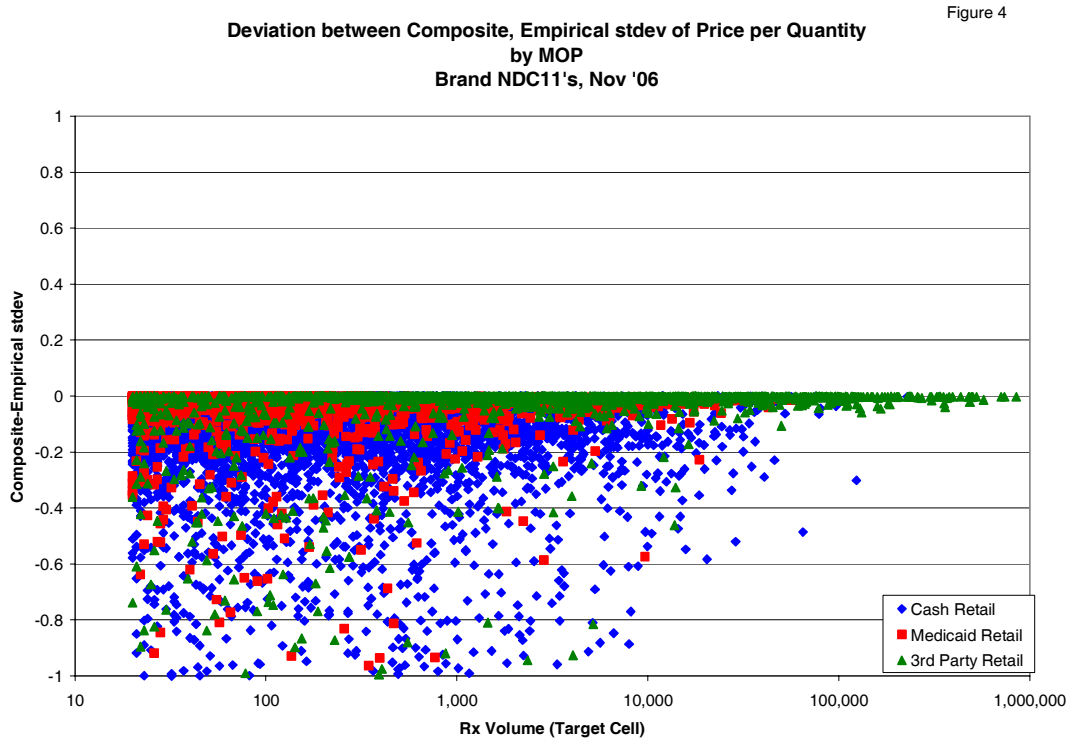
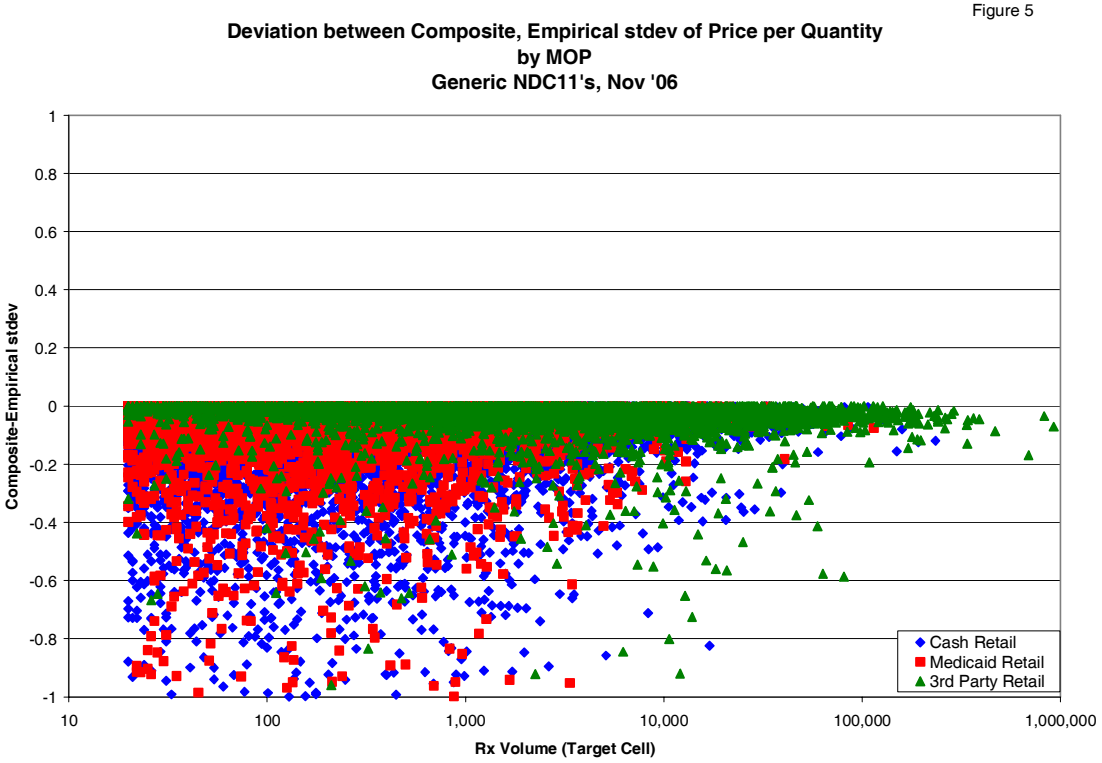


Figure 5


Deviation between Composite, Empirical stdev of Price per Quantity
by MOP
Generic NDC11's, Nov '06
Composite-Empirical stdev
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
10
100
1,000
10,000
100,000
1,000,000
Rx Volume (Target Cell)
Cash Retail
Medicaid Retail
3rd Party Retail