Using a Two-Part Markov Latent Class Model to Examine Expenditure Report Quality

Brian Meekins
Bureau of Labor Statistics
N. Clyde Tucker
American Institutes for Research

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Latent Class Analysis

- Uses repeated measurements from panel survey data to estimate classification error
- Does not require external validation data; estimates of error directly from panel data
- LCA used to study measurement or response error (VandePol and deLeeuw 1986; Tucker 1992; Van de Pol and Langeheine 1997; Bassi et al. 2000; Biemer and Bushery 2000; Tucker, et al. 2002, 2003, 2004, 2005, 2006, and 2008); Meekins et al. (2011)

BLS

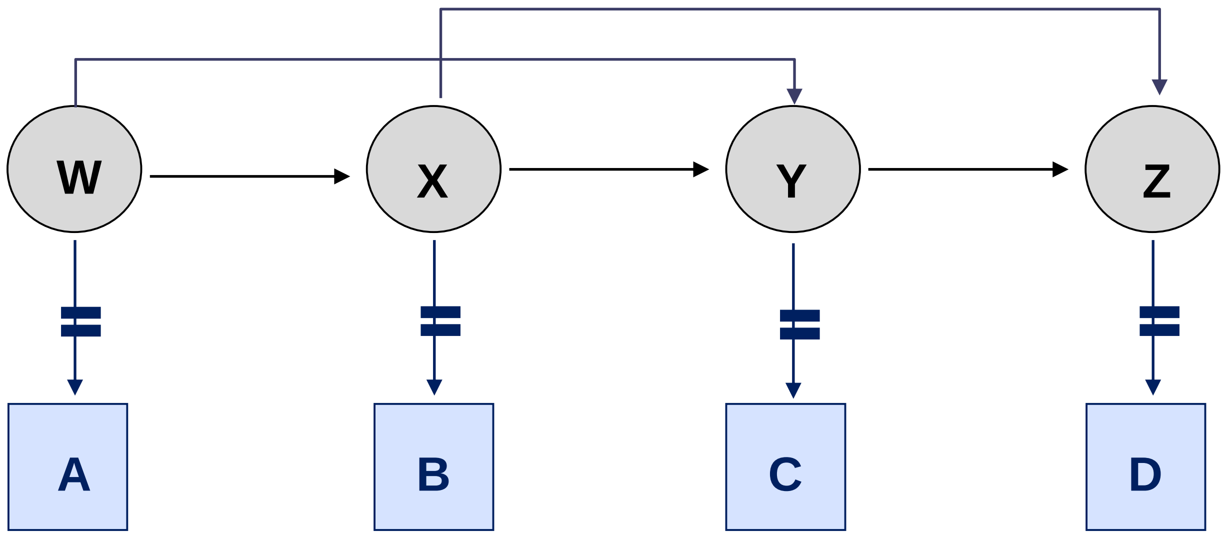
U.S. Consumer Expenditure Interview Survey (CEIS)

- ~ 6,000 CU’s/year
- CU’s interviewed every 3 months about prior 3 months expenditures
- 4 consecutive interviews on each CU
- 15 years of CEIS: 1996-2010
- Unweighted analysis
- 31 commodity categories analyzed

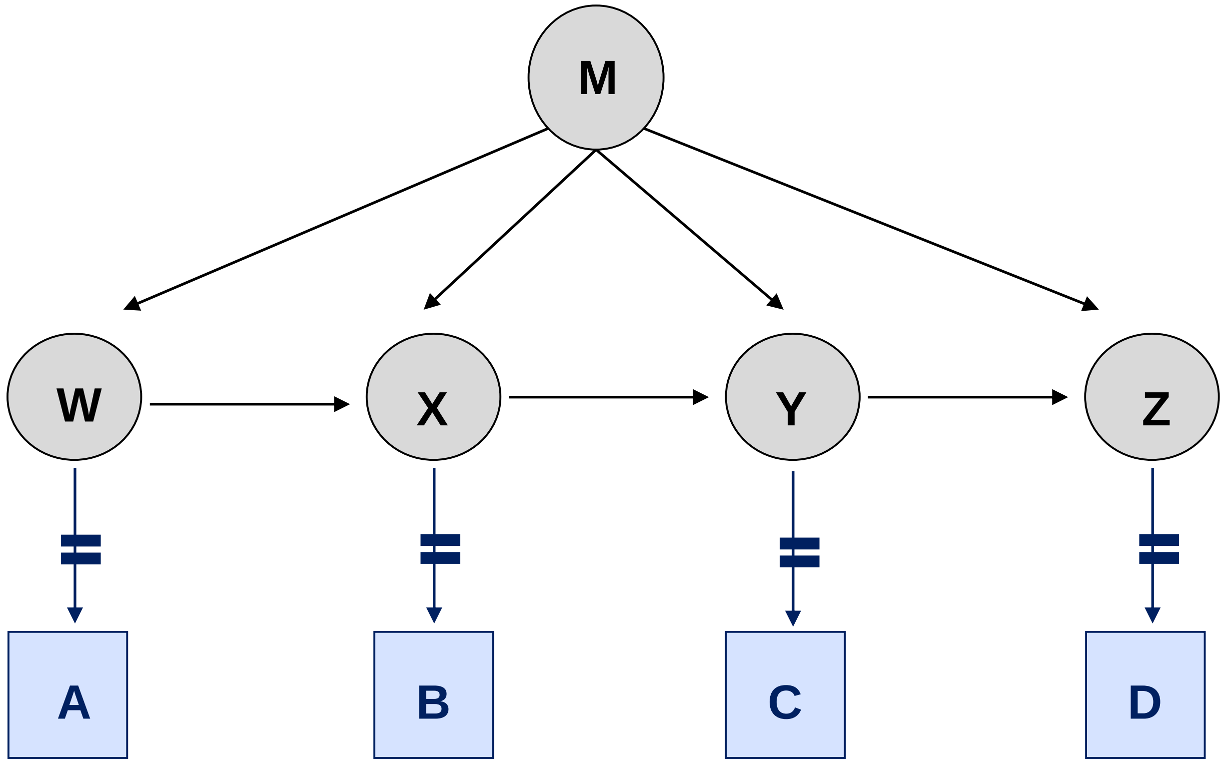
Commodity Categories

- Dental
- Prescription drugs
- Eye care
- Clothing
- Infant clothing
- Clothing accessories
- Clothing services
- Sewing
- Shoes
- Jewelry
- Events (e.g. sporting/theatre)
- Computer games
- Computer equipment
- Books
- Cable
- Music
- Internet (2001+)
- Sports equipment
- Major appliances
- Minor appliances
- Electronics
- Childcare
- Pets and pet supplies
- Major Vehicle Repairs
- Minor Vehicle Repairs
- License/registration
- HH electricity
- HH gas
- HH trash service
- Phone
- HH services

2^{nd} Order Markov



Mover-Stayer



$$M = \begin{cases} 1, & P(W), P(X), P(Y), \text{ and } P(Z) \text{ are unconstrained.} \\ 2, & P(W=1) = P(X=1) = P(Y=1) = P(Z=1) = 1 \\ 3, & P(W=1) = P(X=1) = P(Y=1) = P(Z=1) = 0 \end{cases}$$

Model Assumptions

- Markov or Mover-Stayer model assumptions
- Equal measurement error across all interviews

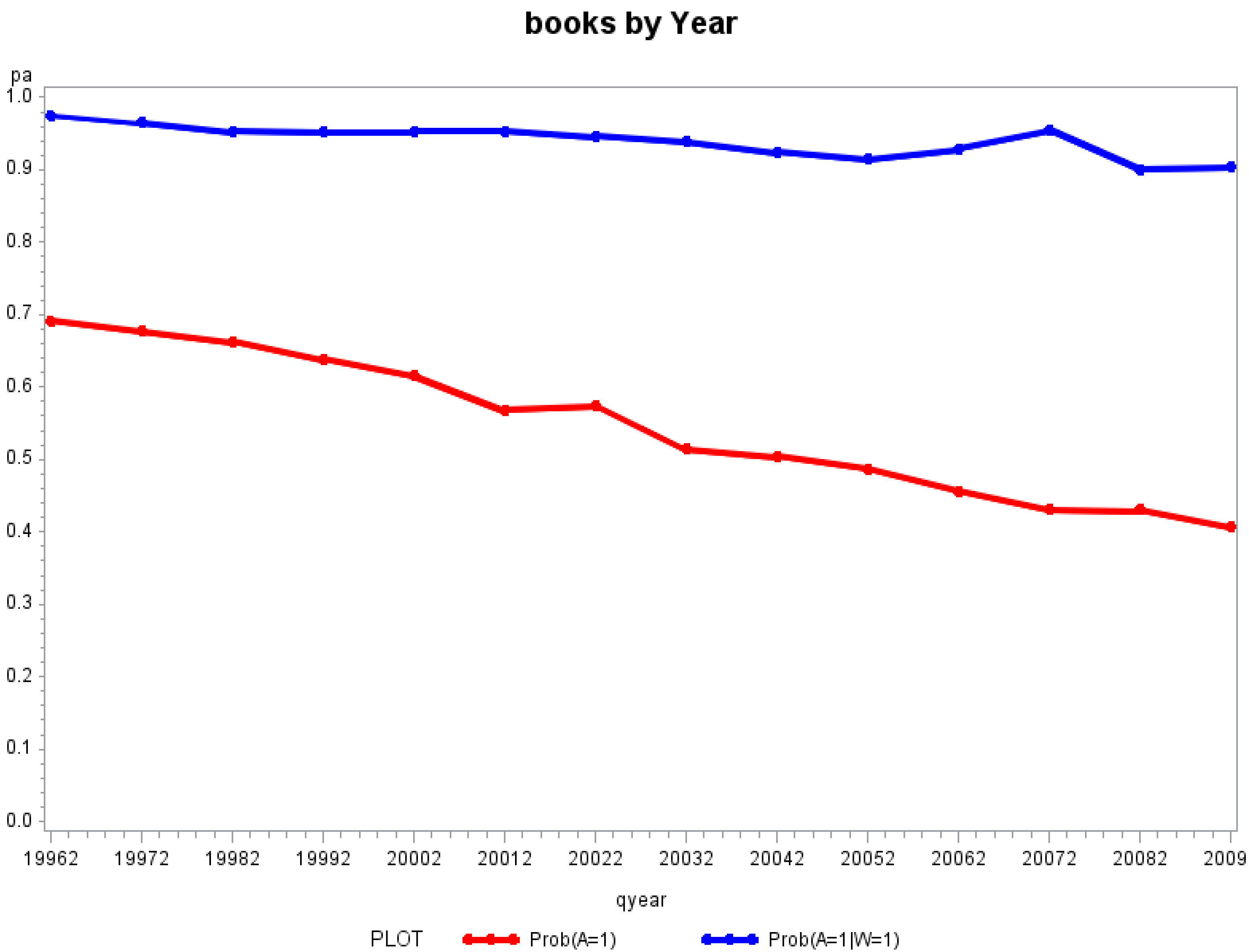
$$P(a_i = j \mid w_i = k) = P(b_i = j \mid x_i = k)$$
$$= P(c_i = j \mid y_i = k) = P(d_i = j \mid z_i = K) = q_{jk}$$

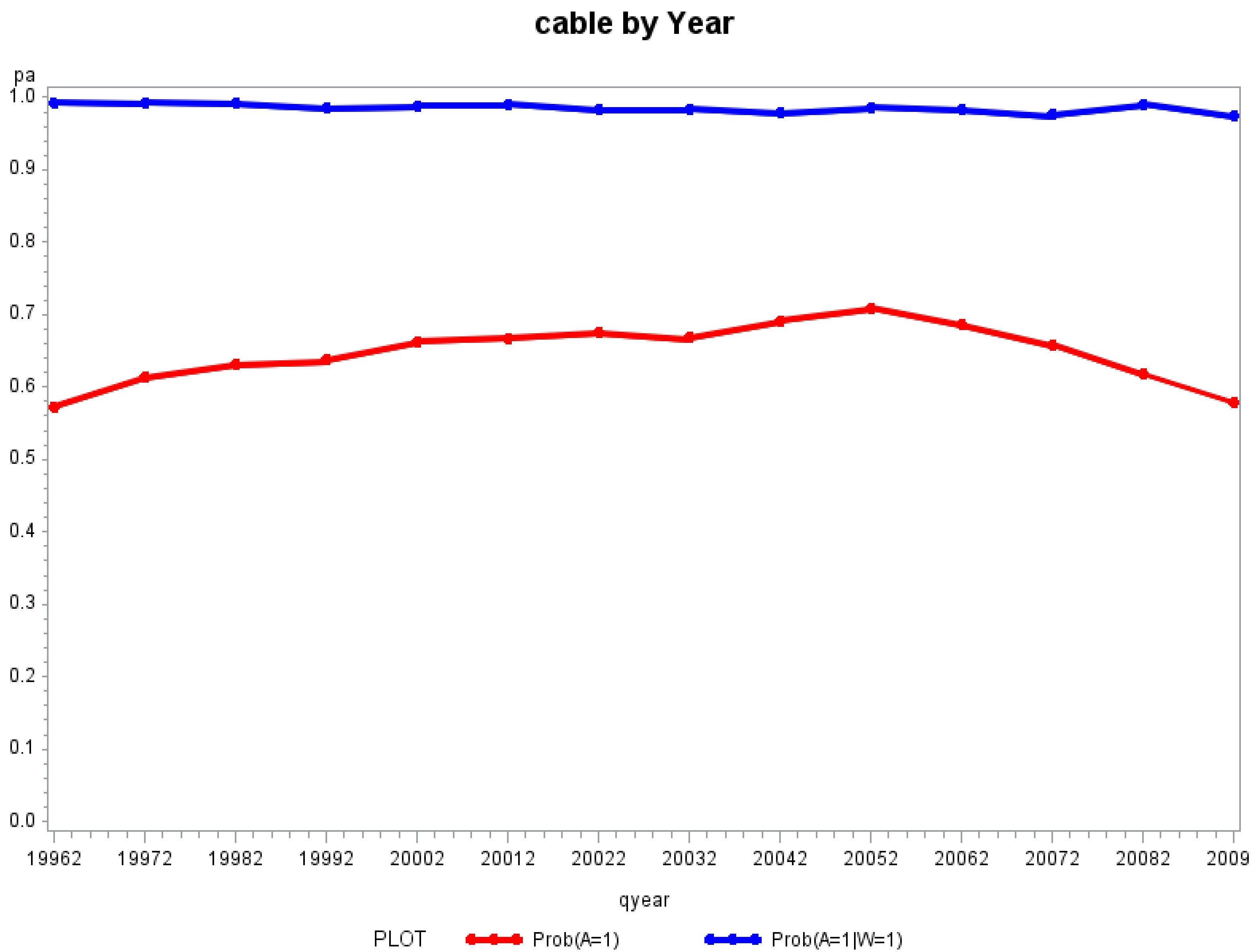
- No False Positives

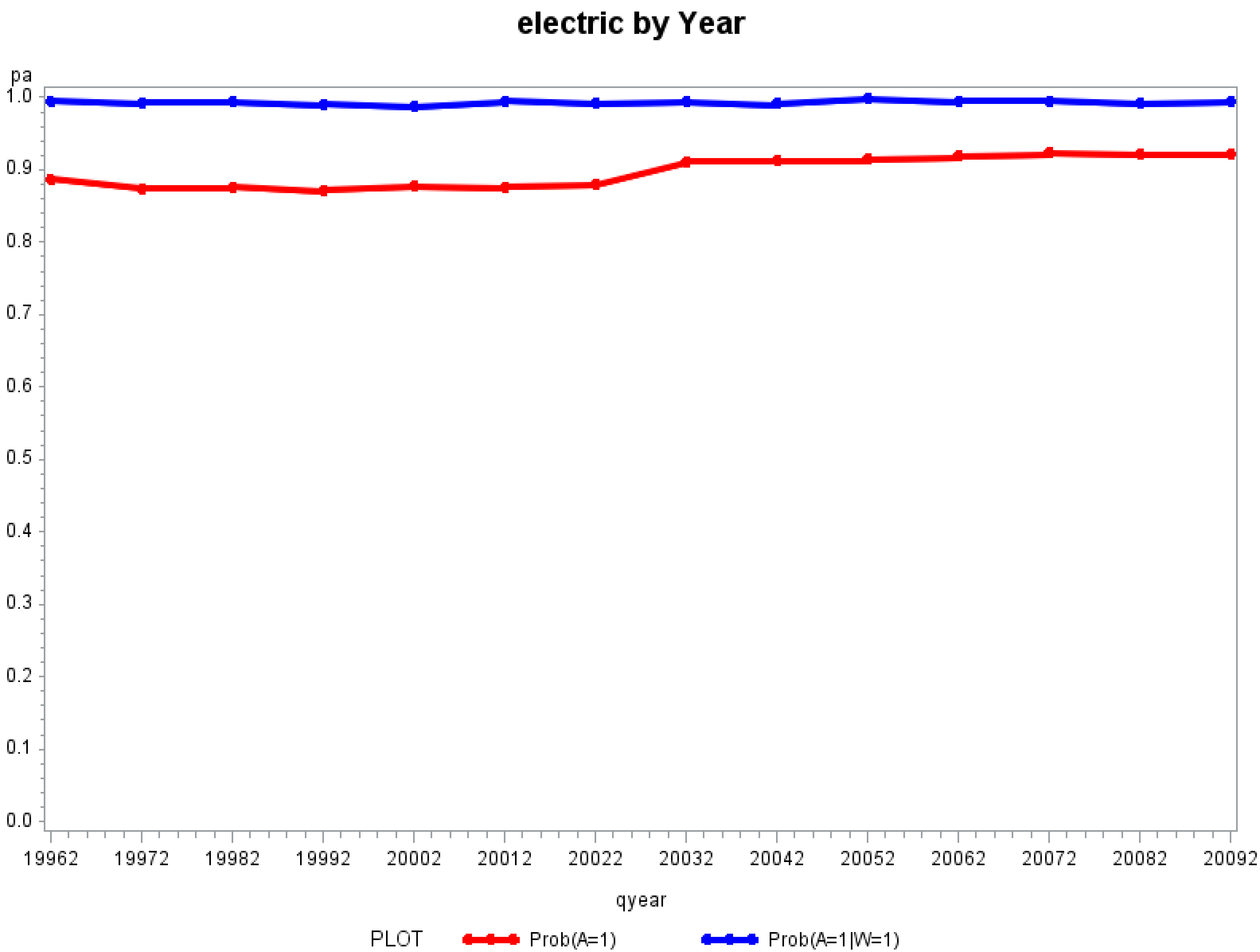
$$P(a_i = 1 \mid w_i = 2) = 0$$
$$P(a_i = 2 \mid w_i = 2) = 1$$

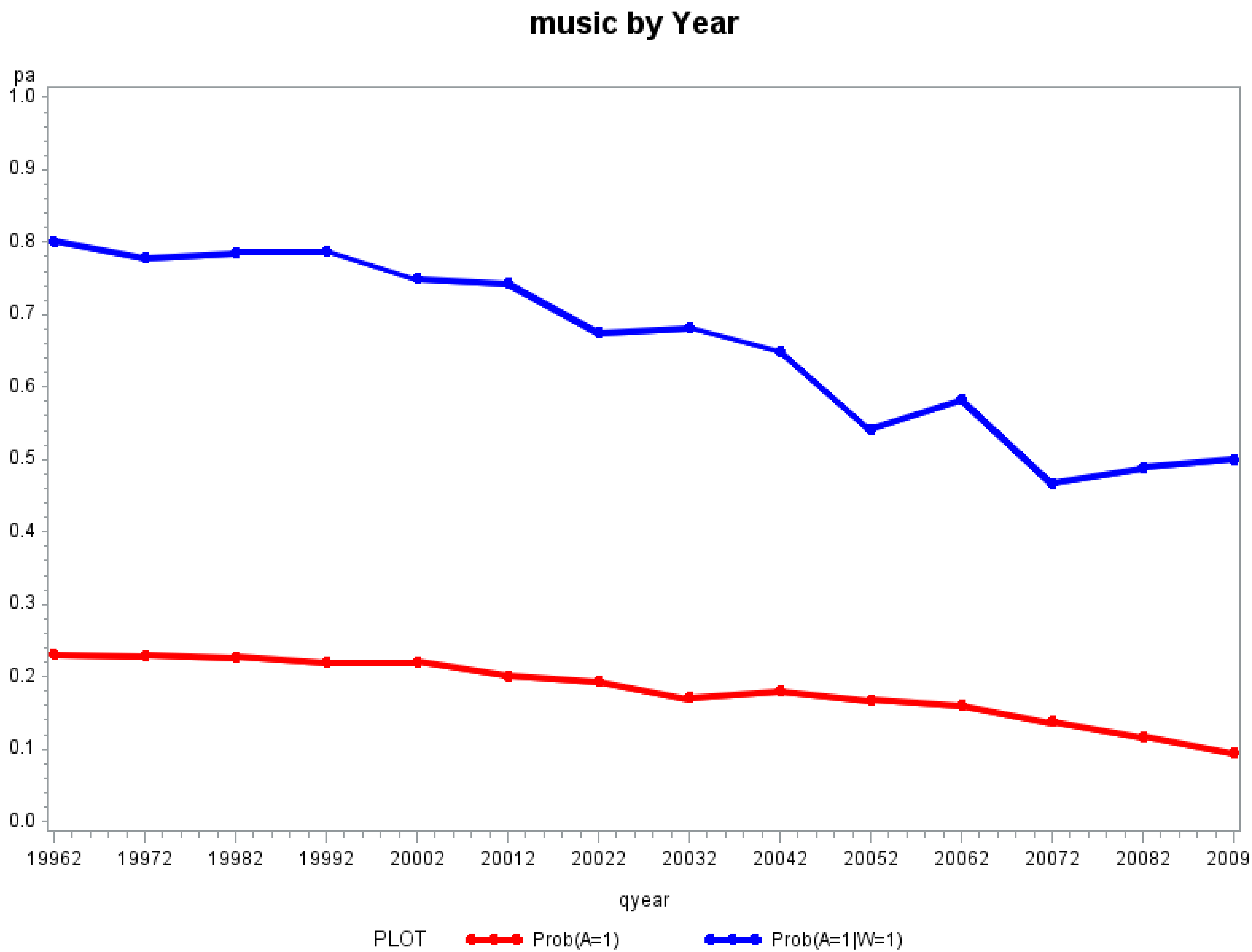
BLS

LCA MOVER-STAYER OVER TIME
(1 YEAR POOLED COHORTS)


books by Year
pa
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
19962
19972
19982
19992
20002
20012
20022
20032
20042
20052
20062
20072
20082
20092
qyear
PLOT
Prob(A=1)
Prob(A=1|W=1)


cable by Year
pa
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
19962
19972
19982
19992
20002
20012
20022
20032
20042
20052
20062
20072
20082
20092
qyear
PLOT
Prob(A=1)
Prob(A=1|W=1)


electric by Year
pa
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
19962
19972
19982
19992
20002
20012
20022
20032
20042
20052
20062
20072
20082
20092
qyear
PLOT
Prob(A=1)
Prob(A=1|W=1)


music by Year
pa
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
19962
19972
19982
19992
20002
20012
20022
20032
20042
20052
20062
20072
20082
20092
qyear
PLOT
Prob(A=1)
Prob(A=1|W=1)

Summary of Previous Findings

- Assessment of MLCA for the detection of *change* in measurement error/time
- Accuracy rates/all estimates noisy
- Estimates are reasonable
- Useful: Sensitive to survey changes

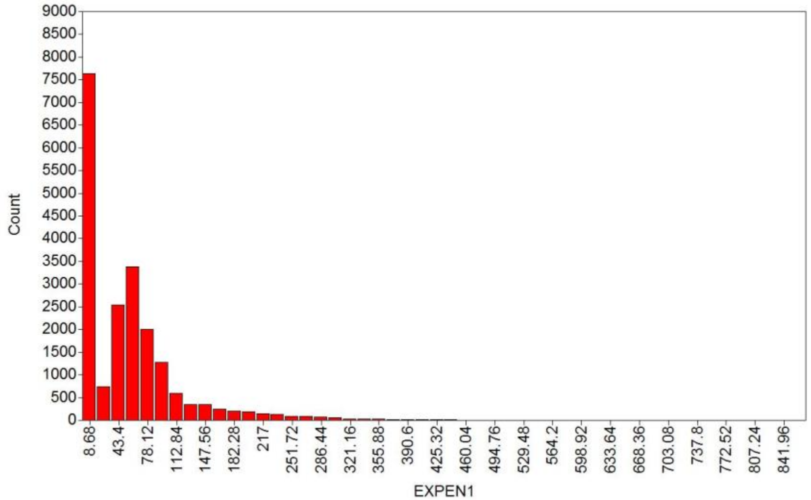
BLS

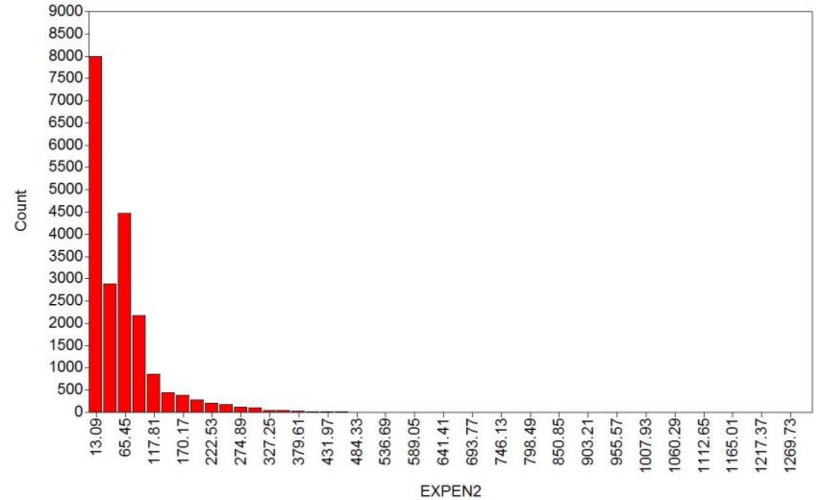
Add Expenditure

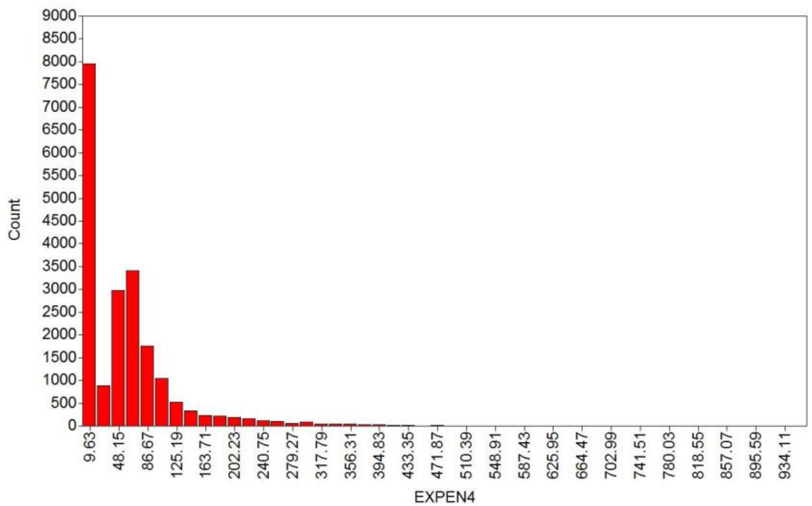
- Adds information to the model
- Allows for simultaneous (as opposed to two-stage estimation of unreported expenditure)
- If auto-correlative effects are large – stabilization of estimates should result

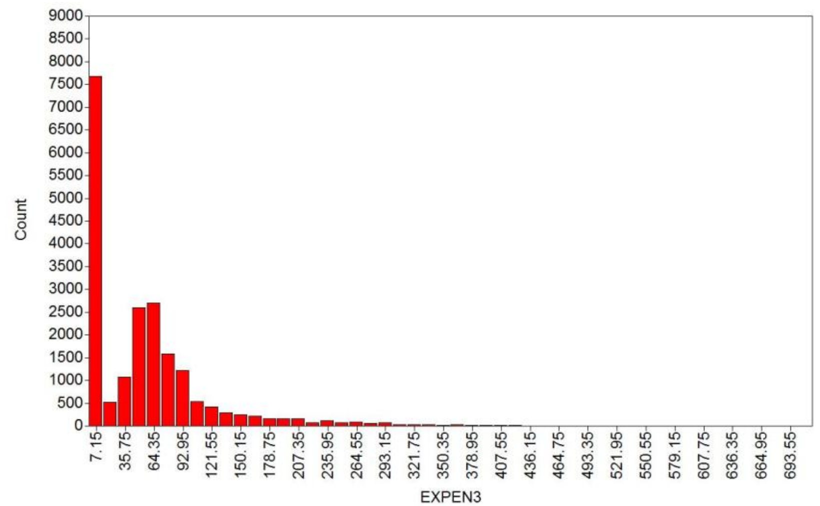
BLS

Expenditure data raw

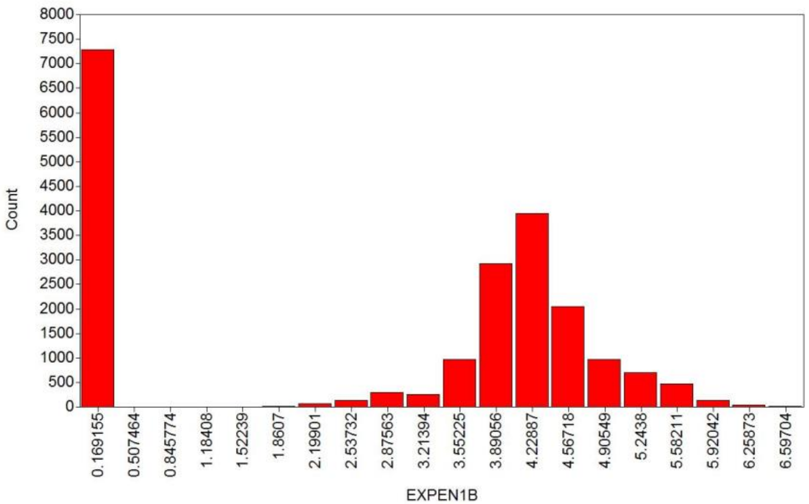


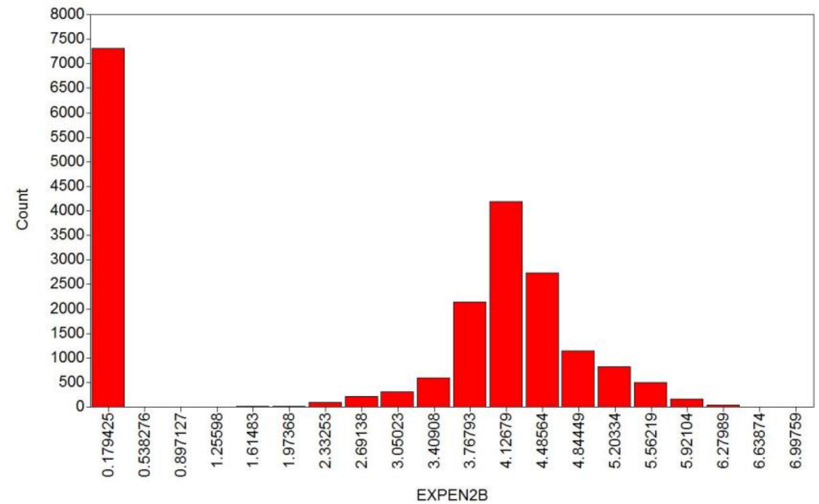


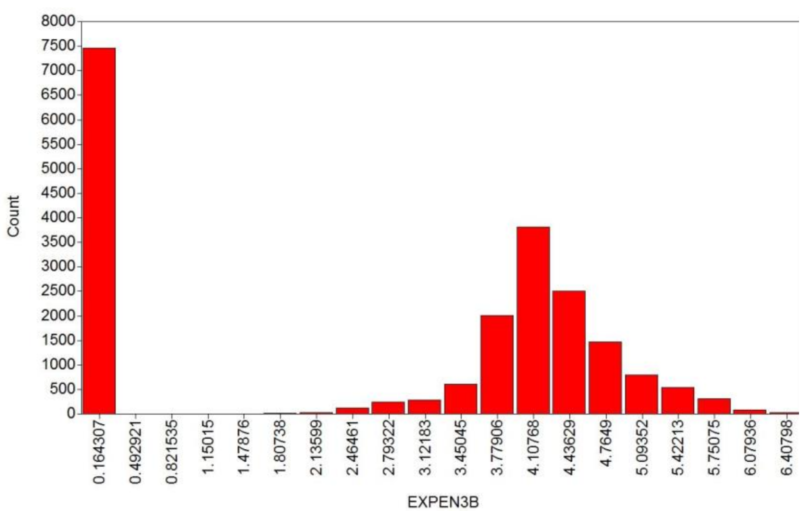


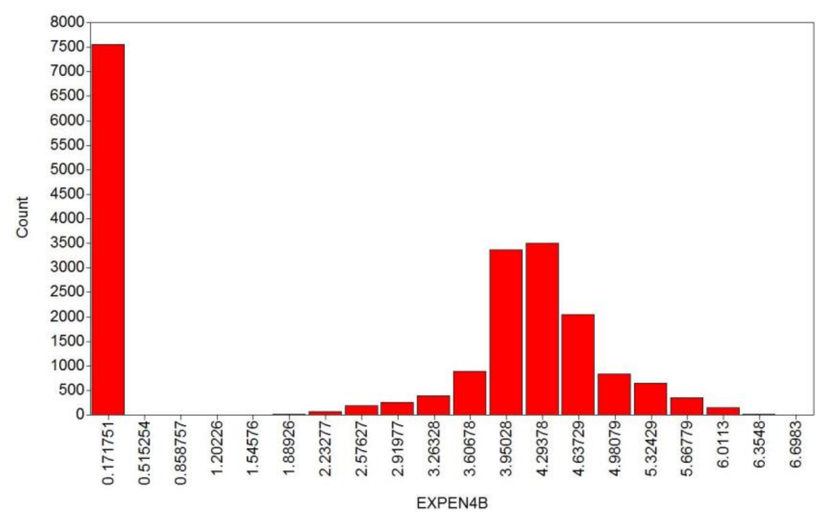


Expenditure Data (ln)

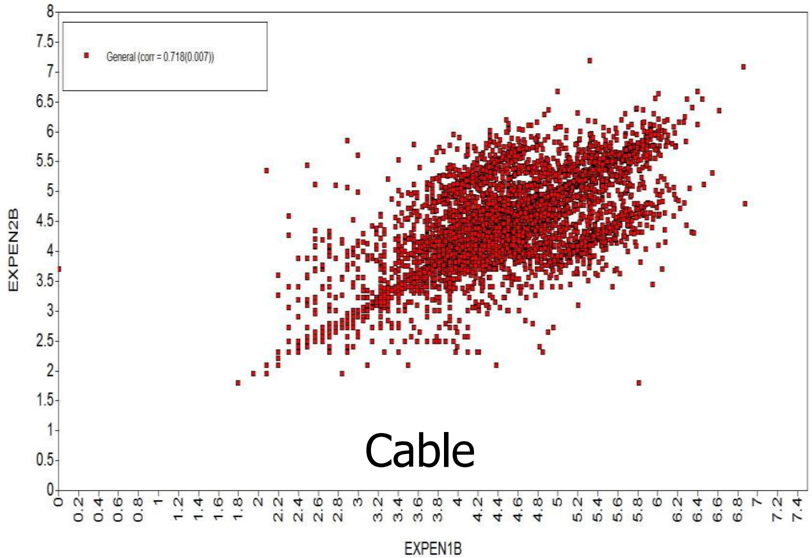


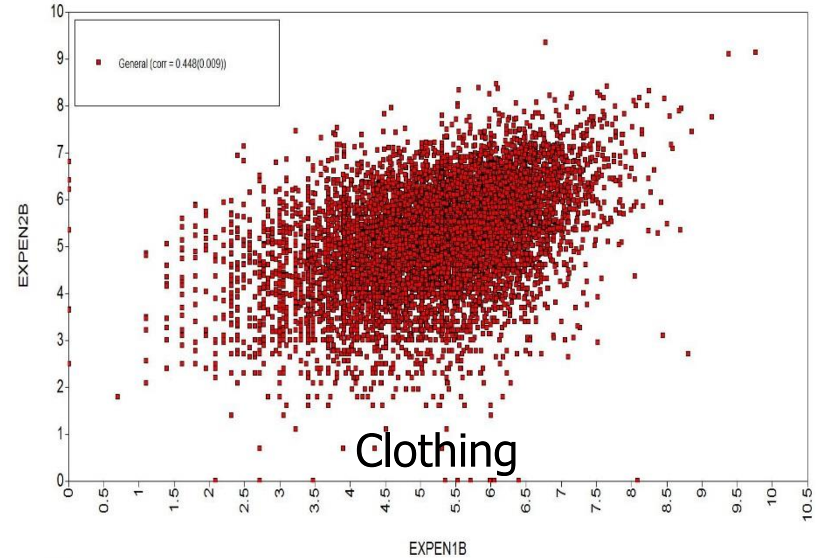


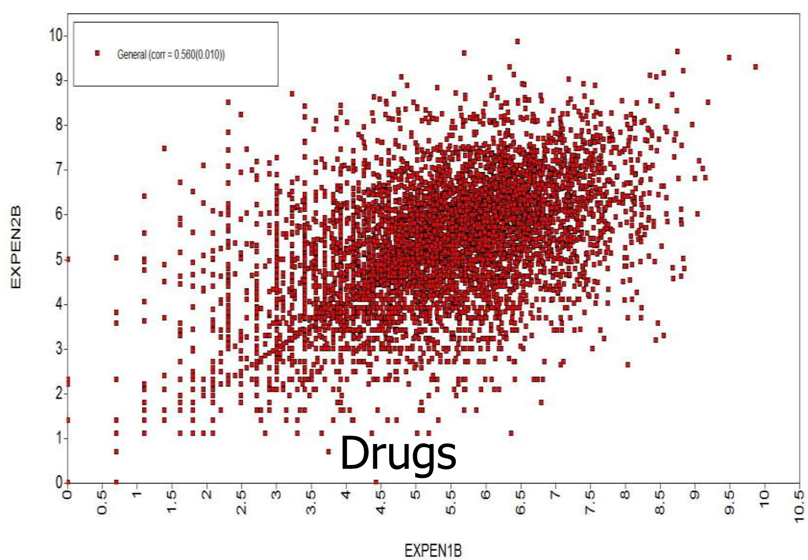


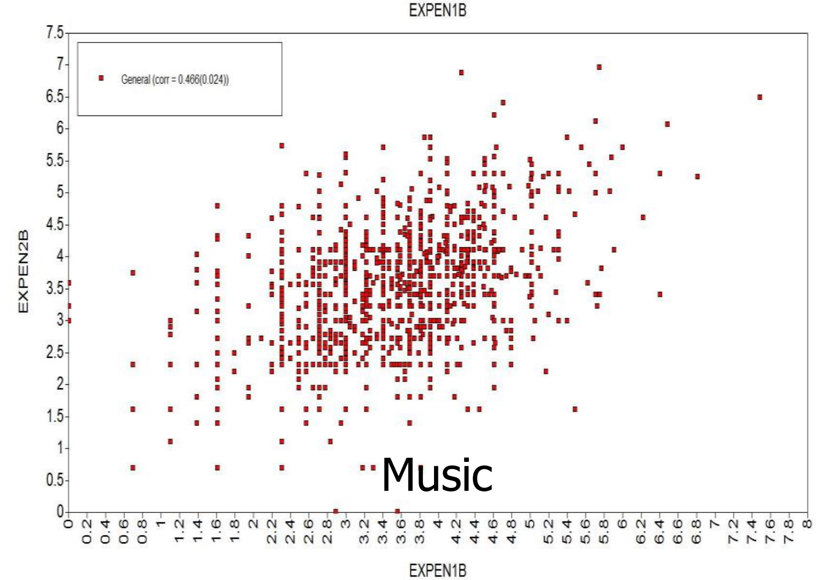


Autoregressive (Observed)

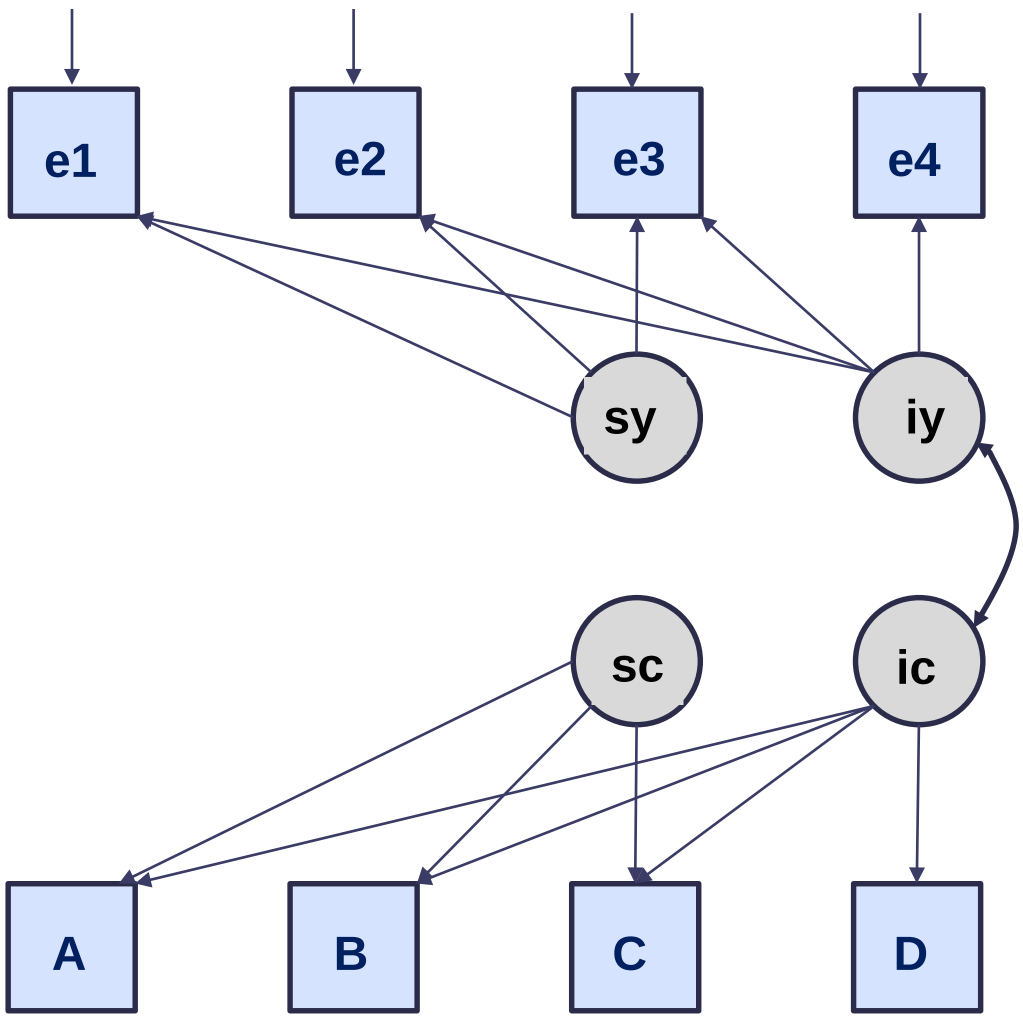






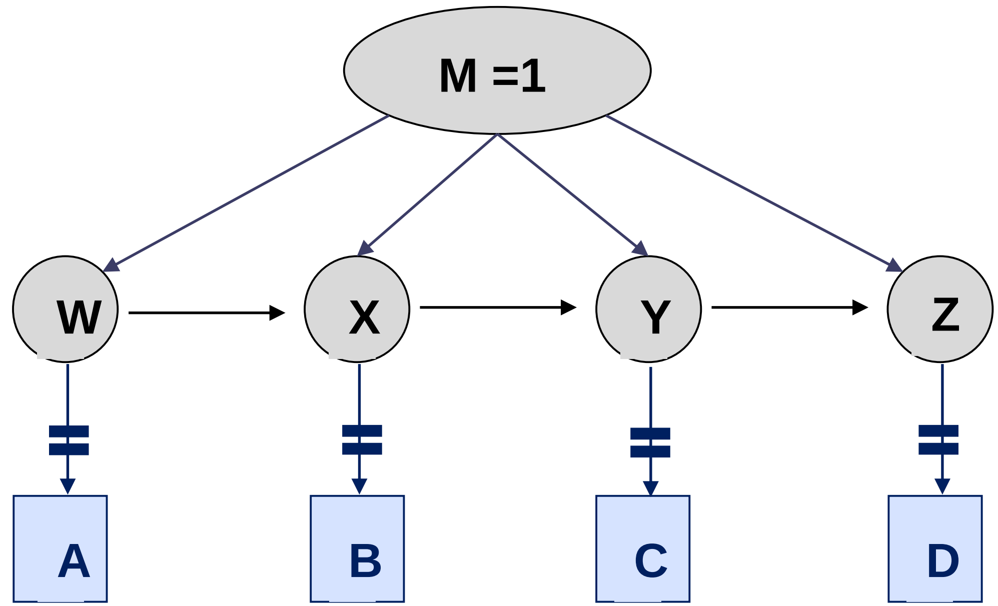


Two-part Model (Olsen and Schaeffer 2001)

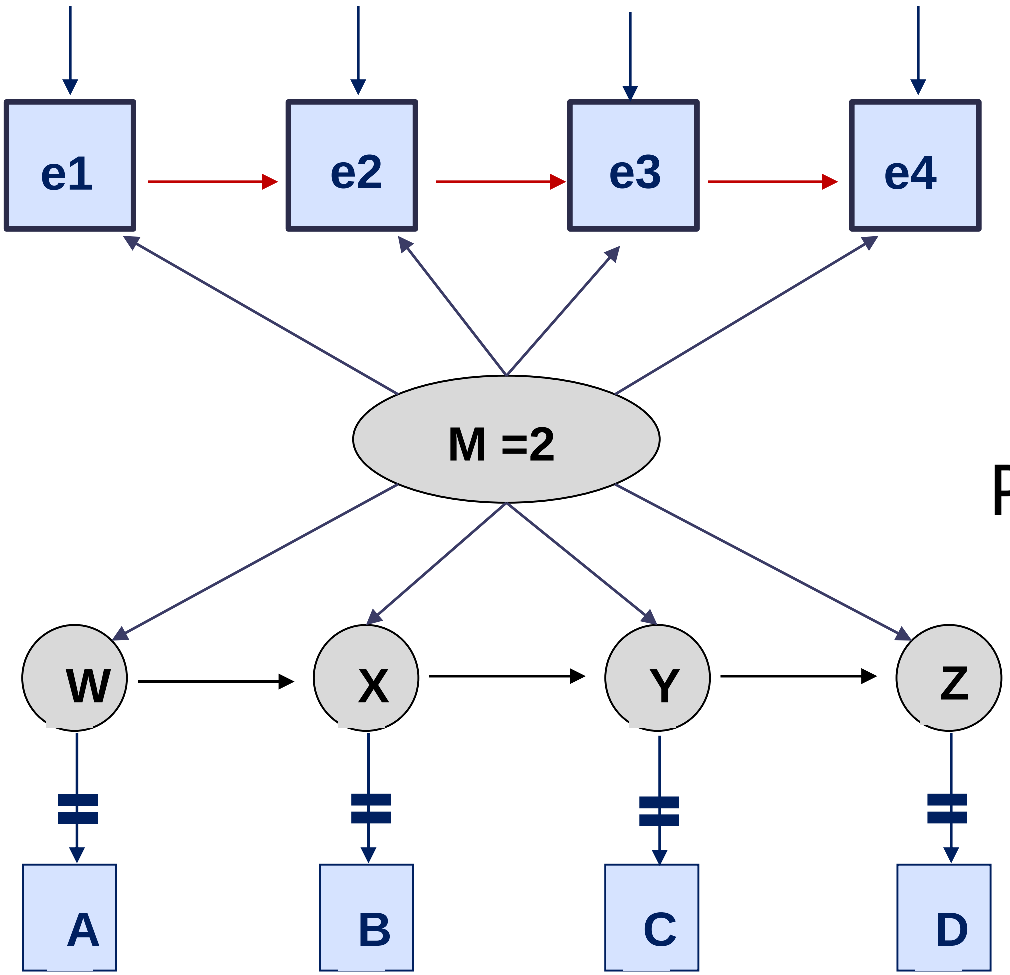


Modified MS

Mover

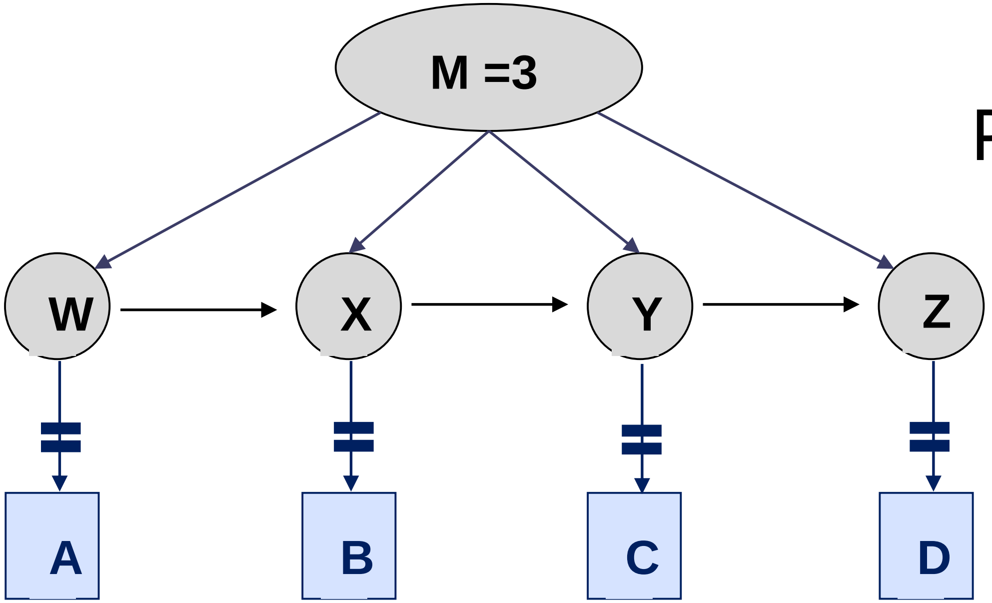


Modified MS



Stayer
Purchaser

Modified MS



Stayer
Non-
Purchaser

Objective Diagnostics

- Fit Statistics
  - L-square

$$L^2 = 2\sum_i n_i \ln\left(\frac{n_i}{\hat{m}_i}\right)$$

  - Entropy

$$en(\alpha) = -\sum_i^N \sum_j^J \alpha_{ij} \log \alpha_{ij}$$

  - BIC

$$BIC = L^2 - df \log N$$

BLS

Results Model Fit

Cable

Model	log L	BIC	Entropy
2nd order	-34769.864	69579.396	.961
Mover-Stayer	-30532.432	61183.868	.845

Two part	-63092.649	126294.386	na
Modified MS	-148951.831	297949.663	.983

BLS

Results Model Fit

Clothing

Model	log L	BIC	Entropy
2^{nd} order	-50629.582	101298.83	.801
Mover-Stayer	-49623.976	99366.956	.793
Two part	-121605.028	243319.143	na
Modified MS	-227619.635	455285.270	.935

BLS

Results Model Fit

Drugs

Model	log L	BIC	Entropy
2nd order	-48549.757	97139.181	.768
Mover-Stayer	-45002.673	90124.351	.766

Two part	-106225.730	212560.548	na
Modified MS	-214846.321	429920.735	.924

BLS

Results Model Fit

Major Appliances

Model	log L	BIC	Entropy
2nd order	-16041.502	32122.673	.200
Mover-Stayer	-15990.318	32099.640	.576
Two part	-22532.573	45174.233	na
Modified MS	-123067.662	246363.416	.759

BLS

Results Model Fit

Music

Model	log L	BIC	Entropy
2nd order	-32777.309	65594.285	.503
Mover-Stayer	-31134.973	62388.951	.769
Two part	-46334.975	92779.037	na
Modified MS	-135818.829	271865.751	.823

BLS

Results, Reports/Expenditure

Cable

Model	P(A=1/W=1)	Missing Expenditure /QTR	Missing Expenditure /CU QTR
2nd order	.984	$1,751.96*	$1.17*
Mover-Stayer	.984	$1,751.96*	$1.17*
Two part	na	$18,358.56	$12.24
Mod MS	.984	$996.16	$.66
**Estimated from P(A/W)*			

Results, Reports/Expenditure

Clothing

Model	P(A=1/W=1)	Missing Expenditure /QTR	Missing Expenditure /CU QTR
2nd order	.910	$25,059.97*	$16.71*
Mover-Stayer	.846	$46,124.29*	$30.75*
Two part	na	-$26,799.50	-$17.87
Mod MS	.742	$33,992.79	$22.66
**Estimated from P(A/W)*			

BLS

Results, Reports/Expenditure

Drugs

Model	P(A=1/W=1)	Missing Expenditure /QTR	Missing Expenditure /CU QTR
2nd order	.817	$60,399.74*	$40.27*
Mover-Stayer	.903	$28,966.10*	$19.31*
Two part	na	-$33,316.50	-$22.21
Mod MS	.877	$17,222.49	$11.48
**Estimated from P(A/W)*			

BLS

Results, Reports/Expenditure

Major Appliances

Model	P(*A=1/W=1)*	Missing Expenditure /QTR	Missing Expenditure /CU QTR
2nd order	?	?	?
Mover-Stayer	?	?	?
Two part	na	$5,182.59	$3.46
Mod MS	< 0	-$27,784.80	$-18.52
**Estimated from P(A/W)*			

BLS

Results, Reports/Expenditure

Music

Model	P(*A=1/W=1)*	Missing Expenditure /QTR	Missing Expenditure /CU QTR
2nd order	.445	$58,551.30*	$39.34*
Mover-Stayer	.741	$16,409.11*	$10.94*
Two part	na	-$1,840.63	-$1.23
Mod MS	.187	$31,017.09	$20.68
**Estimated from P(A/W)*			

BLS

Are Models Worth It?$_1$

- Generally more information is better
- Time consuming – estimation is slow
- Model fit does suffer more than expected
- Are estimates of missing expenditure superior?
  - LCA Mover-Stayer, 2nd order, are vetted, stable over time, estimates make sense, internal validity, validation with external sources

BLS

Are Models Worth It?$_2$

- Estimates are no more believable for difficult expenditure categories (e.g. major appliances)
- Two part latent growth produces very different estimates
- Some support for modified mover-stayer
- Much more testing is needed
  - Grouping variables
  - Examine estimates over time
  - Validation with external sources

BLS

Contact Information

Brian Meekins
Office of Survey Methods Research

202-691-7594
meekins.brian@bls.gov

BLS

Estimating Magnitude of Underreported Expenditures for False Negative: Notation

$\hat{R}_c$ Total *reported* expenditures for persons with characteristics, *c*

$\hat{\pi}_{1|1,c}$ Accuracy rate for persons with characteristics, *c*, estimated from M-S model; i.e. P(A=1|W=1)

T_c True total expenditures persons with characteristics, *c*

$T_{c,+}$ True total expenditures persons with characteristics, *c* for true positives

$T_{c,-}$ True total expenditures persons with characteristics, *c* for false negatives

BLS

Assumptions

- No false positive reports of expenditures
- Reported expenditures are accurate; i.e.,

$$E(\hat{R}_c) = T_{c,+}$$

- Mean expenditures for reporters and mean expenditures for nonreporters are equal

Estimate of Underreports Due to False Negatives

Under these assumption, an estimate of T_c is

$$\hat{T}_c = \frac{\hat{R}_c}{\hat{\pi}_{1|1,c}}$$

Thus, an estimate of $T_{c,-}$ is

$$\hat{T}_{c,-} = \hat{T}_c - \hat{R}_c$$

BLS

Mover-Stayer Model

Assumptions

Population can be divided into:

- Persons who purchase the item in each quarter ("purchase-stayers")
- Persons who do not purchase the item in any quarter ("nonpurchase-stayers")
- Persons whose purchase behavior is not consistent across the quarters ("movers")

Additional Assumption

- No false positive reports. Persons who report a purchase are assumed to have actually made that purchase.

Definition of Latent Variables

Where,

$$W = \begin{cases} 1, & \text{if one or more purchases of an item during the quarter ("purchaser")} \\ 2, & \text{if no purchase ("non-purchaser")} \end{cases}$$

with similar definition for *X*, *Y*, *Z* for 2nd, 3rd, and 4th interview

Definition of Indicator Variables

Define for Interview 1,

A = 1 if reported as a purchaser for the quarter
2 if reported as non-purchaser

with similar definition for *B*, *C*, *D*
for 2nd, 3rd, and 4th interviews

BLS

Grouping Variables

1. Family size
2. Refusal to answer income question
3. Derived variable combining records use and interview length
4. Income class

BLS