Evaluating Hot Deck with Propensity Score Matching For the Advance Monthly Retail Trade Survey (MARTS)

Katherine Jenny Thompson (U.S. Census Bureau)
Laura Bechtel (U.S. Census Bureau)*
Nicole Czaplicki (U.S. Census Bureau)

2018 Federal Committee on Statistical Methodology
March 7, 2018

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

** Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.*

Outline

- Hot Deck
- Propensity Matching
- Relationship between Hot Deck Imputation and Propensity Matching
- Our Application
- Evaluation
- Concluding Remarks
- Related Research

Hot Deck Imputation

- Often described as “model free”
- Donors – reported values
- Recipients – missing values
- Recipient and donor are matched
    - Direct substitution from donor

$$Current\ Month\ Sales_{\ Recipient} = Current\ Month\ Sales_{\ Donor}$$

    - Derived from donor

$$Current\ Month\ Sales_{\ Recipient} =$$

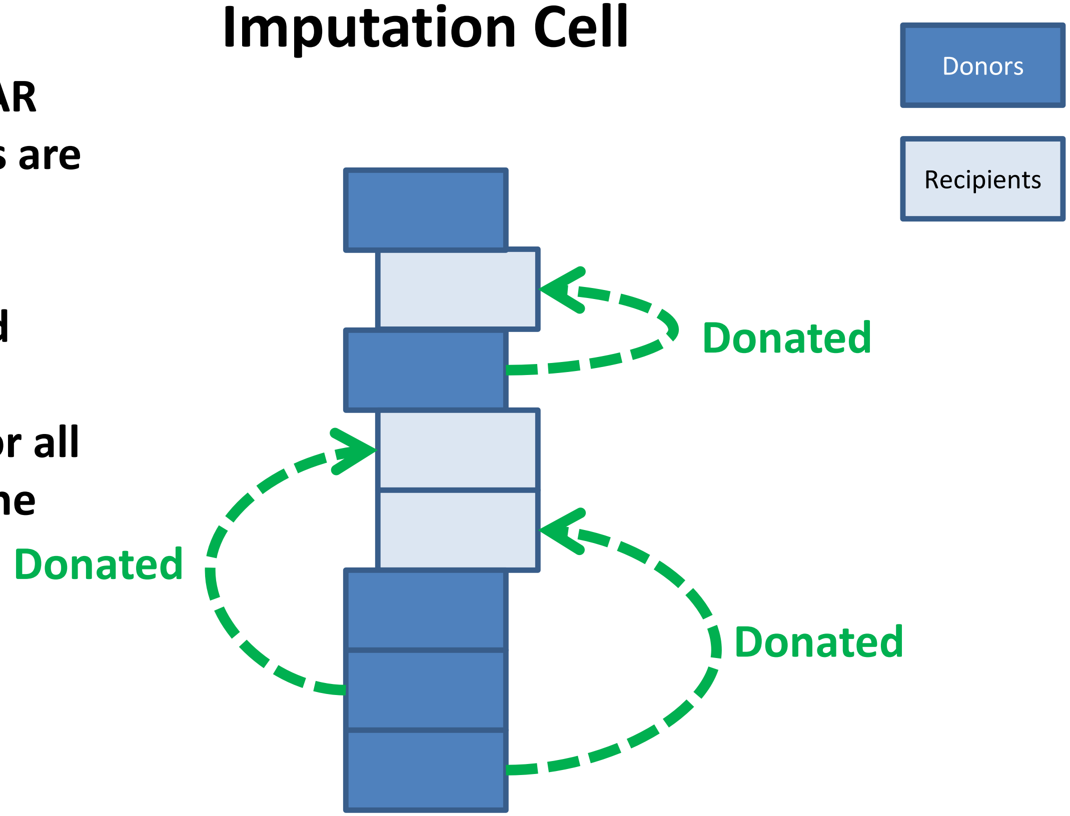
$$\frac{Current\ Month\ Sales_{\ Donor}}{Previous\ Month\ Sales_{\ Donor}} Previous\ Month\ Sales_{\ Recipient}$$

Random Hot Deck

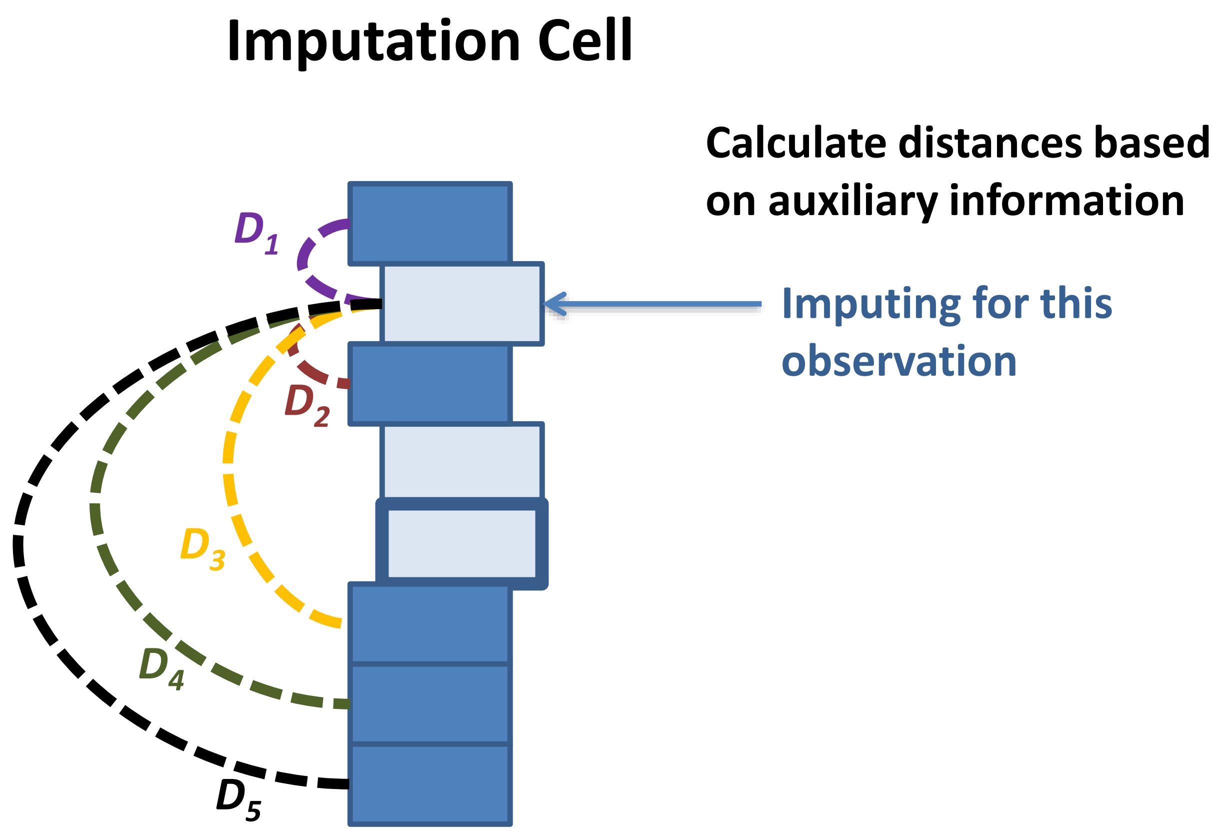
Imputation Cell

Assumes MCAR or MAR when imputation cells are used.

Assumes the expected value of outcome of interest is the same for all observations within the imputation cell.

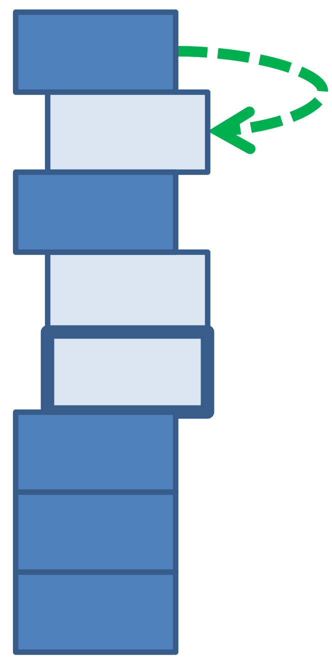


Nearest Neighbor Hot Deck



Nearest Neighbor Hot Deck

Imputation Cell



Min Distance Donated

Assumes the outcome variable can be predicted by the auxiliary variables

Hot Deck With Business Surveys

- Skewed population
  - Direct donation not a good idea for quantitative variables
  - Nearest Neighbor often used (size predictive of response/outcome)

- Derived value – donor ratio

- More recipients than donors

- Seasonal effects/trading day effects

Propensity Score Matching

- Background
  - Causal inference/causal assumptions
  - Predicting outcome variable (response to treatment due to factors that are common to both treatment and control)
- Propensity Score
  - One single score or combinations of variables

What About Propensity Scoring?

How do you develop one appropriate score function?	Compromise between the two methods	What about important continuous variables?
Everything into the score (all variables)	Develop a score within a block	No score (block)

How propensity scoring works

- Matching
  - Need to specify a distance function.
  - Cannot re-use donors (one to one or many to one).
- Greedy matching[1]
  - Pairs donors to recipients sequentially.
  - Sort matters (confounding with distance).
  - Need to have more donors than recipients to use.
- Optimal matching[1]
  - Pairs donors to recipients based on closest distance subject to minimizing total aggregated distance over all recipients.
  - Distance function matters.

[1]Used publicly available SAS code developed by Bergstralh and Kosanke at the Mayo Clinic (http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros)

Greedy matching

Recipients sorted in ascending sequence

Distance between donor W and Recipient A is 7

Donor **Y** is matched with Recipient **A**

	Donors			
Recipients	W	X	Y	Z
A	7	8	**5**	13
B	10	9	4	**6**
C	**11**	17	8	10
D	25	**14**	7	8

Greedy matching

	Donors			
Recipients	W	X	Y	Z
A	7	8	**5**	13
B	10	9	4	**6**
C	**11**	17	8	10
D	25	**14**	7	8

Greedy matching

	Donors			
Recipients	W	X	Y	Z
A	7	8	**5**	13
B	10	9	4	**6**
C	**11**	17	8	10
D	25	**14**	7	8

Greedy matching

	Donors			
Recipients	W	X	Y	Z
A	7	8	**5**	13
B	10	9	4	**6**
C	**11**	17	8	10
D	25	**14**	7	8

Greedy matching –sort matters

Recipients sorted in descending sequence

Donor **W** is matched with Recipient **A**

	Donors			
Recipients	W	X	Y	Z
D	25	14	**7**	8
C	11	17	8	**10**
B	10	**9**	4	6
A	**7**	8	5	13

Optimal matching

Donor **X** is matched with Recipient **A**

	Donors			
Recipients	W	X	Y	Z
A	7	**8**	5	13
B	10	9	**4**	6
C	**11**	17	8	10
D	25	14	7	**8**

Relationship between hot deck and propensity matching

Donors (Control) / Recipients (Treatment)	→ Propensity Match →	Find donors such that $E\left[\frac{x_{t,d}}{x_{t-1,d}}\right] \approx E\left[\frac{x_{t,r}}{x_{t-1,r}}\right]$	→ Hot Deck Imputation →	Donors (Control) / Recipients (Treatment)

Causal inference framework:

- Treatment = donor selection procedure
- Block = imputation cell
- Outcome = M-T-M change

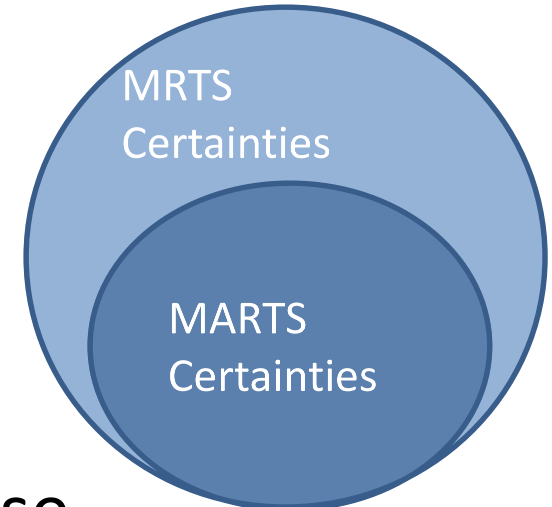
Our application - Advance Monthly Retail Trade Survey (MARTS)

- Monthly Economic Indicator
  - Sales and month-to-month percent change
  - Inputs into the quarterly Gross Domestic Product (GDP) produced by the Bureau of Economic Analysis
- MARTS is a subsample of Monthly Retail Trade Survey (MRTS)
  - Certainties – selected with probability = 1

	MARTS	MRTS
Sample size	5,000 companies	12,000 companies
Sample frame	MRTS sample	Annual Retail Trade Survey sample
Sample design	Stratified PPS -WOR (subsample of MRTS)	Stratified SRS-WOR
Sample redesign cycle	Approximately every 2.5 years	Approximately every 5 years
Time to respond	Approximately 7 business days	Approximately 5 weeks
Imputation	Analyst impute for selected companies	Analyst imputes retained, ratio impute for remaining nonrespondents and edit-failing items
Estimation	Link relative estimator	Horvitz-Thompson estimator
Tabulation industries	30	83

Our application - Advance Monthly Retail Trade Survey (MARTS)

- The largest MRTS Certainties are selected with certainty for MARTS



- Low Response Rates &
  Size is Predictive of Response
- Data are seasonally adjusted
  - Seasonal effects
  - Trading day effects – many series

Simulation Study Design

MRTS Certainty Units
(Not In MARTS)

MARTS Certainty Units

MARTS Noncertainty Units

Source Data:

- In Statistical Period
  - March 2016 – Feb. 2017
- MRTS Certainty Units ONLY
- Responded to MRTS
  - Current Period and Prior Period
  - Both values of sales > 0

Simulation Study Design

MRTS Certainty Units (Not In MARTS) - Donors

MARTS Certainty Units – Donors

MARTS Certainty Units – Recipients

Randomly split within Statistical Period

MARTS Noncertainty Units - Recipients

Relationship between hot deck and propensity matching

Donors (Control) | Recipients (Treatment) → Propensity Match → Find donors such that $\mathrm{E}\left[\frac{x_{t,d}}{x_{t-1,d}}\right] \approx \mathrm{E}\left[\frac{x_{t,r}}{x_{t-1,r}}\right]$ → Hot Deck Imputation → Donors (Control) | Recipients (Treatment)

Causal inference framework:

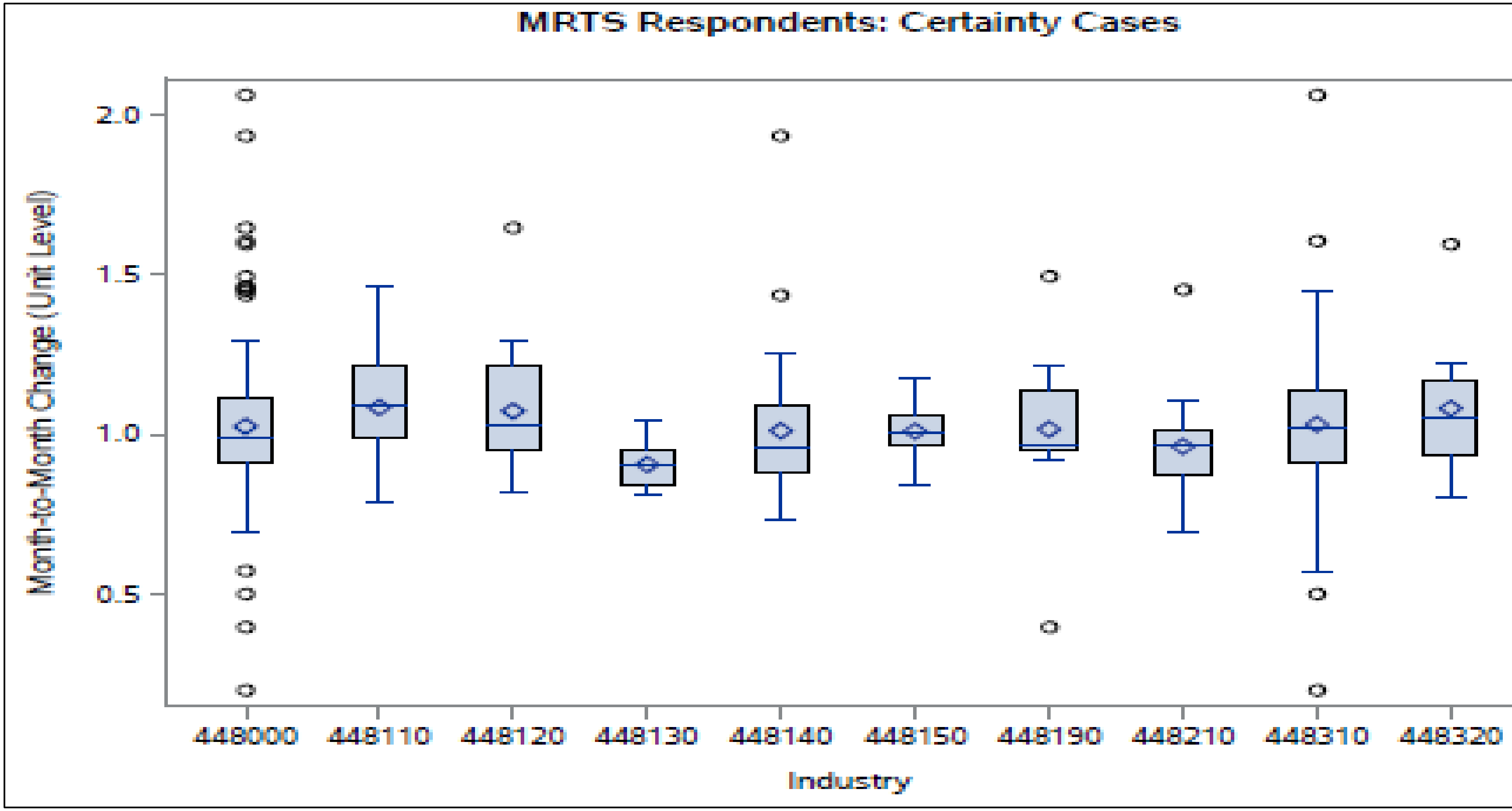
- Treatment = donor selection procedure
- Block = imputation cell
- Outcome = M-T-M change

What should our match variables be?

Finding Matching Variables

- What variables are predictive of month-to-month change?
  - Industry – 6-digit NAICS (North American Industry Classification System) vs 3-digit NAICS

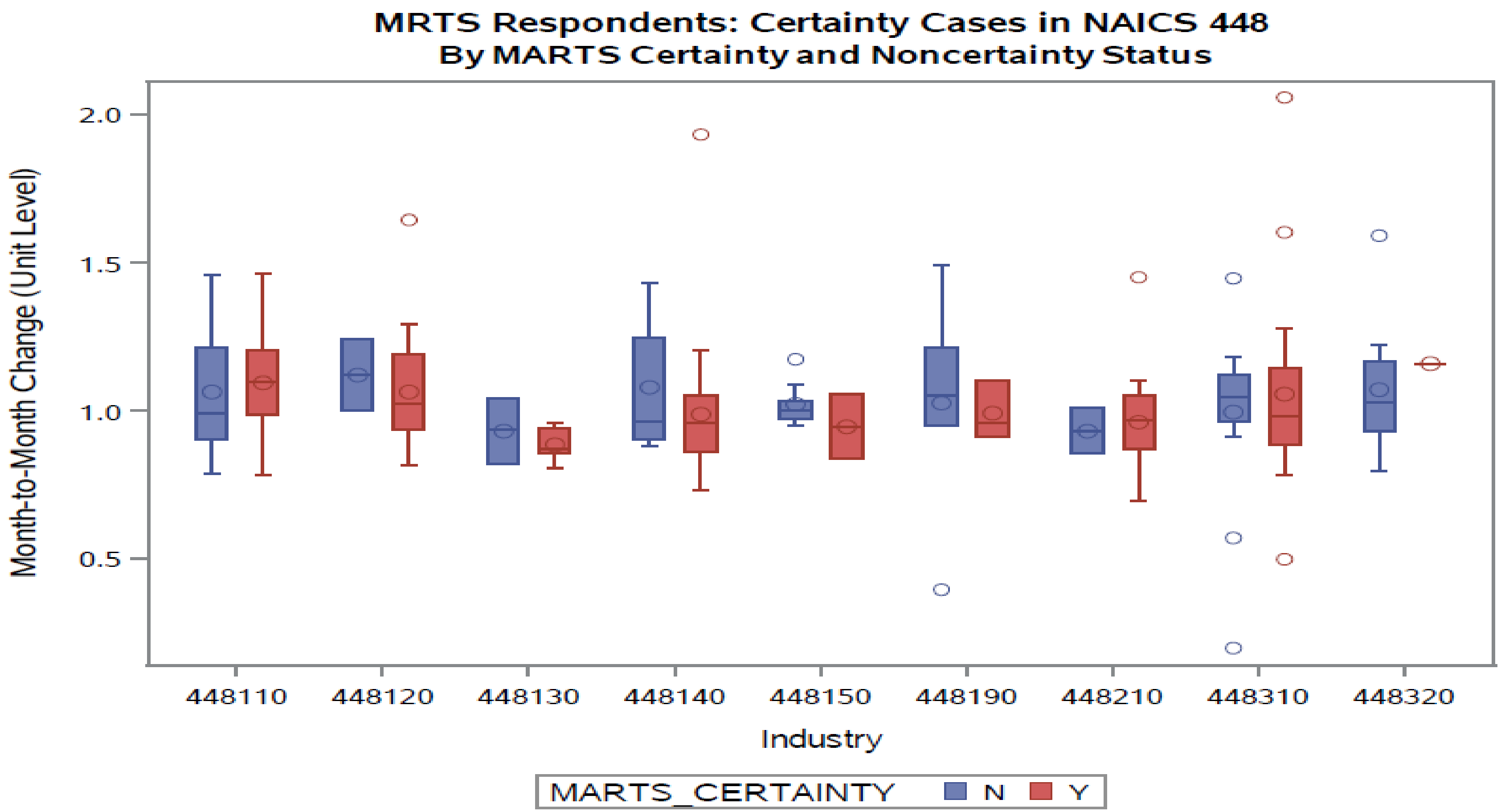
Distributions of Month-to-Month Change in NAICS 448



Finding Matching Variables

- What variables are predictive of month-to-month change?
  - A lot is built into the imputation cells
    - Industry – 6-digit NAICS (North American Industry Classification System) vs 3-digit NAICS
  - Unit size

Distributions of Month-to-Month Change in NAICS 448



Finding Matching Variables

- Predictive of m-t-m change
  - A lot is built into the imputation cells
    - Industry – 6-digit NAICS (North American Industry Classification System) vs 3-digit NAICS
  - Unit size is important – but we are restricted to MRTS certainty only (historic data limitations)
- Other factors investigated
  - Prior month sales (size)
  - Sampling weight (size)
  - Variables predictive of response

Actual Matches

- Blocks/Imputation Cells – 6-digit industry

- Matching variables
  - Prior month sales
  - Number of industries that the company operates in (proxy for complexity of the company)

Evaluation

		Hot deck method	Match Variables	Sort Variables
Greedy	1	Random hot deck	Random number	Random number
	2	Nearest neighbor	Prior Month Sales	Random number
	3	Propensity	Prior Month Sales	Prior months sales (descending)
	4	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	Random number
	5	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	Prior month sales (descending)
Optimal	1	Propensity	Prior Month Sales	N/A
	2	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	N/A

Evaluation Statistics: Mean Absolute Error

- ***Mean Absolute Error (MAE)***

$$MAE_t^{am} = \sum_{i=1}^{n'_{rt}} \left| \underbrace{x_{t,r(i)}}_{\text{Truth}} - \underbrace{\left(\frac{X_{t,d(i)}^{am}}{X_{t-1,d(i)}^{am}} \right) x_{t-1,r(i)}}_{\text{Imputed value}} \right| / n'_{rt}$$

measures the average magnitude of the error per imputed unit.

Evaluation Statistics: Relative Bias

- ***Unconditional Relative Bias(URB)*** – measures the overall effect of the imputation error on the tabulated estimates.

$$RB_t^{am} = \frac{\hat{X}_t^{am}}{X_t} - 1$$

- ***Conditional Relative Bias (CRB)*** – provides the direction of the imputation bias for the imputed units and gives some indication of the magnitude. Extremely sensitive to size.

$$CRB_t^{am} = \frac{\hat{X}_t^{am(R)}}{X_t^R} - 1$$

Two Phases to our Research

Donors (Control) / Recipients (Treatment)	Phase 1	Find donors such that $E\left[\frac{X_{t,d}}{X_{t-1,d}}\right] \approx E\left[\frac{X_{t,r}}{X_{t-1,r}}\right]$	Phase 2	Donors (Control) / Recipients (Treatment)

- Find which matching applications are most effective in selecting donors (imputation constant)
  - Donated ratio - current month/prior month
- Compare statistical performance of the recommended matching algorithm from Phase1 (imputation varied, matching constant)
  - Donated ratios from 1 year ago (seasonality)
  - Donated ratios from most recent calendar with the same working day composition (seasonality & trading day)

One Match Variable Versus Two

- Chi-square tests for independence
- Treatment = two match variables
- Control = one match variable
- Optimal and Greedy match – no improvement with two

Phase 1 Summary

Looking at MAE and CRB

- Random Hot Deck worst performance
- Nearest Neighbor slight underperformance compared to Optimal and Greedy
- Greedy and Optimal similar performance
  - Greedy - needed to “trick” the code
- Phase 2 will focus on Optimal Matching

Phase 2: Selection of Hot Deck Donor Pool

Ratio	Min.	Q1	Med.	Q3	Max
Donors (1 Year Ago) to Recipients	0.89	1.69	2.14	3.19	5.58
Donors (5 Years Ago) to Recipients	0.55	0.97	1.38	1.69	2.70

Phase 2: Chi-Square Test for Independence to Assess Treatment Effect (Donor Choice)

	1 year ago outperformed	5 years ago outperformed	Tie between the 2 treatments
MAE	17	11	2
URB	16	9	5

- Example where p-value is misleading
  - There is a an effect overall…but it ignores differences within industries

Concluding Remarks

- Optimal matching effective
  - Parsimonious model works
  - No need for a single score in our application

- Challenge in determining how to use donors
  - No one-size-fits-all model with for choosing ratios
  - Considering alternative calendar adjustments

Related Research

- Comparison to other missing data treatments as part of a larger study
  - 10:30 tomorrow morning in 145AB Nikki Czaplicki is presenting “Finding an Estimator that Minimizes Revisions in a Monthly Indicator Survey”

Thank you

Laura Bechtel

Laura.Bechtel@census.gov