IEPEC International Energy Program Evaluation Conference

CLEANING UP THE MESS OF ENERGY BILLING DATA

An investigation of Differences in Billing Analysis Results Caused by Data Cleaning Methodologies

Eric Ziemba, Stefanie Wayland, Olivia Patterson
Opinion Dynamics

2017 IEPEC Conference – Baltimore, Maryland
August 9, 2017

Opinion **Dynamics**

"For every complex problem there is an answer that is clear, simple, and wrong."

~H. L. Mencken

Core elements to billing analysis

- Data cleaning and preparation
- Comparison group selection
- Model specification and validation

- Always strive to provide the right answer

- Can we rely on “Quick & Dirty” cleaning methods to yield reliable savings estimates?

- **Hypothesis:** *Data cleaning has a significant effect on regression results … but there is a point at which time spent on data cleaning has diminishing returns.*

- Data cleaning is shown to have a significant effect on regression results
- Very simple data cleaning improves the accuracy of savings estimates

Experiment

- We examine accuracy of point estimates from models using:
  - Data from actual program evaluations performed
  - **Data representing two difference stages in cleaning process**
  - Two period-assignment methodologies
  - Two models to improve the validity of our results and allow us to see if results from each set of data are model dependent
  - Future participant comparison group in each model

Accuracy of Results

- Two measures of data cleaning effects:
  - Model results fall within the error-bounds of the same model that uses the final data.
  - Scaled value that indicates proximity of test data results to final data.

Data Versions

Raw Data

- Data are exactly as delivered from the utility

Deduped Data

- All duplicated records removed from Raw Data

Final Data

- Compilation of the final datasets for each program included
- These data went through detailed and extensive cleaning, with an effort to maintain to most data possible
- We use these data as a benchmark for accurate results

Methods Used

- Period Assignment Methodologies
  - Mid-Point of Bill
  - Calendarized Bills
- Models
  - Linear Fixed Effects Regression (LFER) Model
  - Two-Way LFER Model

Accuracy vs. Time

Accuracy	Raw Data	Deduped Data	Additional Steps …	Final Data
Within Error	44%	63%	47%	100%
Scaled Value	1.81	0.99		0.0
% Full Cleaning Time	1%	5%	95%	100%

- Diminishing returns on time spent?

$$\text{Scaled Value} = \frac{|Estimate - Final|}{Error}$$

Implications

- Data cleaning is shown to have a significant effect on regression results
- Very simple data cleaning improves the accuracy of savings estimates
- Opens the door for additional research on determining with data cleaning steps/methods provide the most “bang for the buck”

Thank you

Eric Ziemba
eziemba@opiniondynamics.com
617-492-1400 ext. 4644

Opinion **Dynamics**