

Data Analysis for the EIA-826: Test results

Nancy Kirkendall and Joe Sedransk, SMG, EIA

EIA Form-826 collects information, monthly, from regulated and unregulated companies that sell or deliver electric power to end users. It collects State-level sales volumes, sales revenues, and number of customers by end-use sector (residential, commercial, industrial, and total). The existing sample and methodology to estimate population totals were described at the fall meeting of the Energy Committee. This session will describe current efforts to form homogeneous subpopulations for estimation, and evaluations based on predictions using annual data (where truth is assumed known). We assess the importance of outlier detection, under alternative scenarios. We will be asking for advice on next steps for analysis and on how to frame convincing recommendations for implementation.

Background

Form EIA-826 collects information from regulated and unregulated companies that sell or deliver electric power to end users, including electric utilities, energy service providers, and distribution companies. The Form EIA-826 is a monthly survey that prior to 2004 collected state-level sales volumes, sales revenues, and number of customers by end-use sector (residential, commercial, industrial, other (including public street and highway lighting), and total).

The Form EIA-826 uses three Schedules to collect information: Schedule A collects from full service providers (bundled electricity and delivery service to end users); Schedule B collects from marketers that provide electricity only service to end users (without delivery service); and Schedule C collects from utilities that own distribution lines that provide delivery only service to end users.

The respondent list for the EIA-826 consists of the following groups:

Respondent Classifications

Schedule	Respondent Group
Schedule A Electricity Generators	Census of IOU's
	Sample of non-IOUs
Schedule B Power Marketers	Census of Units Selling to End-Users
Schedule C Distribution	Census of Utilities

Form EIA-861 is used to collect retail sales of electricity and associated revenue by sector from all electric utilities, electricity service providers and distribution companies in the United States on an annual basis. It provides the frame for the monthly EIA-826. Hence the respondents to the EIA-826 are a subset of the respondents to the EIA-861.

Sample Design and Estimation for Non-sampled Companies

Schedule A is completed by a combination of a cut-off sample of full service providers and a census of Investor Owned Utilities (IOUs). Schedule B is completed by a census of marketers, and Schedule C is completed by a census of Utilities that provide delivery only service.

The number of companies reporting on the EIA-861 was 3,214 in 2002 and 3,215 in 2003. The companies reporting on Schedule A of the EIA-826 included 259 IOUs and 149 sampled units in 2002, and 261 IOUs and 170 sampled units in 2003.

The cut-off sample (i.e., sample of largest units) was selected using annual data for two different years to demonstrate that relative standard errors (RSE's) were smaller than 1% for residential, commercial, and industrial revenues, sales, and prices. Initially (in the late 80's and early 90's) estimates were done by State and virtually all units in the cut-off sample were used to estimate for the nonsampled units. Since that time, with changes in the industry and the addition of IOUs and power marketers as certainty units, the cut-off sample was adjusted to maintain a total number on the respondent list of fewer than 450.

Instead of making estimates separately by State, estimates are now made within 11 estimation regions that have similar weather and economic conditions. The 11 estimation regions are Alaska, Hawaii, NEA (CT,DE,DC,ME,MD,MA,NH,NJ,PA,RI,NY,VT); NEC (IA,MI,MN,WI); CEN (IL,IN,KY,MO,OH,TN,WV), NWC (MT,NE,ND,SD,WY); WES (CA,NV); NEW (OR,WA,ID); SEA (AL,FL,GA,NC,SC,VA); SOU (AR,KS,LA,MS,OK,TX); and SWE (AZ,CO,MN,UT).

By region, the sample coverage rate (data reported as a percent of monthly estimated total) is presented for each region in the table below.

AK	88.65%
CEN	78.68%
NEA	95.21%
NEC	83.12%
NWC	79.51%
NWE	75.23%
SEA	76.66%
SOU	68.17%
SWE	86.72%
WES	92.21%

Based on the data published for August 2003, the sample coverage rate by sector is 81% for residential revenue, 79% for residential sales, 88% for commercial revenue, 86% for commercial sales, 84% for industrial revenue and sales, 78% for other revenue, 76% for other sales, 84% for total revenue, and 82% for total sales.

Estimation for nonsampled companies in an estimation group is done using a regression equation of the form

$$y_{is} = \beta_s x_{is} + \varepsilon_{is} \quad (1)$$

Here y_{is} is the current EIA-826 data for company (i) in estimation region (s), x_{is} is the past EIA-861 data for company (i) in region (s), and ε_{is} is the error term, assumed to be normally distributed with mean 0 and variance $\sigma_s^2 x_{is}^{2\gamma}$. Based on comparisons that were conducted during the late 1980's and early 1990's, γ is currently taken to be 0.8. The coefficient β_s represents the seasonal or business cycle change from the annual data to the current monthly data in estimation region s. The seasonal or business patterns estimated by $\hat{\beta}_s$ are assumed to apply to the nonsampled companies, as well as the sampled ones. The estimated regression coefficient $\hat{\beta}_s$ is used to predict the monthly data for nonsampled companies, based on their EIA-861 data. Hence, $\hat{y}_{ks} = \hat{\beta}_s x_{ks}$ for the kth nonsampled company in region s. Once the estimated values are available for the nonsampled companies, the estimates and reported values together are used to prepare aggregates for States and other regions. (Note that estimation groups are strata of companies with similar seasonality, and are used for estimating for non-sampled companies. Once estimates are prepared for each company, aggregates can be prepared for publication for any region (State or larger)).

In 2003 most of the sampling error was associated with the "Other" category. This is due to the fact that the emphasis was on the three main sectors: residential, commercial, and industrial. The RSE tables in the *Electric Power Monthly* (EPM) publication provide information on sampling errors. The total numbers of RSEs greater than 10 in the EPM tables for November 2003 were:

Category	Revenue (\$-Mil)	Sales (MWh)	Price (¢/KWh)
Total data elements	300	300	300
Residential RSE >10	1	2	0
Commercial RSE > 10	1	2	0
Industrial RSE > 10	6	5	3
Other RSE >10	19	23	15
Total RSE > 10	1	2	0
Total	28	34	18

The RSEs appear to be relatively stable over time. While the RSE for "Other" tends to be large, it is no longer collected beginning in January 2004. (It has been replaced by a new category, "transportation.")

Any company that appears to have valid monthly data, but whose annual (EIA-861) data is inconsistent with the monthly data is treated as an "additive outlier." An additive outlier company's data are used to form monthly totals but are not used in estimation for other companies.

Evaluation Based on Predicting Annual Data from Annual Data

At the last meeting of the ASA Energy Committee we displayed scatterplots, standardized residual plots and other diagnostic tools in a thorough exploratory analysis to assess the quality of the fit of the model in (1). In that exercise we used both ten geographical regions (those described above except for HI), and a smaller set of four geographical regions North East (NEA, CEN, and NEC); North West (AK, NEW, NWC); South East (SEA, SOU); and South West (SWE, WES, HI). The diagnostic evaluation included regressions by region with all data included and with outliers and influential observations removed. We declared an observation to be an outlier if the absolute value of the standardized residual exceeds 3.5, and deemed an observation to be influential if DFFITS exceeds $2/n^{1/2}$.

These diagnostics indicated that the model (1) actually fits very well. But we had some residual questions to address: a) what is the value of using an automatic outlier detection and removal scheme? b) does the current stratification into estimation groups lead to acceptable results, and is there a better choice? c) should the monthly data from the IOUs be used in estimation for the small non-sampled companies? d) what value of gamma should be used routinely?

To illuminate the answers to these questions we conducted a comparison using annual data from the EIA-861. The data files from 2002 and 2003 were matched by respondent, and companies that report on the EIA-826 were identified as “sampled” companies. Companies that are IOU’s were also identified. With this data file, the regression equation in (1) was used with y_{is} as the 2003 data for company (i) in estimation region (s), and x_{is} as the 2002 data for company (i) in region (s). The data from the “sampled” companies is used to estimate $\hat{\beta}_s$. The model and estimated $\hat{\beta}_s$ are used to predict the 2003 data for nonsampled companies, based on their 2002 data. Hence, $\hat{y}_{ks} = \hat{\beta}_s x_{ks}$ for the k^{th} nonsampled company in region s. Once the estimated values are available for the nonsampled companies, the estimates and reported values together are used to prepare aggregates for States and other regions.

The advantage of this approach is that we can compute summaries concerning the estimation error for the nonsampled companies under various hypotheses. Several statistics were computed to summarize results. Let , $\hat{e}_{ks} = y_{ks} - \hat{y}_{ks}$ represent the error in using the approach above to estimate for the nonsampled company, k, in region, s. Let n_s represent the number of sampled companies in stratum s; and N_s represent the total number of companies within the population in stratum s.

1) The average estimation error within stratum, s, is given by
$$\bar{e}_s = \frac{1}{N_s - n_s} \sum_{k=1}^{N_s - n_s} \hat{e}_{ks} .$$

This is expressed as a percent by dividing by the total of the nonsampled companies in stratum s , $Tns_s = \sum_{k=1}^{N_s-n_s} y_{ks}$.

2) The RMSE of the estimation errors in stratum s is computed as

$$RMSE_s = \text{sqr}t\left(\frac{1}{N_s - n_s} \sum_{k=1}^{N_s-n_s} \hat{e}_{ks}^2\right).$$

This is expressed as a percent by dividing by Tns_s

3) The CV is the RMSE divided by average of nonsampled units $Tns_s / (N_s - n_s)$.

4) Percent error in estimating the nonsampled companies in the stratum:

$$E_s = (N_s - n_s) \bar{e}_s * 100 / Tns_s.$$

5) And the percent error in estimating the stratum total.

$$E_s = (N_s - n_s) \bar{e}_s * 100 / (Tns_s + Ts_s). \quad \text{Where } Ts_s = \sum_{k=N_s-n_s+1}^{N_s} y_{ks}$$

These summary statistics are also computed at the US level, and these are displayed for residential sales, commercial sales, and industrial sales in the attached tables. There are still some anomalies in the data that will be investigated before the meeting.

The table in each attachment shows the baseline case with the current stratification and data from IOUs not used in estimating for the nonsampled companies. There are 6 alternatives shown in the row of that top table.

- a. Estimation with gamma=.5
- b. Estimation with gamma=.8
- c. Estimation with gamma=.5 and one pass of outlier detection and removal
- d. Estimation with gamma=.8 and one pass of outlier detection and removal
- e. Estimation of gamma using a two stage least squares procedure
- f. Estimation of gamma using a two stage least squares procedure and one pass of outlier detection and removal.

The second table in each attachment shows the results using current stratification but including the IOUs among the sampled companies in estimation for the nonsampled companies. The six alternatives above are repeated.

The third table uses an alternative stratification of the States into estimation groups as discussed below, and does not include IOUs in the estimation for non-sampled companies. The six alternatives above are repeated.

Finally the fourth table uses the alternative stratification and includes data from the IOUs among the sampled companies to estimate for the nonsampled companies.

Composition of “estimation groups” (i.e., post-strata)

We have been considering the ten region groups that have been used to define estimation groups. For this study we developed new estimation groups based on the seasonality evident in the EIA-826 data by State. For two years, for each company in the EIA-826 sample, the company level β 's by month were computed as the monthly data divided by the average of the monthly data over the year from that company. Hence the company level β 's, like the regression coefficient vary around 1. We based the stratification analysis on the company level β 's for February and for August because these months seem to best characterize differences in seasonality. For each state the average of the company level β 's for February and for August was computed, averaging over 2002 and 2003.

For the stratification used in this analysis, States were grouped in a heuristic approach roughly based on the following: First the value of the February number was used to define groups – of particular importance is whether it was greater than 1 or less than 1. In the Pacific Northwest, for example, there is no peak during the winter months. Once that grouping was done, the difference between February and August was used to classify the extremes of seasonal swings. The new stratification groups will be illustrated in presentation materials.

The assessment described to date does have its limitations. For example, the real purpose of the stratification into estimation groups is to capture regional seasonality. An evaluation based on annual data cannot really demonstrate that aspect of the stratification. We are considering a simulation study based on monthly data.

Questions for the Committee

1. Do you have alternative suggestions for the summary statistics to use for assessing alternatives?
2. Do you have any suggestions for follow-on studies, particularly the simulation study?
3. Any insights into the methodology we used to develop the new stratification?
4. Any suggestions for how to best package the results of this study to convince managers to implement changes.

Residential sales -- baseline (current stratification, IOUs not used in estimation)

experiment	Percent Ave Est Error	Percent RMSE Est Error	Percent CV - Est Error	Percent Error in NS	Percent Error in US total
Gamma=.5	0.00	0.00	12.11	1.56	0.32
Gamma=.8	0.00	0.02	66.75	11.35	2.36
Gamma=.5 with outlier	0.00	0.00	10.54	-0.20	-0.04
Gamma=.8 with outlier	0.00	0.00	10.66	0.46	0.09
Est gamma	0.00	0.00	11.91	1.17	0.24
Est gamma with outlier	0.00	0.00	9.70	0.02	0.00

Residential sales -- current stratification, IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.00	11.51	1.72	0.36
Gamma=.8	0.00	0.02	45.02	8.17	1.70
Gamma=.5 with outlier	0.00	0.00	11.24	1.22	0.25
Gamma=.8 with outlier	0.00	0.00	11.65	1.30	0.27
Est gamma	0.00	0.00	11.47	1.32	0.27
Est gamma with outlier	0.00	0.00	10.98	0.98	0.20

Residential sales -- new stratification

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.00	9.67	0.28	0.06
Gamma=.8	0.00	0.02	56.52	6.55	1.36
Gamma=.5 with outlier	0.00	0.00	7.57	-0.63	-0.13
Gamma=.8 with outlier	0.00	0.00	7.96	-0.16	-0.03
Est gamma	0.00	0.06	151.83	12.14	2.53
Est gamma with outlier	0.00	0.06	151.72	11.06	2.30

residential sales -- new stratification, IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.00	9.97	1.61	0.33
Gamma=.8	0.00	0.02	46.15	6.40	1.33
Gamma=.5 with outlier	0.00	0.00	8.65	0.96	0.20
Gamma=.8 with outlier	0.00	0.00	8.88	0.93	0.19
Est gamma	0.00	0.00	11.17	1.11	0.23
Est gamma with outlier	0.00	0.00	9.22	0.91	0.19

Commercial sales -- baseline (IOUs not used for estimation, current stratification)

	Percent	Percent	Percent	Percent	Percent	Percent
experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total	Error in US total
Gamma=.5	0.00	0.03	83.96	8.44		1.21
Gamma=.8	0.01	0.05	121.31	13.94		2.01
Gamma=.5 with outlier	0.00	0.03	66.87	6.23		0.90
Gamma=.8 with outlier	0.00	0.03	88.36	9.67		1.39
Est gamma	0.00	0.03	67.20	2.17		0.31
Est gamma with outlier	0.00	0.02	64.42	6.03		0.87

Commercial sales -- IOUs included in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.03	73.28	8.72	1.26
Gamma=.8	0.01	0.04	107.99	15.31	2.20
Gamma=.5 with outlier	0.00	0.02	57.34	4.91	0.71
Gamma=.8 with outlier	0.00	0.02	61.70	5.99	0.86
Est gamma	0.00	0.02	57.44	1.97	0.28
Est gamma with outlier	0.00	0.02	53.11	3.17	0.46

Commercial sales -- new stratification

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.03	80.21	7.39	1.06
Gamma=.8	0.00	0.04	114.51	11.70	1.68
Gamma=.5 with outlier	0.00	0.03	66.24	6.90	0.99
Gamma=.8 with outlier	0.00	0.03	75.37	7.53	1.08
Est gamma	0.01	0.08	203.86	15.65	2.25
Est gamma with outlier	0.01	0.08	203.96	18.81	2.71

Commercial sales -- new stratification, IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Gamma=.5	0.00	0.03	75.58	9.99	1.44
Gamma=.8	0.01	0.04	105.23	14.17	2.04
Gamma=.5 with outlier	0.00	0.02	61.10	6.08	0.88
Gamma=.8 with outlier	0.00	0.02	64.33	6.39	0.92
Est gamma	0.00	0.02	58.05	1.26	0.18
Est gamma with outlier	0.00	0.02	56.68	5.14	0.74

Industrial sales -- baseline (current stratification, IOUs not used in estimation)

experiment	Percent Ave Est Error	Percent RMSE Est Error	Percent CV - Est Error	Percent Error for NS	Percent Error in US total
Gamma=.5	0.00	0.03	60.57	6.06	0.94
Gamma=.8	0.01	0.05	79.86	13.60	2.11
Gamma=.5 with outlier	0.00	0.03	57.44	1.76	0.27
Gamma=.8 with outlier	0.00	0.03	57.76	1.60	0.25
Est gamma	0.00	0.03	56.32	0.53	0.08
Est gamma with outlier	0.00	0.03	57.75	0.69	0.11

Industrial sales -- IOUs included in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error for NS	Error in US Total
Gamma=.5	0.00	0.03	56.95	5.15	0.80
Gamma=.8	0.01	0.04	64.20	9.65	1.49
Gamma=.5 with outlier	0.00	0.03	56.19	3.39	0.53
Gamma=.8 with outlier	0.00	0.03	56.39	3.15	0.49
Est gamma	0.00	0.03	56.05	1.71	0.26
Est gamma with outlier	0.00	0.03	56.60	1.91	0.30

Industrial sales -- new stratification

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error for NS	Error in US Total
Gamma=.5	0.00	0.03	59.94	5.06	0.78
Gamma=.8	0.01	0.04	73.71	10.39	1.61
Gamma=.5 with outlier	0.00	0.03	56.62	1.76	0.27
Gamma=.8 with outlier	0.00	0.03	56.23	1.14	0.18
Est gamma	0.01	0.08	139.71	10.05	1.56
Est gamma with outlier	0.01	0.08	139.09	11.05	1.71

Industrial sales -- new stratification, IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error for NS	Error in US Total
Gamma=.5	0.00	0.03	59.21	5.95	0.92
Gamma=.8	0.01	0.05	78.80	15.74	2.44
Gamma=.5 with outlier	0.00	0.03	57.42	3.50	0.54
Gamma=.8 with outlier	0.00	0.03	57.96	4.34	0.67
Est gamma	0.00	0.04	61.27	0.83	0.13
Est gamma with outlier	0.00	0.03	55.86	2.64	0.41