Some Interesting Statistics Problems Arising from Analyses of Energy Data from Surveys

Carol Joyce Blumberg
cblumberg@gmail.com

Abstract

The author has recently encountered many interesting problems when working with energy data that involved development of new statistical techniques or novel adaptations of existing techniques. The purpose of this talk is to present some of these problems. Both solved and unsolved problems will be discussed. The hope is that the presentation of these problems will serve as seeds for research by others. Some of the problems that will be discussed are (i) how to best measure a rate of change when the data used to form the numerator and denominator come from different, but comparable, sources; (ii) how to estimate weekly values from data collected at a monthly level; and (iii) how to estimate values, that came in the past from aggregating data together from two surveys, when one of the surveys is discontinued temporarily

Keywords: correlation, establishment surveys, industrial statistics, measurement, regression, small area estimation

1. Introduction

Recently the author has had the opportunity to work on several studies involving energy data from establishment surveys that involved some interesting statistical problems. The studies and statistical methodologies employed will be presented in the next four sections. In addition, in each section the unsolved statistics problems will be discussed in the hope that they will serve as seeds for research by others.

2. Study 1: Determining which of nine methods of measuring rate of change is best

The U.S. Energy Information Administration (EIA) collects data weekly and monthly on volumes of certain petroleum products. The weekly volumes are collected from a sample of companies in the sampling frame and the monthly volumes are collected from all of the companies in the frame. These volumes are often reported in print in units of thousand barrels per day. What EIA calls Monthly-from-Weekly (MFW) weighted averages (with the weights for each week being the proportion of days in a particular month that fall in that week) are calculated using the data collected weekly and published within 12 days of the end of the month. In some instances, EIA also publishes running 4-week averages. Since the units are in thousand barrels per day, the 4-week (4W) average that covers the most days of the month of interest can be used as a measure of monthly volumes. EIA also calculates preliminary monthly (PM) volumes based on the data received by the 20th of the following month. The data used for the PM volumes are carefully examined and discrepancies discussed with the respondents before being published approximately 60 days after the end of the month of interest. However, EIA then waits until approximately July of the following year and publishes final revised volumes (FV).

The rate of change that is of high importance to financial analysts is $R = \left[\left(\frac{Volume\ for\ a\ certain\ month\ in\ year\ t}{Volumes\ for\ a\ certain\ month\ in\ year\ t-1}\right) - 1\right] * 100\%$. The gold standard is to use the FV's in both the numerator and denominator of R. However, the financial analysts do not want to wait until the FV's are published (which can be from 7 to 19 months after the

end of the month of interest) to calculate this ratio. The question investigated in Study 1 was as to whether MFW, 4W or PM could be used instead in the numerator and/or denominator to obtain valid estimates of the quantity R. The nine estimates of R, including the gold standard of Formula 1, that were compared are listed in Table 1.

Table 1. The 11 Estimates of R Investigated in Study 1

			Numerator	
Denominator	**FV**	**MV**	**MFW**	**4W**
FV	Formula 1	Formula 2	Formula 4	Formula 7
MV		Formula 3	Formula 5	Formula 8
MFW			Formula 6	
4W				Formula 9

The interesting statistical issue that is still an open question and needs more research is:
Question 1: What measure(s) are the best statistically to determine which formulas best approximate Formula 1? The measures could be any of those listed in the next paragraph or perhaps other measures.

The following were used in this study: (i) bias—as defined by the mean for Formula 1 minus mean of each of the other 8 formulas; (ii) other simple descriptive statistics such as median, quartiles, and range; (iii) standard deviation of the difference between Formula 1 and each of the other 8 formulas; (iv) mean square error of the differences between Formula 1 and the other formulas as defined by $\sqrt{(Bias)^2 + (SD\ of\ differences)^2}$; (v) paired t-tests using the estimates of R from Formula 1 versus using each of the other formulas; (vi) percentage of the time that each of the other formulas was within x% of Formula 1 (1% and 2% were used in this study); (vii) percentage of time Formula 1 and each of the other 8 formulas have the same sign. This last measure is important because for this study the ratio being investigated was a growth rate in supply of certain petroleum products. Even if two formulas are close by measures (i) to (vi), having a wrong sign is a major problem psychologically; and (viii) correlation of Formula 1 with each of the other formulas. See Blumberg (2007) for more details.

It should be noted that data were available for all months for 28 years (i.e., 336 months) for this study. But, in many datasets similar to this one, fewer months may be available. So, another question worth investigating is:
Question 2: Does the answer to Question 1 depend on sample size?

3. Study 2: Estimating weekly values when only monthly values are available
This purpose of this study was to try to improve the autoregressive models being used (Burdette & Zyren, 2002 and 2003) to predict three days in advance and eight hours in advance (using the same models both times) what the EIA published weekly changes in the retail prices for regular grade motor gasoline (all formulations combined) and diesel fuel would be for 10 regions of the USA. The Burdette & Zyren models were based on changes in closing spot prices. But, 18 out of the 20 regional predictions involved more than one spot price. Hence, before building each of these models the relevant spot prices had to be volume-weighted. The problem was that the changes in spot prices were measured weekly, but the appropriate volumes were based on data collected monthly by EIA. In the models being used at the time, the volumes used were fixed at a certain year's total volumes and never changed. In updating the models it was desired to try to have the volumes change over time, perhaps even weekly by using estimated volumes based on

the previous year's volumes for the same time period. Some literature could be found on disaggregation of the main data from a lower frequency (in this case, monthly) to a higher frequency (in this case, weekly). However, no literature could be found on disaggregation involving weights. For this study it was decided to try cubic splines and quadratic matching methods of disaggregation on the volumes being used as weights. Neither method provided any meaningful improvement over using the monthly weights and the use of monthly weights provided no meaningful improvement over using the yearly volumes for a fixed year. This failure to find an appropriate method leads to two open questions:

Question 3: What methods of disaggregation work best for energy price (e.g., crude oil spot prices) and volume (e.g., volumes of residual fuel oil) data that can have high volatility over short periods of time?

Question 4: Does the answer to Question 3 change if the disaggregation is for a main variable of interest or for a weighting variable?

4. Study 3: Estimating prices from the combination of two surveys when one of the surveys is discontinued temporarily

Until 2011, EIA collected price and volume data on sales to consumers for a variety of petroleum products from the producers, resellers, wholesalers, and retailers using two forms: EIA-782A, "Refiners'/Gas Plant Operators' Monthly Petroleum Product Sales Report" and Form EIA-782B, "Resellers'/Retailers' Monthly Petroleum Product Sales Report". The EIA-782A (abbreviated as Form A) was a monthly census. The EIA-782B (abbreviated as Form B) was completed monthly by a stratified random sample of respondents, with its published estimates being computed using appropriate estimation formulas. By law, both surveys were mandatory for respondents to complete. Hence, each had high response rates every month and very little data needed to be imputed each month. Until 2011, the published prices for ten products in EIA's State Energy Data Systems (SEDS—see http://www.eia.gov/beta/state/seds/seds-data-fuel.cfm?sid=US) were computed at the state (or regional levels when publication would violate confidentiality due to a very low number of sellers of a particular product in a state) for each of the 50 States, the District of Columbia (referred to as 51 states for the rest of this paper) and for 9 regions using the following formula: SEDS Published Price =

$$\frac{(\text{Estimated Price}-\text{Form A})\cdot(\text{Estimated Volume}-\text{Form A})+(\text{Estimated Price}-\text{Form B})\cdot(\text{Estimated Volume}-\text{Form B})}{(\text{Estimated Volume from Form A}+(\text{Estimated Volume from Form B})}$$

In March 2011, Form B was discontinued. The study discussed in the rest of this section was undertaken to answer the question of how to derive reasonable published prices for SEDS for 2011 and 2012 from only Form A data that closely approximate the SEDS prices that would have been derived from the combination of Forms A and B, if both still existed. It was decided to use linear regression with Form A estimated prices as the independent variable and computed SEDS prices for Forms A and B combined for the years of 1994 to 2010 as the dependent variable. It was also decided to form these regressions equations using the non-suppressed (i.e., both the published and unpublished) prices and volumes, which were made available to the author for this study.

Although, for each of the 10 products, it would be most desirable to have developed a regression equation for each of the 51 states separately, this would have required development of 510 regression equations. Hence, it was decided to develop only 90 regression equations for the 9 regions (for each of the 10 products) and then apply those equations to the states within each region. The only publication that could be found

where a method similar to this was used was Kim, Kim, and Park (2012). This leads to the following open question:

Question 5: Under what statistical conditions is it valid to develop regression equations on a larger region and then apply those regression equations to data collected on subdivisions of the larger region?

The published values in SEDS are at the annual level. However, the data from Form A and Form B go back only as far as 1994, for a total of only 17 years plus 2 months of data. This is clearly not enough observations to develop equations using the annual-level data. However, the annual prices are weighted averages (by volume) of the monthly level data collected for and published in *Petroleum Marketing Monthly* (http://www.eia.gov/petroleum/marketing/monthly/). Hence, it was decided to use the Form A and Form B non-suppressed estimates at the monthly level to form the regression equations. This gave 204 observations per region for development of the equations, except in a few cases where a particular product was not sold in a particular month in a particular region (e.g., distillate fuel used for residential heating in the Gulf Coast region in some summer months). Forming the regression equations using monthly level data leads to the following open question:

Question 6: Under what statistical conditions is it valid to develop a regression equation on monthly-level data and then apply those regression equations to data that have been aggregated to the annual level?

Although the monthly observations are clearly a time series, for this study it was felt that the time-series nature of the data could be ignored. Since outliers can have major effects on regression equations, it was decided to use cross-validation. In addition, since the regression equations would be used on 2011 and 2012 data, it was decided that having the models fit the monthly data for 2009 and 2010 was of high importance. Hence, the 206 months were divided into three mutually exclusive subsets: Set 1-A simple random sample of 90 of the first 180 months (from Jan. 1994 to Dec. 2008); Set 2-The remaining 90 of the first 180 months; and Set 3-The 26 months from Jan. 2009 to Feb. 2011. The models were developed separately for Set 1 and Set 2. In the cases where the models built separately for Set 1 and Set 2 did not agree, additional outliers were deleted until the models agreed. The models were then validated using Set 3. In some instances further modifications were made so that the models fit Set 3 also. Two open questions related to this type of model building are:

Question 7: Under what statistical conditions is it valid to develop a regression equation on time-series data when the time-series nature of the data is ignored?

Question 8: What methods of cross-validation are statistically viable for regression models built on time-series data when the time-series nature of the data is ignored?

5. Some additional open questions that seem to arise regularly in analyzing energy data from establishment surveys

In the statistics literature there have been several methods suggested as to how to handle the situation that some summary data (that is, sensitive summary data) need to be suppressed in order to protect the confidentiality of corporations or individuals and at the same time minimize the proportion of the non-sensitive data that are suppressed.

Question 9: Under what conditions are each of the present methods of suppression best for energy data?

Question 10: Can new suppression methods be developed that are best for energy data under a wider variety of conditions than the present methods?

As the use of the Internet, rather than paper, for presenting published data has increased, the question of what metadata and other auxiliary information should accompany the published data has become more important. While one could use pdf or similar files and keep the information in the same format as paper, one major advantage of the Internet is that more flexible formats are possible. However, the amount of metadata and other auxiliary information that can be presented along with the data may be reduced. Hence, an interesting question for study is:

Question 11: What types of metadata and auxiliary information about the data presented on a webpage should be on the webpage itself and what types should be provided in links from the webpage?

6. Disclaimer and Acknowledgements

The views expressed in this article are those of the author. The author wishes to thank Irena Ograjenšek of the University of Ljubljana for suggesting the topic for this paper.

References

Blumberg, C. (2007). *Comparison of Methods for Computing Yearly Growth Rates from Weekly and Monthly Data, 1978 to 2005*. Presented at Federal Committee on Statistical Methodology 2007 Research Conference. Paper available at http://www.fcsm.gov/07papers/Blumberg.II-B.pdf.

Burdette, M. & Zyren, J. (2002). *Diesel Fuel Price Pass-through.* Available at http://www.eia.gov/pub/oil_gas/petroleum/feature_articles/2002/diesel/diesel.html.

Burdette, M. & Zyren, J. (2003). *Gasoline Price Pass-through.* Available at http://www.eia.gov/pub/oil_gas/petroleum/feature_articles/2003/gasolinepass/gasolinepass.htm.

Kim, J., Kim, S., & Park, S. (2012). *Small-Area Estimation Combining Information from Several Sources*. Joint Statistical Meetings 2012 Proceedings, 3489-3499. Online at https://www.amstat.org/membersonly/proceedings/2012/data/presinfo/presinfo34079.cfm