IMPUTING PRICE AS OPPOSED TO REVENUE IN THE EIA-782 PETROLEUM SURVEY

Pedro J. Saavedra, Paula Weir and Michael Errecart
Pedro J. Saavedra, Macro International, 8630 Fenton St., Silver Spring, MD 20910

Imputation, survey, exponential smoothing, nonresponse.

The EIA-782B is a monthly price and volume survey of petroleum resellers and retailers. Every month, preliminary results are published for the current month and final results for the previous month. Missing data and data that fail the edits are imputed using a combination method which involves predictive ratios of forecasted volumes and revenues. The predicted value is obtained using exponential smoothing for volume and revenue. This approach has the disadvantage that the same alpha coefficient affects both revenue and volume, while there is some evidence that more recent values are better predictors for price than for volume, thus suggesting that different alpha coefficients for price and volume would be more efficient. An alternate approach—predicting price and volume separately, and using an imputed price for exponential smoothing when the volume is zero—was simulated under different conditions and compared with the current approach.

The Current Procedure and its Alternatives

The EIA-782B is a monthly price and volume petroleum survey through which data are gathered for a variety of product/end-user combinations. Prices are reported at the State level, though prices and volumes are not published for every product--State combination. The EIA-782B focuses on resellers, but its estimates are combined with the EIA-782A—a census of refiners. Though the sampling design is complex (Saavedra, 1988), for purposes of this discussion one can think of the design as a set of stratified samples, where in each State there are different stratifications for different variables in the survey.

The imputation procedures are described in the Petroleum Marketing Monthly publication. We will briefly describe the approach. For every Company-State Unit (CSU), separate time series of volume and revenue values are carried for each product/end-user combination reported in the survey to yield a historical volume and a historical revenue. These historical values are obtained through exponential smoothing and updating for changes in the market so that the previous month dominates the value, but other months also contribute. The degree to which the previous month dominates the value varies by product, so that in some cases the previous month is most important and in others the adjusted historical average is the critical figure. A ratio of reported to historical values is then obtained for each cell for volume and a similar one for revenue, using only respondents (cells that do not have at least two respondents selling the product are collapsed). This ratio is multiplied by the historical value for nonrespondents, thus yielding an imputed value. This approach is called a chain link.

The chain link as a general principle seems like a reasonable method of combining the individual Company-State unit's historical value with the changes in the market since the previous month. Indeed any system must take into account both the historical value and the changes since that value was reported. However, even if one accepts the chain link as a basic approach, there are many issues that need to be resolved. We identified the following issues as critical in this study:

- What variables should be used to obtain the historical values? Should price and volume be used instead of revenue and volume?
- Should exponential smoothing be used, or merely previous values?
- What exponential smoothing coefficients should be used?
- Should exponential smoothing coefficients be obtained separately for each product? For each product and form (EIA-782A and EIA-782B)?

- Should volumes be treated as a two stage equation—taking into account probability of reporting and volume if the company reports a positive volume?

- How should the ratio of current value to previous value be estimated? If by PADD, should it be done separately by form?

- Within what unit should the ratio be estimated, and if a small unit (e.g. a cell) how should one collapse if there are too few respondents?

- What initial values should be used for price and volume?

Clearly the possibility exists that the optimal answer for each of these questions will be different depending on how other questions are answered. For instance, the optimal unit for estimating the ratio could depend on the estimation approach, which in turn could depend on the variables chosen. Furthermore, the solution may be different for different products and States. In order to make the analysis manageable, the different issues were resolved one at a time, and each issue was considered settled as the next one was considered.

In a number of instances the decision was dictated by ease of programming in the absence of a clear cut statistical reason for preferring the more difficult method. For example an analysis indicated that carrying six months of volumetric data may be marginally superior to exponential smoothing, but far more complicated. The simpler method was chosen.

The following sections discuss each one of the issues presented.

Domain Used for Updating the Ratio

This issue was investigated in 1991 and presented in a previous paper (Saavedra and Weir, 1991). If one needs to determine whether companies like a particular respondent have changed their volume or price since the prior month, is it better to aggregate data within the cell (and if so how should the cells be combined if there is not a sufficient number of respondents in the cell), a stratification level (whether geographic or volumetric), the State, or the PADD? The smaller domains can be expected to be more homogeneous, but the larger domains can be expected to provide more stable results. An empirical simulation showed stability to be more important than homogeneity and, thus, the PADD was chosen as the unit for which the ratios would be calculated.

The issue of whether we should estimate separately for each form remains to be determined. Simulations at various stages suggested that we should estimate the ratio separately by form for price, but not for volume. The final decision was made as part of the final analysis, presented in a subsequent section.

Units to be Imputed

The current approach uses revenue and volume separately. This has the advantage that there is always a value present, even if it is zero. An analysis was conducted using companies that reported positive volumes for four consecutive years. An analysis of individual company results with product-PADD aggregate data indicated that correlations were relatively high for all companies when it came to price, but there were many negative correlations when it came to volumes.

These results suggests that one should be able to make relatively good predictions for aggregate prices, but not for aggregate volumes. Thus we can expect within-company fluctuations of volume to be less dependent on what other companies are doing. On the other hand, when one company raises prices, all other companies raise prices. This suggests that a different strategy should be followed for price than for volume.

It should be noted that while there was a more uniform pattern for prices than for volumes, this varied by the product. For example, for residential fuel oil the pattern for volumes was much more uniform than for other products, but still less uniform than for prices.

Indeed if exponential smoothing or some other approach were used, one might wish to use different coefficients for price and volume. One would still weigh imputed price by imputed volume, but each would be achieved independently. This leaves open the question of how to treat months where there were no sales of the product in a given State. The answer is given to us by the approach currently used for imputing for chronic nonrespondents. A price would be imputed for the Company-State Product Unit (CSPU) when it reports zero volume, and this imputed price would be treated as if it were the reported price. Naturally, since the reported volume would be zero, this imputed price would have no effect on the estimated price for that month.

Exponential Smoothing and Other Decisions

Simulations with various approaches suggested that exponential smoothing including zeros for volume worked as well as any other approach, and had the virtue of preserving continuity with what had been previously done. It was also found that not using initial values, but assuming the product was not sold until the first report, was as effective as any other approach.

Parameters for Exponential Smoothing

The use of ARIMAS, given the indication that different companies were in different cycles was deemed inappropriate. Regression was considered, and would have been appropriate had the parameters fallen within specified bounds (requiring descending parameters between zero and one). While a number of equations fit these requirements, many products yielded regression equations that did not.

At last the following procedure was decided upon. First, a database was created with all records for which six lags existed for price, and a separate database was created with all records for which six lags existed for volumes (using four years data, but including respondents for only part of this period). Each lag was multiplied by the appropriate ratio (the ratio of current value to the lag for the product and PADD) in order to update it to the equivalent of the criterion month. Then the equation for the imputed value was calculated using one alpha parameter. This equation was:

$$HV_t=V_{t-1}+\alpha(V_{t-2}-V_{t-1})+\alpha^2(V_{t-3}-V_{t-2})+\ldots+\alpha^5(V_{t-6}-V_{t-5})-\alpha^6V_{t-6}$$

where the lags are updated by the appropriate ratio and A corresponds to the alpha coefficient.

The first attempt to obtain an optimal value for alpha was to use nonlinear regression. However, the SAS nonlinear regression algorithm proved to be prohibitively slow, particularly given the large number of product-form combinations. Instead the squares of the difference between predicted and actual value was minimized for ten values (.05 to .95) and then a second search for a minimum within .05 of the minimum identified first was carried out. Thus an optimal value of alpha was identified for each product-form combination.

Obtaining the ratio

Up to now it had been assumed that obtaining an estimate for the entire cell, State or PADD, should be used to estimate the likely increase of a given CSPU. But it may not be the appropriate estimator. For example suppose a company sold 60% of the volume of a certain product in a PADD. Suppose further that twenty other companies sold the remaining 40%. It can be argued that the price increase of the nineteen other small companies is a better estimator for a nonresponding one's increase than the price increase for the PADD, controlled as it is by one atypical company.

Thus there are four ways of weighing prices (both historical and reported) which can be compared. One is a simple average of all CSPUs selling the product in a given PADD. The second is an average weighted by the company adjusted weight (but not by volume). A third is weighing by weight times reported volume, which is essentially obtaining the average PADD price after eliminating nonrespondents. The fourth is using volume as if it were a weight, but ignoring the weights.

In addition there are three ways of estimating the ratio. One approach uses averages (weighted or not) for the historical and reported prices. In other words, a historical price is obtained for the PADD and a reported price is obtained for the PADD and then the ratio is calculated.

A second approach obtains a ratio for each CSPU with both a historical price and a reported price, and then obtains a (possibly weighted) average of the ratios.

Finally, an approach similar to the first is implemented, but the estimates are based on averages (possibly weighted) of the logs of the prices, taking then an exponential to estimate the PADD price. This is something similar to the approach that was found optimal in the 1991 study.

For volumes only the simple average and the weighted average make sense.

In order to estimate the precision of these approaches the difference between imputed and reported value was obtained where both existed and its mean square error (mean squared plus variance) was calculated. In all cases the simple unweighted volume proved to be the better predictor. For the volumes the ratio of mean reported volumes to mean historical volumes (including zeroes in the calculation of both means) proved optimal. For prices the simple unweighted price average also proved best. However, these results were obtained at the company level, and the use of weights was sufficiently close that it was necessary to examine the various methods at the publication cell level.

Not every combination of approaches could be subjected to evaluation. Instead, only the variation of the calculation of the updating ratios was preserved in the evaluations. Thus the evaluations used four approaches in their final form (at least two other approaches were dismissed after they proved worse than any of the others). These were:

- Ratios obtained using unweighted averages of respondents' new and historical volumes and prices by the Product-PADD-Form combination.
- Ratios obtained using weighted averages of respondents' new and historical volumes and prices by the Product-PADD-Form combination.
- Ratios obtained using unweighted averages of respondents' new and historical volumes and prices by the Product-PADD combination.
- Ratios obtained using weighted volume averages for the Product-PADD combination and using unweighted price averages for the Product-PADD-Form combination.

The analysis used nonrespondents at the time the preliminary file was prepared who became respondents by the time the final file was completed. The old imputed values were obtained from the preliminary blended file, and the reported values from the final file. The absolute value of the difference of the two estimates at the cell level (using only cells for which there were nonrespondents) were then used as the criterion. It should be noted that the cells are not independent, not only because the same companies report for different products and months, but also because PADD estimates were included as well as State estimates (using only PADD estimates for distillate data from nonpublication States).

The first important result was that all of the new methods performed better than the existing method. This was true whether one used the mean absolute difference, the mean squared difference or looked at the maximum difference. True, there were individual cells where the old method performed better, but these were not systematically grouped. For 52 percent of the volume estimates and 44 percent of the price estimates the mixed method outperformed the existing one, with the existing method doing better for 37% of the volumes and 29% of the prices. The remaining values yielded identical results.

It soon became apparent that unweighted averages, separately by PADD and Form worked best for prices, but weighted estimates done for the PADD only worked best for volumes. The combination of these two was then examined. This led to a price result slightly worse than the first (but not significantly so) and a better volume result. Furthermore, when the number of cells in which one method or the other was better was examined, this mixed method worked best. Thus, it is the recommended method.

Summary

An examination of the imputation procedures of the EIA-782B Petroleum Survey
indicated that while the chain link was an appropriate procedure, several changes were appropriate. A previous study had shown that the PADD or the State were the appropriate units for which to obtain the aggregate statistics used in the imputation. The present study indicates that rather than imputing volume and revenue using the same coefficients to obtain a historical value, it is preferable to impute volume and price, using different coefficients for each. In addition the best estimates for prices to be used in calculating ratios were derived from unweighted averages, taken separately from the EIA-782A and the EIA-782B. However, weighted averages combining the two forms provided the better estimates for volumes.

Bibliography

Saavedra, P.J. 1988. Linking multiple stratifications: Two petroleum surveys. Proceedings of the 1988 Joint Statistical Meetings, American Statistical Association Survey Section, 777-781.

Saavedra, P. J. and Weir. P. 1992 Imputation in the EIA-782B Petroleum Product Survey -- a simulation study. Paper presented at the 1992 Joint Statistical Meetings, Boston Mass.

Acknowledgements

The authors wish to thank Richard Mantovani and Benita O'Colmain who acted as peer reviewers for this paper.