RESULTS OF MODEL-BASED APPROACH TO SAMPLING

Yahia Z. Ahmed and Nancy J. Kirkendall
Energy Information Administration
U.S. Department of Energy

1.0 Introduction

The Department of Energy, Energy Information Administration, Office of Oil and Gas is responsible for publishing timely and accurate weekly estimates of total inputs to refineries, total production of individual petroleum products, and total inventories of crude oil and individual petroleum products at the national level. At present these weekly estimates are produced using weekly data from 202 refineries in conjunction with monthly census survey data (available with a three month lag) from the 321 U.S. refineries. See Kirkendall and Kolmar (1980) for a more complete description.

A project is underway to redesign the weekly survey system to enable estimation of all of the above items in each of the 13 refining districts which make up the United States. The sample design constraints are:

1. Forty-four companies (154 refineries) must be included in the sample.
2. Estimates must be produced for totals of 17 items in 13 refining districts, with 5% margin of error.

The sampling problem is complex because of the large number of estimates desired and the small, skewed nature of the population. There are 321 refineries located in the United States (See Table 1). Not all of these refineries produce or store all products. Because the distributions of the individual item volumes in a month are highly variable and skewed, traditional sampling techniques produced a very large sample, almost a complete census. Even after relaxing the precision constraint, the traditionally selected sample encompassed approximately 85% of the population.

A relatively large sample is undesirable because of data processing problems (final estimates are made two days after the forms are due) and respondent burden. Hence, the model-based (superpopulation) approach was suggested to select the sample and produce estimates. Preliminary results indicate that this approach will lead to a smaller sample size than traditional probability sampling, and should provide the desired level of precision.

This paper presents the results of this phase of the study. Section 2 summarizes the predictive model-based approach as presented in Royall (1970). Section 3 describes the application of this approach to the weekly problem, and describes the procedures used to select the appropriate model. Section 4 describes the two methods used for sample selection. Section 5 presents the results of the study. This paper is an abbreviated version of a more comprehensive paper which can be obtained from the authors.

2.0 Background - The Predictive Model-Based Approach

The model-based approach assumes that the finite population under study has itself been generated as a random sample from an infinite superpopulation. In mathematical terms, the actual population $(y_1, y_2, \ldots, y_N)$ is the realized outcome of the random vector $(Y_1, Y_2, \ldots, Y_N)$, having an N-dimensional joint probability distribution ξ. The superpopulation distribution ξ is usually modeled to reflect the available background knowledge. Specification of ξ can vary from something crude and basic to a very detailed description, depending on what assumptions the analyst feels are legitimate.

Following Royall (1970), the population of interest consisted of N indentifiable units labelled 1, ..., N. Associated with unit i are two numbers, x_i and y_i, with x_i known and y_i fixed but unknown. A sample, denoted s, of size n is to be selected from the N units, and the y values associated with the sample units are to be observed. The objective is to estimate the total

$$T = \sum_{i=1}^{N} y_i.$$

In Royall's approach, the values $y_i, \ldots, y_N$ are presented as realizations of independent random variables, with

$$E(Y_i) = \beta x_i,$$

$$V(Y_i) = \sigma^2 v(x_i),$$

$$C(Y_i, Y_j) = 0, \qquad i \neq j,$$

where the operators E, V and C denote expectation, variance and covariance, respectively, with respect to the probability distribution ξ. The function $v(x_i)$ is assumed known with $v(x_i)>0$ for $x_i>0$; the constants β and σ^2 are unknown. The usual alternatives considered for $v(x_i)$ are $v(x_i) = 1$, $v(x_i) = x_i$ or $v(x_i) = x_i^2$.

Royall's basic approach is derived from recognition of the fact that after the sample is observed, the population total can be written as

$$T = \sum_{i \in s} y_i + \sum_{i \notin s} y_i,$$

where the first sum is known and the second must be estimated from the sample. Now any estimator, $\hat{T}$ of T, can be written as

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i,$$

where $\hat{y}_i$ is the implied predictor of y_i. Thus, under the regression model, any estimator $\hat{T}$ of T can be uniquely expressed in the form

$$\hat{T} = \sum_{i \in s} y_i + \hat{\beta} \sum_{i \notin s} x_i \quad .$$

Royall showed that under this model, the best linear unbiased estimator of T is given by the above expression with $\hat{\beta}$ the weighted least squares estimator of β,

$$\hat{\beta} = \left(\sum_{i \in s} \frac{y_i \, x_i}{v(x_i)} \right) / \left(\sum_{i \in s} \frac{x_i^2}{v(x_i)} \right).$$

It follows that in the three cases $v(x_i)=1$, $v(x_i) = x_i$ and $v(x_i) = x_i^2$, the best linear unbiased estimators for T are respectively

$$\hat{T}_0 = \sum_{i \in s} y_i + \left(\sum_{i \in s} x_i \, y_i \, / \sum_{i \in s} x_i^2 \right) \sum_{i \notin s} x_i,$$

$$\hat{T}_1 = \sum_{i \in s} y_i + \left(\sum_{i \in s} y_i \, / \sum_{i \in s} x_i \right) \sum_{i \notin s} x_i,$$

$$\hat{T}_2 = \sum_{i \in s} y_i + \left[(1/n) \sum_{i \in s} (y_i \, / x_i) \right] \sum_{i \notin s} x_i.$$

Furthermore these estimators are best for any sampling plan. Whatever sampling plan is used in selecting s, the error variance is

$$E(\hat{T} - T)^2 = \sigma^2 \left[\sum_{i \notin s} v(x_i) + \left(\sum_{i \notin s} x_i \right)^2 \frac{\sum_{i \in s} v(x_i)}{\left(\sum_{i \in s} x_i \right)^2} \right].$$

Royall showed theoretically and empirically that for producing an estimate close to the true population total under this model random sampling plans are often inferior to strategies which call for purposive (non-random) selection of samples. Namely, the optimal sampling design (minimum mean square error) is the one which entails selection of s* with certainty, where s* is the set of n units for which

$$\sum_{i \in s^*} x_i = \max_{s \in S} \sum_{i \in s} x_i,$$

where S is the totality of all possible samples.

3.0 Model Specification

In the weekly survey problem, for a given product in a given region, the known data, $(x_1,\ldots,x_N)$ will be the three month old reported values from the monthly census and the vector $(y_1,\ldots,y_N)$ will be the realization of the random vector of weekly data, $(Y_1,\ldots,Y_N)$. Since the weekly data are not now available, the sample selection and testing were done by letting the vector $(y_1,\ldots,y_N)$ be presented by the monthly data, with $(x_1,\ldots,x_N)$ the monthly data three months older.

The model for this application was selected from the three models listed above based on the monthly data from all refineries between January 1979 and March 1980. The three regression models were fit for each product at the U.S. level for each of 12 pairs of months. For each model and each product the residuals of the 12 regressions were plotted as a function of the size of the the company. For each product the size of the company was defined to be the average of that company's volume of that product over 12 months (April 1979 - March 1980). The mean square error for each model and for each product was also calculated using the residuals from the 12 regressions.

If the data fit a particular model, one would expect the spread of the residuals from that regression model to be independent of company size. By this criterion, the plots for all products showed that the third alternative, $v(x_i) = x_i^2$, was a poor fit. A visual examination of the plots for the other two models was inconclusive.

However, for each of the 17 products, the mean square error for the unweighted regression, $v(x_i) = 1$, was slightly smaller than the mean square error for the weighted regression, $v(x_i) = x_i$. Hence, the unweighted regression model was selected as the appropriate model.

The refineries operated by the major companies (154 of the total of 321 refineries) are prespecified as being in the sample. Hence, the refineries of the major companies form the nucleus of the sample. A statistical test was performed at the national level, to determine whether the same model parameters apply to both the major companies' refineries and the nonmajor refineries. This test was done via the regression model with a dummy variable added for the difference in slope between majors and nonmajors (Gujarati (1970)). The test of the null hypothesis, that the coefficient of the dummy variable was zero could not be rejected at the 5% level of significance for all products except production and stocks of kerosene jet fuel.*

For production and stocks of kerosene jet fuel, the major companies accounted for more than 96% of the total during this time period. Hence, we can be reassured that the error calculated as a percent of the total will be small, even though the error calculated as a percent of the total of the nonsampled companies may be larger than one would like.

4.0 Sample Selection

Two approaches for sample selection were considered. Both approaches began with the 154 refineries belonging to the major companies as the nucleus of the sample. In both cases, estimates for the total volume of each product in each district for each of 12 months (April 1979 through March 1980) were calculated based on the sample containing only the majors. These estimates were compared to the true totals and the relative error for each was calculated.

* Results of the test would have been the same at the .5% level for all products with the additional exception of production and stocks of kerosene.

On the basis of these results the largest nonmajor refineries were added to the sample until the errors in calculating monthly totals were suitably small.

In the first approach, the implied estimator $\hat{\beta}$ for each product in each district was estimated separately, resulting in 221 regressions per month. Refineries in a problem district were added to the sample based on their total volume of the products for which the sample of majors gave large errors. Companies were frequently added to the sample for small districts or for small products to obtain a census.

This approach has the advantage that different patterns of behavior in different districts can theoretically be accounted for. In the smaller districts for the smaller products, however, it has the disadvantage of estimating the implied estimator of β, with a small number of observations. For these cases, accuracy of the estimation is questionable even with all sampled company's reporting. The situation would be even worse with nonresponse.

The second approach is to estimate the regression parameter for all products at the national level. In this case refineries were also added to the sample to improve estimation for these products where the sample of majors gave large errors. However, in this case there were two philosphies for adding refineries. 1) If there were large errors for that product in many districts, refineries were added based on their volume of the "problem product" according to a total U.S. ranking. This procedure was intended to improve the estimate of β. 2) If there were large errors for a product in only one or two districts, refineries in the problem districts were added based on the volume of the product according to a ranking within the problem district. This procedure was intended to provide better sample coverage in that district.

This second approach should provide more reliable estimates for individual parameters than the first approach because of the larger number of observations. This procedure would be preferred as long as the difference between the "true" parameter values for a product in individual districts is not too great. The larger number of observations used for parameter estimation will also considerably reduce the impact of nonresponse.

For both alternatives, after companies were added to the sample, all totals were reestimated using the expanded samples and the errors were examined. If the desired accuracy was still not obtained, more companies were added to the sample as decribed above. This procedure was repeated until adequate accuracy at all levels of estimation was obtained.

The first approach resulted in a sample of the 154 major refineries plus 43 additional refineries; giving a sample of size 197. The second approach in a sample consisting of the 155 major refineries plus 27 additional refineries giving a sample of size 181. Since at the present time the weekly system is operating with a sample of 202 refineries, these results are encouraging.

5.0 Results

For a given sample and a given estimation procedure, estimates for each product and each district were calculated for each month from April 1979 through December 1980. Errors for each month were calculated as the difference between the estimated total and the true total. Relative errors for each month were calculated as the ratio of the error to the true total.

For each sample and each estimation procedure, the summary statistics described below were calculated over the 12 months from April 1979 through March 1980, the time period which was used for sample selection. For the sample drawn through Approach 2, the summary statistics were also calculated over the 9 months from April 1980 through December 1980.

The summary statistics and the acceptability criteria used for this study are described below:

a. Mean error (mean relative error) is defined to be the average of the errors (relative errors) over either the 12 month period from April 1979 through March 1980 or the 9 month period from April 1980 through December 1980.

   Mean error was somewhat arbitrarily defined to be acceptable if either the absolute value of the relative error was less than 3% or the absolute value of the mean error in thousand barrels (the publication units) was less than 10.

b. Root mean square error (root mean square relative error) is defined to be the square root of the mean of the squares of the errors (relative errors) over either the 12 month period or the 9 month period.

   Root mean square error was defined to be acceptable if either the root mean square relative error was less than 8% on the root mean square error was less than 80 thousand barrels.

c. Maximum error (maximum relative error) is defined to be the maximum of the absolute value of the error (relative error) observed over either the 12 month period or the 9 month period.

   The maximum error was defined to be acceptable if either the maximum relative error was less than 12% or the maximum error was less than 150 thousand barrels.

5.1 Comparisons Based on the 12 Month Period

The purpose of the comparison based on the twelve month period from April 1979 through March 1980 is twofold: 1) to determine whether the smaller sample selected through approach 2 is adequate; 2) to determine whether district based estimation or nationally based estimation yields more accurate results.

These two goals are addressed by computing the summary statistics described above for the 221 product-district estimates. Those which failed any of the three criteria introduced above for approach 1 (a larger sample, district based estimation) and for approach 2 (the smaller sample, nationally based estimation) are examined and explained. The number of failures of the criteria for approaches 1 and 2 are summarized in table 2.

There was only one product-district estimate which failed for both approaches. Production of kerosene in district 3C had a maximum relative error of 16.2% and a maximum error of 252 thousand barrels using approach 1. It had a maximum relative error of 14.3% and a maximum error of 221 thousand barrels using approach 2. In this case, the sample for approach 2 included one more company than the sample for approach 1. This was the only product-district to fail any of the 3 criteria for approach 1. All estimates for production of kerosene in this district failed because of irregularities in the patterns of production of kerosene. In January 1979, the base month for estimating April 1979, the nonsampled companies only produced 29 thousand barrels (approach 1) or 9 thousand barrels (approach 2) out of a total of 1193. In April 1979 the nonsampled companies produced 266 thousand barrels (approach 1) or 226 thousand barrels (approach 2) out of a total of 1548. The magnitude of this variation in the non-sampled companies would be impossible to predict.

The additional failures of the criteria for approach 2 are summarized and discussed below.

a. Stocks of distillate fuel oil in district 3A failed all three criteria. The mean relative error was - 7.8% , and the mean error was -90 thousand barrels. The root mean square relative error was 14.5%, and the root mean square error was 164 thousand barrels. The maximum relative error was 34.2%, and the maximum error was 396 thousand barrels.

   In district 3A one refinery reported zero stocks of distillate fuel oil for all but 5 months during the winter of 79-80. For these 5 months their stocks were about one fifth of the total in the district. During the two year period from January 1979 through December 1980, they had no inputs, no production and only one or two months with small stocks of other products. The sample drawn through approach 1, included this company. However, it is routinely small enough that it should not be included as a sampled company in the weekly system. During this same 5 month period one other small nonsampled company showed larger stocks of distillate fuel oil than usual, while for two major companies, stocks of distillate fuel oil fell abnormally low. The net result was that total stocks of fuel oil were not significantly different, but the percentage held by sampled and nonsampled companies changed radically. Since stocks are reported on a custody basis, it may be that the major companies arranged for the nonmajor companies to hold some of their stocks of distillate for that period.

b. Stocks of naphtha jet fuel in district 2B had a mean relative error of 3.2% and a mean error of 13 thousand barrels.

   In this district the samples drawn through both approaches were the same. The mean error associated with estimation at the national level exceeded the mean error criterion. The mean error associated with estimation at the district level did not. However, the root mean square error and the maximum error were smaller with estimation at the national level.

c. Production of residual fuel oil in district 2D had a maximum relative error of 14.8% and a maximum error of 153 thousand barrels.

   For the sample drawn in the second approach, the district based estimate for production of residual fuel oil was acceptable. Only the nationally based estimate violated the criterion.

In order to compare nationally based estimation with district based estimation the summary statistics described above were also calculated using district based estimation and the sample drawn in approach 2. A summary of the number of failures of the three criteria for the two estimation schemes is presented in Table 3. It appears that district based estimation fails the three criteria more frequently.

5.2 Comparisons Based on the 9 Month Period

Although the comparisons above are useful, they are not necessarily illustrative of expected precision since the samples were selected to minimize errors for the April 1979 - March 1980 time period. For this reason, the summary statistics for the nationally based estimates and for the district based estimates were calculated for the 221 product - districts using the sample drawn in approach 2 for a later time period

(April 1980 through December 1980). As in the previous section, comparisons are made by examining failures of the three criteria. These are summarized in table 3. Only one failure of national level estimation was critical. In district 5A for production of residual fuel oil the mean relative error was -3.7%, and the mean error was -583 thousand barrels. The magnitude of the mean error is unacceptable. District level estimation gave approximately the same results. This situation should be investigated.

In general, however, for the later time period nationally based estimation results in fewer violations of the three criteria.

5.3 Conclusions

In general, it appears that estimating the model parameter for a product at the national level is more reliable than computing a parameter for each district. In two cases for the 12 month data, the district level estimates were better. It may be that the optimal procedure would involve basing estimates on district data when there are a relatively large number of companies reporting, and on national level data when only a small number are reporting.

In the first approach to sample selection, companies were added to small districts or small products to obtain a census. When basing results on monthly data only, the accuracy for these districts or products is perfect. However, care must be taken to avoid adding to the weekly system very small companies or companies which have trouble reporting in a timely fashion on the monthly system. It is felt that these companies will contribute to errors in the weekly system by increasing the nonresponse rate. For these reasons, and because the observed accuracy was acceptable, it is felt that the smaller sample, derived through the second approach is preferable. For application in 1981 or 1982, however, the sample would have to be augmented to resolve the large mean error in the estimation of residual fuel oil in district 5A.

It appears that the estimation accuracy does not deteriorate too badly the year following sample selection. To maintain accuracy, operators of the system will have to be conscious of large changes in reporting patterns and of births and deaths in the system.

In conclusion, we believe that this exercise has demonstrated that the model-based approach will lead to a manageable sample size and acceptable estimates.

REFERENCES

Gujarati, D. (1970). Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Note. The American Statistician, 24, 50-52.

Kirkendall, N. J. and Kolmer, W. K. (1980). EIA Weekly Petroleum Data: Data Collection and Methods of Estimation. Monthly Energy Review. US Department of Energy, DOE/EIA-0035 (80/11).

Royall, R.M.(1970). On finite population sampling theory under certain linear regression models. Biometrika, 57,377-387.

Särndal, C. E. (1978). Design-based and model-based Inference in Survey Sampling. Scandinavian Journal of Statistics, 5, 27-52.

Table (1) REFINERY POPULATION SIZE BY REFINING DISTRICT AND TYPE

Refining District	Type		
	Major	Nonmajor	Total
1A	13	7	20
1B	5	5	10
2A	2	0	2
2B	22	15	37
2C	5	2	7
2D	17	8	25
3A	8	19	27
3B	19	20	39
3C	13	19	32
3D	5	19	24
3E	1	6	7
4A	14	17	31
5A	30	30	60
Total	154	167	321

Table (2) Number of Product-District Estimates Failing Criteria
(Approach 1 Versus Approach 2)

	CRITERIA FAILED*			
	Mean Error	Root Mean Square Error	Maximum Error	No Failures
Approach 1	0	0	1	220
Approach 2	2	1	3	217

* Product-district estimates may fail more than one criterion

Table (3) Number of Product-District Estimates Failing Criteria
(Nationally Based Versus District Based Estimates)

	CRITERIA FAILED*			
	Mean Error	Root Mean Square Error	Maximum Error	No Failures
Apr 79 - Mar 80:				
Nationally Based	2	1	3	217
District Based	2	4	11	209
Apr 80 - Dec 80:				
Nationally Based	7	1	1	214
District Based	14	3	4	204

* Product-district estimates may fail more than one criterion