An Iterative Cutoff Sampling Method Applied to an Annual Oil and Gas Reserves Report

Jason Worrall, Samson Adeshiyan
U.S. Energy Information Administration, Forrestal Bldg., Washington DC 20585

July 19, 2015

1 Introduction

This paper presents a procedure for selecting a cutoff sample of multiple data items using noncontiguous estimation and publication groups. The procedure is intended to select a sample of minimal size needed to obtain given target Relative Standard Errors (RSEs) for each publication item. It builds on work by Jim Knaub at the Energy Information Administration in model based sampling to present a generalizeable framework for cutoff sample selection. Knaub 2013 [1] presents a methodology in the case of the Classical Ratio Estimator (CRE) for estimating the approximate sample coverage needed to achieve a target RSE, but no general procedure has been described for optimizing sample size accross multiple data items or estimation and publication groups, nor for weighted least squares linear models aside from the CRE. This work attempts to do that.

The case study for this methodology is EIA's Form EIA-23, "Annual Survey of Domestic Oil and Gas Reserves". This survey collects annual production and year ending reserves from a sample of oil and gas well operators. Estimates of total reserves are then published in EIA's "Annual Domestic Oil and Gas Reserves Report" by state, subdivision, and reservoir type. The sampling frame is provided by "DrillingInfo" (DI), a subscription database of monthly oil and gas production by well, which has information on well charactaristics and current operator identification.

2 The Problem

Respondents to the EIA-23 are oil and gas well operators. They may operate many different wells in many different regions all accross the country. The EIA-23 estimates are based on the gamma super population model (equation 1):

$$y_i = \beta x_i + x_i^{\gamma}\epsilon_i \tag{1}$$

where y_i represents an operator's year ending reserves, x_i their annual production, and ϵ_i is a random disturbance with mean zero and variance σ^2. This model is estimated by weighted least squares with weights equal to $x_i^{-2\gamma}$. Oil and gas wells are not homogenous, and different regions and reservoir types have different charactaristics that effect the relationship between reported production and reserves. So equation 1 must be estimated separately for each region and reservoir type, yeilding distinct estimation groups. In addition, the Annual Reserves Report publishes reserves by state, subdivision, and reservoir type, yeilding distinct publication groups. These groups are noncontiguous, meaning estimation groups are not necessarily contained within a single publication group, nor are publication groups contained within a single estimation group. The situation is illustrated in figure PUT A FIGURE HERE.

As a result, the total reserves to be estimated for a given publication group p made up of estimation groups $e \in (1, E)$ is given by equation 2.

$$\hat{T_p} = \sum_{e=1}^{E} \left(\sum_{i=1}^{n} y_{i,e,p} + \sum_{i=n+1}^{N} \hat{\beta_e} x_{i,e,p} \right) \tag{2}$$

If there were only a single estimation group and publication group, [WHO SHOWS THIS?] shows that the RSE of the error in this total is

$$RSE = \frac{\sqrt{V(T - \hat{T})}}{\hat{T}}$$

$$V(T - \hat{T}) = \hat{\sigma}^2 \left(\sum_{i=n+1}^{N} x_i^{2\gamma} + \left(\sum_{i=n+1}^{N} x_i \right)^2 \left(\sum_{i=1}^{n} x_i^{2-2\gamma} \right)^{-1} \right) \tag{3}$$

with

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{x_i^{2\gamma}(n-1)}$$

With arbitrary estimation and publication groups, the variance of a particular publication group will be given by the sum of the variance of each piece, as we are assuming independence in ϵ_i accross observations, and so independence of $\hat{\beta}$ accross different estimation groups.

$$V(T_p - \hat{T_p}) = \sum_{e=1}^{E} \left(\hat{\sigma_e}^2 \left(\sum_{i=n+1}^{N} x_{i,e,p}^{2\gamma} + \left(\sum_{i=n+1}^{N} x_{i,e,p} \right)^2 \left(\sum_{i=1}^{n} x_{i,e}^{2-2\gamma} \right)^{-1} \right) \right) \tag{4}$$

With this formula in hand, the question becomes: what is the smallest sample that will achieve RSE targets? Given a target RSE, reference [1] shows in the case of $\gamma = 0.5$ and a single estimation group that the needed sample coverage to achieve a target RSE is a function of only three quantities: the anticipated value of of σ^2, the anticipated total y_i across all population units (T_y), and the total x_i across all population units (T_x).

One could simply apply that methodology to each estimation/publication (ep) group combination individually, using the RSE target for the publication group for each ep group, and then combine the samples from each ep peice to get the sample for the publication group. Unfortunately, this will yield suboptimal results, especially since a particular estimation group may make up a relatively small portion of the publiacation group, and so sampling to achieve such low RSEs in such a small area is not be required. In the next section we present a procedure for selecting a cutoff sample in this general case that attempts to avoid such inefficiencies.

3 Proposed Procedure

Examining equation 4, we may follow in the footsteps of reference [1] and consider the optimal sample S to be a function of the target RSE, R, the expected total dependent variable T_p, the anticipated σ_e^2, and the full regressor data, $\mathbf{X}$. In other words there exists a function, $f : (R, T_p, \sigma_e^2, \mathbf{X}) \rightarrow S$. However, this function will be completely intractable in most cases. So, we propose a methodology that iteratively adds to the sample for a given publication group, always adding the operator that will reduce the anticipated RSE by the most. Explicitely:

1. Obtain anticipated σ_e^2 and β_e for each estimation group using whatever means are available.
2. Consider a particular publication group p.
3. Add to the sample the largest unit from each estimation/publication group combination $(e_1, p)...(e_E, p)$ that makes up p.
4. Calculate what the anticipated RSE on T_p would be if the largest unit from (e_1, p) were added to the sample. Repeat for $(e_2, p)...(e_E, p)$.
5. Add to the sample the unit that lowered the RSE by the most.
6. Repeat 3-5 until the RSE has been brought below the cutoff.
7. Repeat 2-6 for every publication group.

This procedure attempts to minimize ineffiencies by accounting for all estimation group within a publication group while building the sample, but does not make considerations accross publication groups. That is the subject of further work.

The reason for step (3) is to avoid having *ep* groups that are entirely unrepresented, and to enable the use of equation 4. If there is an excessively large number of *ep* groups, one may consider only adding the largest operator from each estimation group, rather than from each *ep* group.

4 Results

Consider the case of sampling Colorado gas reserves. The EIA-23 divides Colorado into eight geological provinces/estimation gorups, labeled as $A - H$ here. Estimates of β and σ^2 based on historical data are presented in table 1, where T_{COx} is the total Colorado x for the estimation group.

Estimation Group	T_{COx}	$\hat{\sigma}^2$	$\hat{\beta}$
A	1,225	68.2	11.08
B	832,684	55.8	9.0
C	365,067	115.6	13.0
D	6,129	25.3	7.3
E	27,906	74.9	13.3
F	14,688	99.0	10.2
G	9,347	100.0	13.1
H	764,569	90.5	13.2

Table 1: Estimation Groups

Looking at the numbers in table 1, groups B and H make up the majority of the production. It stands to reason that more respondents from those groups will be necessary. In addtion, group H has both higher σ^2 and β estimates, which will increase the needed sample size. Applying the procedure outlined in the last section, a sample of 14 respondents is necessary to achieve an expected RSE of 5% in Colorado. Group B has 5 respondents sampled, group C has 2, and group H has 4. All the remaining groups have only a single operator sampled (remember that the largest operator from each group is always sampled). Two operators sampled had production in two regions.

If each *ep* group is targeted at the 5% RSE level seperately, adding respondents until the RSE for each falls below 5%, the sample size calculated is 51. This is a dramatic difference, understood as a result of the five smallest estimation groups making up only 3% of the Colorado total. If the analysis is restricted to the two largest estimation groups (B, H), seperately sampling from the two groups yields a sample size of 12 while the proposed procedure yields a sample size of only 9.

In practice, statisticians selecting cutoff samples will often use a cutoff that has been shown to give satisfactory RSEs, but the cutoff is often *not determined by the RSE.* It is specified exogenously or determined based on some other criteria. The proposed approach allows statisticians to ground cutoff sample selection explicitely on the expected variance of the final estimate for arbitrary estimation/publication groupings.

Bibliography

[1] James R. Knaub. Projected variance for the model-based classical ratio estimator: Estimating sample size requirements. In *Joint Statistical Meeting 2013*. American Statistical Association, 2013.