Effects of Measurement Errors and Heteroscedasticity in Estimation of Constrained Parameters

Melinda Holt and Cecil Hallum
Sam Houston State University, Box 2206, Huntsville, TX 77341-2206

Abstract

A generalized, weighted multiple regression model with no intercept and equality-constrained coefficients is investigated and modified to arrive at best linear estimators (BLE) in direct support of subsurface mineral composition studies in oil exploration. Subsurface analyses have historically required use of a model relating a spectral response vector, *Y*, to a matrix of standard yields, *X*, for elemental standards such as iron, silicon, quartz, calcium, etc. A primary motivation for the resulting model approach is the need to more appropriately account for heteroscedasticity within the independent variables, which is prevalent in petroleum environments. Monte Carlo simulations are utilized to explore the performance of these estimators in the presence of measurement error and heteroscedasticity.

Key Words: Parameter constraints, Multivariate linear measurement error, Coverage.

1. Motivation

Since their development, gamma ray spectrometry tools have been used (Hertzog (1980)) to conduct elemental analyses with a variety of applications (Pasternack, B. S. (1962)). One goal is to estimate the percentage of the gamma ray signal contributed by each of k elements and, thus, to estimate parameters subject to the constraint that they sum to 1. To do so in oil exploration, the tool is lowered into a borehole and gamma ray readings are measured. The observed total counts, *Y*, are then regressed against laboratory standards, *X*, for each of the contributing elements. While laboratory standard measurements carefully reflect borehole conditions, the actual element readings in the field are unobservable and subject to both random and measurement error. Several authors have published studies that employ ordinary weighted least squares (WLS) regression for this and similar problems. See, for example, Pasternack (1962), Hertzog (1980), Roscoe, Grau and Wraight (1986), Galford, et al. (1988), Gartner and Jacobson (1990), and Wensheng, Wei and Li (2014). Although Chhikara and Hallum (unpublished manuscript, 1986) developed constrained ordinary least squares (COLS) estimators, no one appears to formally build inherent parameter constraints into the WLS model but rather attempt to normalize the data. Neither do they consider the impact of measurement error on the resulting estimators. Wu, Zhang and Luo (2014) offer a nonlinear optimization method to address the constraint in this context and compare the results to existing WLS methods by comparing the correlation of estimates to parameter values, but do not consider impacts of potential measurement error. We provide the best linear estimators (BLE) for COLS and constrained weighted ordinary least squares (CWLS). We also provide a performance analysis that considers the effects on interval coverage of both inevitable measurement errors and statistical errors that appear in the weighting process. Although the derivations herein are motivated by the collection of gamma ray data in industry and associated laboratory standards, both measurements are

highly proprietary. Thus, the data used to drive the performance analysis are purely hypothetical. Efforts to obtain industry values similar to those in Wu, Zhang and Luo (2014) are ongoing.

2. Equation Development

2.1 Constrained Parameter Model

First we consider the following model:

$$Y = Xp + \varepsilon,$$

where Y is the $n\times1$ vector of gamma ray counts, X is the $n\times k$ matrix of laboratory standards, p is the $k\times1$ vector of element proportions, ε is the $n\times1$ error vector and, in the unweighted case, $\varepsilon_i \sim iidN(0, \sigma^2)$ for each i. Here p is also subject to the inherent equality constraint

$$\sum_j^k p_j = 1 \text{ or } Rp = t,$$

where R is the $1\times k$ vector, $R = [1,1,1,\ldots,1]$, and $t = 1$.

The COLS estimators take the form of

$$\hat{p} = R^+t + X_0^+Y_0 \tag{1}$$

where $^+$ is the Moore-Penrose generalized inverse (Boullion and Odell (1971)), $X_0 = X(I - R^+R)$ and $Y_0 = Y - R^+t$.

The associated intervals become

$$\hat{p} \pm z_{\propto/2}V^{1/2},$$

where $V = \text{vecdiag}([X_0'(\hat{\sigma}^2 I)^{-1} X_0]^+)$, $\hat{\sigma}^2$ is the mean squared error, and the vecdiag function simply selects the diagonal elements from its argument to form a $k\times1$ vector of variances for the estimated proportions. Moreover, $V^{1/2}$ is the element-wise square roots of the coordinates in V.

2.2 Constrained Generalized Linear Model with Relative Homoscedasticity

Next we consider the situation in which the coefficient of variation remains constant, but the standard deviations do not. This is sometimes referred to as proportional measurement errors or *relative homoscedasticity*. Such errors have been considered in clinical chemistry (Linnet, 1993), but are not subject to parameter constraints.

Here the model continues to take the form

$$Y = X^*p + \varepsilon,$$

where Y is the $n\times1$ vector of gamma ray counts, p is the $k\times1$ vector of element proportions, ε is the $n\times1$ error vector and $\varepsilon_i \sim iidN(0, \sigma^2)$. Again p is subject to the inherent constraint that $\sum_j^k p_j = 1$ or $Rp = t$. Now, however, X^* is an $n\times k$ matrix such that X_j is a $n\times1$ random

column vector with mean vector $\bar{X}_j$ and covariance $Cov(X_j)$, for $j = 1, \ldots, k$. $Cov(X_j)$ meets an assumption of relative homoscedasticity, meaning that the coefficient of variation is constant or $CV_j = s_{ij}/\bar{X}_{ij}$ is constant for each row i in column j. In addition, we assume independence of X_j.

Parameter estimation here requires significant prior information. Pilot studies may produce $\hat{p}_0$, using the constrained estimators in (1). Expert knowledge of *CV* yields

$$Cov(X_j) = (CV_j)^2 D_j,$$

where D_j is an $n \times n$ matrix with the squared means on the diagonal. The resulting CWLS estimators are

$$\hat{p} = R^+ t + [(X_0^*)'(V^*)^{-1}X_0^*]^+ (X_0^*)'(V^*)^{-1}Y_0, \quad (2)$$

where

$$S = \hat{p}_{01}^2 Cov(X_1) + \hat{p}_{02}^2 Cov(X_2) + \cdots + \hat{p}_{0k}^2 Cov(X_k) + \hat{\sigma}^2 I,$$

$$V^* = \text{vecdiag}[(X_0^*)'S^{-1}(X_0^*)]^+,$$

$$X_0 = X(I - R^+R) \text{ and } Y_0 = Y - R^+\text{t}.$$

The associated intervals then become $\hat{p} \pm z_{\propto/2}(V^*)^{1/2}$.

3. Performance Analysis

3.1 The Data

An example of the data collected, without scale information is presented in Figure 1.

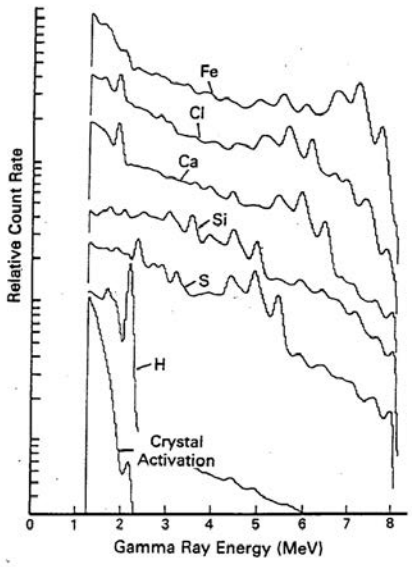

Figure 1: Gamma Ray Data (Herzog, 2016)

A defining characteristic of each element is the considerable range in the relative count rate across the spectrum. This is one motivation for the incorporating a relative homoscedasticity. Specifically, if one knows a reasonable *CV* that is applicable to an element in the field, then a reasonable variance can be derived from the lab-measured element values.

The data in Figure 2 is motivated by natural gamma ray spectra, which seek to estimate potassium, thorium and uranium in shale. They are, however, purely hypothetical. Here k = 3 elements, using n = 11 observations (or channels). Let p_1 = 0.10, p_2 = 0.30 and p_3 = 0.60. Data with error in X, both with and without relative homoscedasticity are considered below.

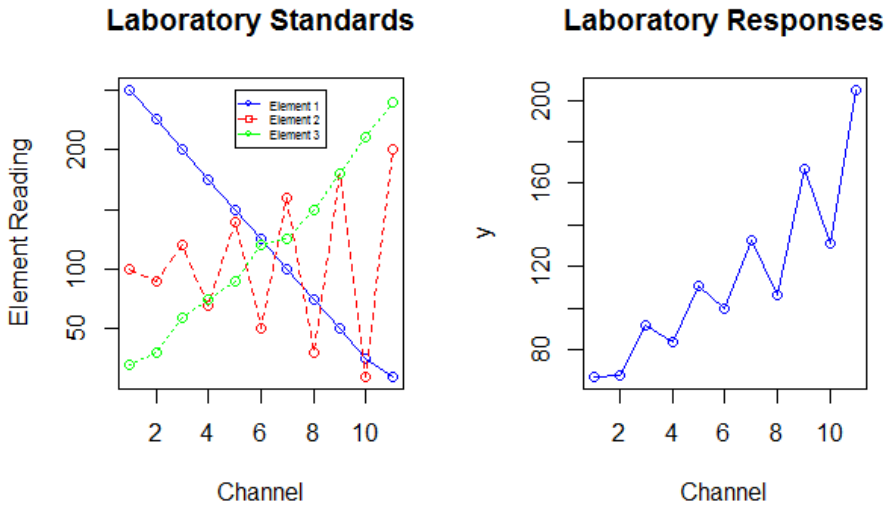

Figure 2: Example Data

3.2 Coverage in the Presence of Errors in X

To consider the impact of measurement error on the estimators in (1), we generated m = 3000 Monte Carlo samples, each containing n = 11 channels. The values presented in Figure 2 were taken as n×1 mean vectors $\bar{X}_j$, or laboratory standard vectors, for each of the three elements, so that $X_j \sim N(\bar{X}_j, \sigma_X^2)$, j = 1 – 3. The model error was simulated using $\varepsilon_i \sim iidN(0, \sigma^2)$ and 95% interval estimates calculated. The results appear in Figure 3 as a function of $\lambda = \sigma_X/\sigma$. Here σ and σ_X take on values of 3, 5, 6, 9, 10, 12, 15, 20, and 25. Thus, λ takes on the values 1/4, 1/3, 1/2, 1, 2, 3, and 4. This allows consideration of circumstances where measurement error is considerably smaller than model error and vice versa. Results are presented only for estimates of p_1, as the performance of the other two parameters is virtually identical.

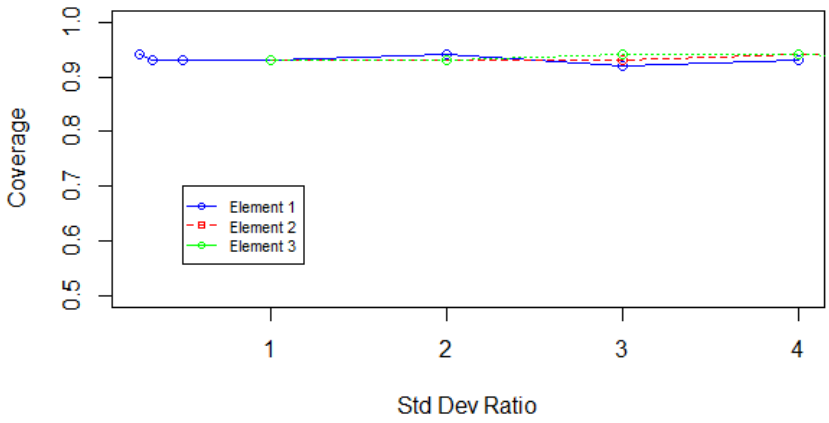

Figure 3: Coverage as a Function of Error Ratios for p_1, p_2 and p_3.

Figure 3 presents coverage for the interval estimates for each of the three elements. Despite the errors in *X*, the constrained estimators performed consistently at or near 93%. That is, the 95% interval estimates contain the true elemental parameter value approximately 93% of the time. This is true for each of the $k = 3$ parameters.

Table 1 provides the average widths for each of the parameters. Clearly, as expected, increases in either measurement error or model error variance increase interval width.

σ	σ_X	λ	p_1	p_2	p_3
3	3	1	0.042	0.073	0.053
3	6	1/2	0.074	0.128	0.093
3	9	1/3	0.109	0.188	0.136
3	12	1/4	0.143	0.248	0.179
6	3	2	0.059	0.103	0.074
9	3	3	0.080	0.138	0.100
12	3	4	0.102	0.177	0.128
5	5	1	0.071	0.123	0.089
5	10	1/2	0.123	0.213	0.154
5	15	1/3	0.181	0.312	0.226
5	20	1/4	0.236	0.409	0.296
10	5	2	0.099	0.172	0.124
15	5	3	0.133	0.230	0.166
20	5	4	0.170	0.294	0.213

Table 1: Interval widths

3.3 Coverage in the Presence of Relative Homoscedasticity in X

Figures 4 and 5 compare the performance of the estimators in (1) and (2) as a function of $CV = CV_j$, for j = 1 – 3. Here independent variables are measured with error such that each $X_j \sim N(\mu_j, \sigma_{ij}^2)$, $j = 1$ - 3, and σ_{ij}^2 meets an assumption of relative homoscedasticity. We simulate five possible scenarios:

1) Relative homoscedasticity exists but is not modeled.
2) Parameter values necessary to the calculation of S, p_j, CV_j, $\bar{X}_j$, σ^2 are known. That is, V^* is known without error.
3) A simulated pilot study produces $\hat{p}$ and $\hat{\sigma}^2$ but CV_j is known.
4) A simulated pilot study produces $\hat{p}$ and $\hat{\sigma}^2$ and CV_j is misidentified as 90% of its true value.
5) A simulated pilot study produces $\hat{p}$ and $\hat{\sigma}^2$ and CV_j is misidentified as 110% of its true value.

We again calculate 95% intervals with m = 3000, n = 11, and k =3. Figure 4 shows simulated interval coverage for each of the five scenarios when σ = 10. As expected, the estimates produced by (1), which do not incorporate relative homoscedasticity, lose coverage as *CV* increases. Again not surprisingly, the estimates produced by (2) maintain coverage near 95% when all necessary values are known perfectly. If each parameter in V is known, the weighted constrained estimators in (2) improve coverage between 4 and 13% over the unweighted ones in (1). While coverage is improved by the use of (2), widths changed very little between the two scenarios and are not presented herein.

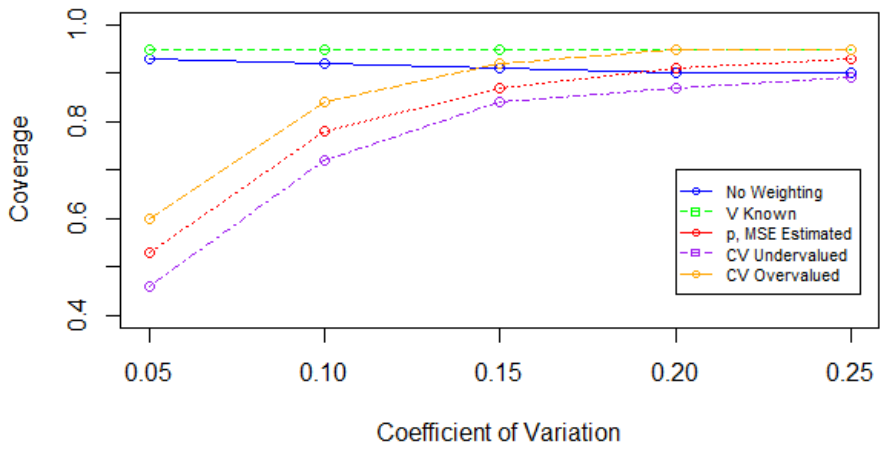

Figure 4: Coverage as a Function of *CV* (σ = 10)

Figure 5 shows analogous results for σ = 5. In this case, a lower model error variance in the model produces more accurate estimates for use in the weighted least squares. Thus, method (2) outperforms ordinary least squares for most *CV* values. This is true even when values necessary to the weighting are estimated through pilot studies.

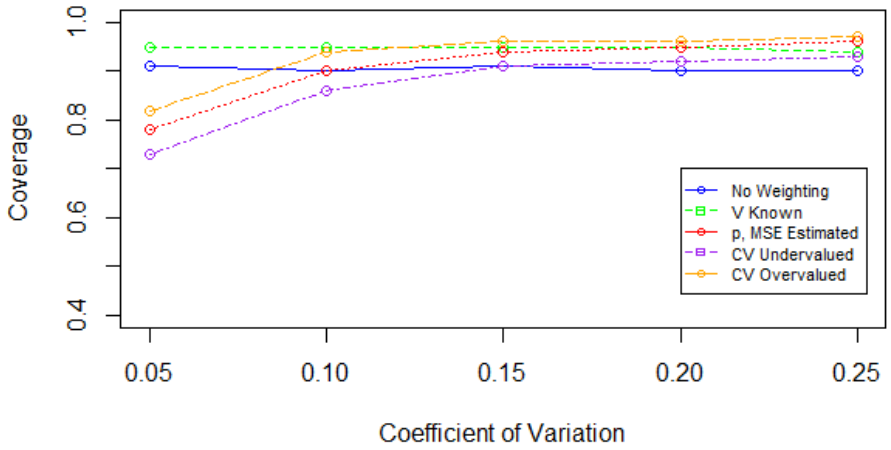

Figure 5: Coverage as a Function of *CV* ($\sigma = 5$)

4. Conclusions and Future Research

The COLS estimators in (1), that assume constant variance, are not developed under a formal errors-in-variables model but maintain coverage of approximately 92% to 94%. This is true despite the error introduced by substituting laboratory standards, *X*, for the unobservable field values. As expected, when variances are derivable from the *CV*s under the assumption of relative homoscedasticity using accurate prior information, the CWLS estimator in (2) outperforms that of equation (1). This simulation indicates that, even when all the values necessary to the weights in CWLS are estimated, the CWLS outperforms constrained OLS whenever *CV* is fairly large relative to σ. The results to date appear to warrant further investigation of the CWLS method.

Because the CWLS estimators rely heavily on *a priori* knowledge about *p*, σ and *CV*, it is worthwhile to develop and investigate the properties of analogous Bayesian estimators. Likewise, the authors plan to develop constrained the estimators under an errors-in-variables model for comparison with the COLS and CWLS models investigated in this paper.

Lastly, as mentioned in a number of the papers referenced herein, the effect of model misspecification is also of special interest. If the practitioner has either an error of omission, ignoring an element that should be included in the model, or of commission, including an element that should be excluded, the estimates of the element proportions can be impacted in a pronounced way.

Acknowledgements

The authors would like to thank Dr. Russel Hertzog for introducing us to this problem. Dr. Hertzog, a pioneer in gamma ray spectrometer development, helps to ensure that we pursue statistical questions relevant to industry.

References

Boullion, T. L. and Odell, P. L. (1971) *Generalized Inverse Matrices*, Wiley, New York.

Galford, J. E., Hertzog, R. C. Flaum, C. and Galindo, G. (1988) "Improving Pulsed Neutron Gamma Ray Spectroscopy Elemental Weight Percent Estimates through Automatic Dimensioning of the Spectral Fitting Process," In: Proceedings of the Society of Petroleum Engineering Annual Technical Conference and Exhibition, Houston, TX, 2 – 5 October, SPE 18151.

Gartner, M. L. and Jacobson, L. A. (1990) "The Dependence of Elemental Yield Variance on Detector Type through Mathematical Modeling," IEEE Transaction on Nuclear Science **37** (2), pp. 931 – 935.

Hertzog, R. C. (1980) "Laboratory and Field Evaluation of an Inelastic Neutron Scattering and Capture Gamma Ray Spectrometry Tool." Society of Petroleum Engineers Journal **20**, pp. 327 – 340.

Hertzog, R. C. (2016) "Neutron Capture Gamma Ray Spectroscopy." Invited Lecture, Presented to SHSU, May 12, 2016.

Linnet, K. (1993) "Evaluation of Regression Procedures for Methods Comparison Studies," Clinical Chemistry **39**(3), pp. 424 – 432.

Pasternack, B. S. (1962) "Linear Estimation and the Analysis of Gamma Ray Pulse-Height Spectra," Technometrics **4** (4), pp. 565 – 571.

Roscoe, B. A., Grau, J. A., Wraight, P. D. (1987) "Statistical Precision of Neutron-Induced Gamma Ray Sectroscopy Measurements," The Log Analyst 28(6), pp. 538 – 545.

Wenshung, W., Wei, N., and Li, L. (2014) "Quantitative Analysis of Neutron-Capture Gamma-Ray Energy Spectra using Direct Demodulation," Geophysics **79** (2), D91 – D96.

Wu, W., Zhang, L. Luo, L. (2014) "Quantitative Analysis of Neutron Captured γ Spectra Using a Sequential Quadratic Programming Method," Journal of Petroleum Science and Engineering **124**, pp. 1 – 6.