Incorporating User Input Into Optimal Constraining Procedures for Survey Estimates

Matthew Williams[1] and Emily Berg[2]

We examine the incorporation of analyst input into the constrained estimation process. In the calibration literature, there are numerous examples of estimators with "optimal" properties. We show that many of these can be derived from first principles. Furthermore, we provide mechanisms for injecting user input to create user-constrained optimal estimates. We include derivations for common deviance measures with linear and nonlinear constraints and we demonstrate these methods on a contingency table and a simulated survey data set. R code and examples are available at https://github.com/mwilli/Constrained-estimation.git.

Key words: Calibration; general deviance measures; nonlinear constraints; raking; user feedback.

1. Introduction

Constrained estimation has diverse applications in survey estimation. In the presence of auxiliary information, calibration of survey weights can improve the efficiency of a design consistent estimator. Deville and Särndal (1992) define calibrated weights as the weights that minimize a deviance function subject to the restriction that the weighted sum of a vector of auxiliary variables is equal to a known population total. They suggest a family of deviance functions and demonstrate that the resulting calibration estimators are asymptotically equivalent to a generalized regression estimator, a particular type of calibration estimator that arises from a quadratic deviance function. Chen and Sitter (1999) formulate the calibration problem using an empirical likelihood. Calibration can also be used to reduce a bias due to undercoverage of the sampling frame or nonresponse (for example, see Kott 2006, Chang and Kott 2008, and D'Arrigo and Skinner 2010). In a seminal paper, Deming and Stephan (1940) use iterative proportional fitting to enforce a restriction that the estimated marginal totals of a two-way table agree with census margins.

Whether the purpose of the calibration is to improve the efficiency of a design-unbiased estimator or reduce a bias due to nonsampling errors, care is often needed to avoid negative or extreme weights. Deville and Särndal (1992), Chen et al. (2002), and Singh and Mohl (1996) discuss methods for imposing range restrictions on calibrated weights.

These ensure that each sampled unit represents a reasonable positive number of units in the population.

Another application of constrained estimation is benchmarking of small area estimates to ensure that aggregated model-based estimates agree with a direct estimator or a previously published statistic for a larger region. Wang et al. (2008) review benchmarking methods in the context of a linear mixed model. They define a class of benchmarked estimators by minimizing a quadratic form subject to the benchmarking restriction. Nandram and Sayit (2011) incorporate linear constraints for small area probabilities using hierarchical Bayes and the standard beta-binomial model. In related work with shrinkage estimators, Ghosh (1992) imposes constraints on the mean and variance of Bayes estimates for a quadratic loss function. While variance constraints are quadratic (nonlinear), the use of a quadratic loss function leads to a closed form solution.

Many of the applications of constrained estimation discussed above apply linear constraints (see Särndal 2007; Estevao et al. 1995, who mention ratios of totals) to a set of initial estimates or initial weights by solving a constrained optimization problem. While the methods serve different purposes and have distinct interpretations, the functional forms are similar and derivations can be based on fundamental mathematical principles (such as the method of Lagrange multipliers). Because of the similarities between methods, constrained estimation in the survey world can seem like a tangle of overlapping terms and concepts. One of the objectives of this article is to clarify some of these associated concepts.

What is missing in the literature is a framework to create an interface between a user and the automated constraining procedure. Such a framework is essential for a statistical agency which is tasked with establishing estimates that are timely and accurate with the expectation of being compatible with subject or commodity knowledge and administrative data with partial coverage. Incorporating constraints into such a process must go beyond default settings and a choice of deviance measures. In addition to clarifying concepts, the purpose of this work is to establish such a framework.

1.1. Motivating Example

For statistical agencies, data often occur in triplets of *numerator* (**n**), *denominator* (**d**), and the *ratio* (**r**) of the two. Suppose we have a set of such triplets which must agree in aggregation with known targets (Table 1). Most methods in the literature would use linear constraints on the totals for **n** and **d**. But if **r** represents an agricultural rate of yield, which is production (**n**) per harvested area (**d**), then biological and industry knowledge would suggest adjusting the ratio directly (using nonlinear constraints) rather than the total production. The choice of which two of the three items in each triplet to adjust will often give distinct solutions. Figure 1 compares the relative adjustments made to each initial estimate when applying equivalent methods for constrained estimation to **n** and **d** versus **d** and **r**. The linear approach applies a constant proportional adjustment (decreasing for **n** and increasing for **d**). The nonlinear approach decreases **r** and increases **d**, but not at the same rates across all rows.

Constrained estimation provides a way for an analyst to incorporate external knowledge of the process that generated the basic estimators (either the direct survey estimators or

Table 1. Simulated Survey Data (rounded). Targets increase (light) and decrease (dark)

	Num (**n**)	Den (**d**)	Ratio (**r**)
1	2,586.20	56.55	45.73
2	30,491.31	913.17	33.39
3	4,141.68	78.83	52.54
4	1,975.41	68.59	28.80
5	18,827.87	362.00	52.01
6	6,280.19	137.20	45.77
7	8,597.05	182.03	47.23
8	4,995.37	242.78	20.58
9	7,402.01	216.61	34.17
10	1,168.46	52.52	22.25
11	5,455.36	243.30	22.42
12	1,778.24	60.79	29.25
13	3,208.09	195.24	16.43
14	2,249.00	56.44	39.85
15	2,215.65	72.80	30.44
16	14,297.99	454.96	31.43
17	3,948.72	190.49	20.73
18	1,653.01	77.39	21.36
19	2,545.01	86.12	29.55
20	2,749.02	72.91	37.70
Total	126,565.63	3,820.71	33.13
Target	120,237.35	3,935.33	30.55

estimators based on a subsequent model). For instance, contributions of large operators in establishment surveys, sizes of nonresponse and bias adjustments, administrative records, historical data, and qualitative information about the data-generating mechanism can be difficult to integrate into the basic estimation procedure, but might factor into an analyst's decision to set some values and reweight the adjustments on others. The analyst would then need a procedure to enforce these additional "user" constraints. For example, we can use analyst knowledge to fix entire rows in Table 1 *and* fix individual ratios r_i, *and*
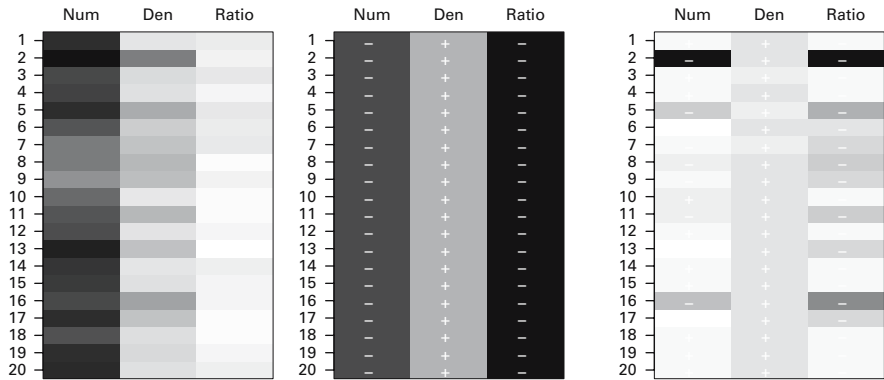

Fig. 1. Heat Map for Default Constraint of Triplets: (left to right) Log(Data), % Change (Linear), % Change (Nonlinear). White to black increases counts or size of change. Signs (−/+) show direction of change
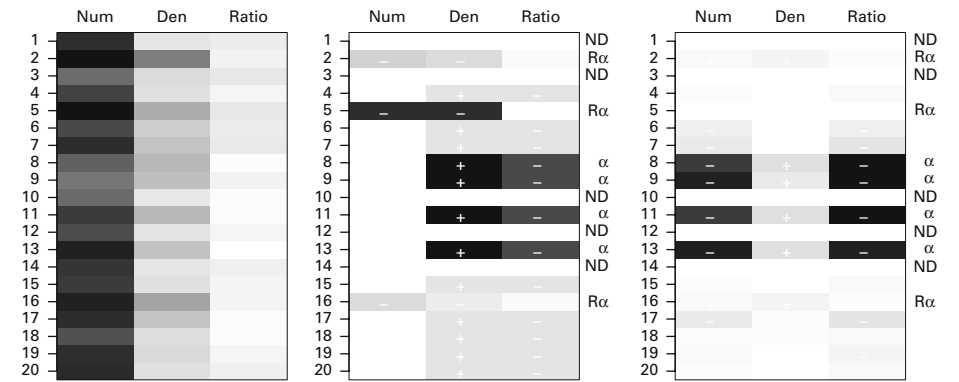
Fig. 2. *Heat Map for User Constraint of Triplets: (left to right) Log(Data), % Change (Linear), % Change (Nonlinear). White to black increases counts or size of change. Signs (−/+) show direction of change. Num and Den fixed (ND), Ratio set (R), rows reweighted (α)*

reweight to redistribute the amount of change absorbed by some rows. In heat map representation (Figure 2), these adjustments take the form of white cells (no change) and increases in intensity (darker up-weighted cells).

In the next section, we review the relationship between constraints and deviance measures. We introduce the concept of user interaction with an optimal procedure and explore several examples that might occur. Section 3 contains the details for a Newton-type method to generate solutions. In Section 4, we revisit the data set from Deming and Stephan (1940), applying our framework to incorporate user interaction. In Section 5 we elaborate on the example of linear and nonlinear constraints for triplets described in Subsection 1.1. Finally, we conclude in Section 6 with a summary and implications for further research. R code for methods and examples is available at https://github.com/mwilli/Constrained-estimation.

2.   Constrained Estimation

We consider the vector of observations (or unrestricted estimates) $\mathbf{y}$ of length n. We may wish to impose $k < n$ linear constraints $\mathbf{Ax} = \mathbf{q}$, where $\mathbf{x}$ is a constrained version of $\mathbf{y}$. Linear constraints take the form of weighted sums $\mathbf{a}_i\mathbf{x} = \sum_j a_{ij}x_j = q_i$ for $i \in 1, \ldots, k$ where $\mathbf{a}_i$ is the *ith* row of the $k \times n$ coefficient matrix $\mathbf{A}$. We restrict $\mathbf{A}$ to have full row rank k. Otherwise at least one $\mathbf{a}_i$ leads to a redundant constraint or creates a conflicting constraint. Consider an example in which constraints are imposed on all marginal totals of a two-way table with R rows and C columns. Because both row and column margins sum to the total for the table, a coefficient matrix $\mathbf{A}$ containing $R + C$ rows, one for each column and row sum, will contain one redundant row. This creates a deficient row rank for $\mathbf{A}$ of $R + C - 1$. A row associated with one of the row or column sums can be removed to produce a coefficient matrix with $R + C - 1$ rows and thus full row rank. (See Section 4 for further discussion of restrictions on the marginal totals of a two-way table).

We also consider $k < n$ *nonlinear* constraints $g(\mathbf{x}) = \mathbf{q}$. While the general class of nonlinear functions (all functions which are *not* necessarily linear) is extremely broad, we limit consideration to those that are well defined and have $n \times k$ continuous derivatives

$\mathbf{D}_g(\mathbf{x})$. In practice we consider polynomial, rational, and transcendental functions and compositions of them. These are generally wellknown and wellbehaved nonlinear functions. For example, the ratio of numerator and denominator from our motivating example (Subsection 1.1) is simple and wellbehaved when the denominator is nonzero. The variance constraint imposed by Ghosh (1992) for shrinkage estimates is a basic quadratic function used to counteract the overshrinking which occurs commonly in applications such as small area estimation.

We can no longer appeal to matrix rank to ensure that we do not have any conflicting constraints. However, it is clear that equations such as $x_1 + x_2 = q_1$, $x_1 - x_2 = q_2$, and $x_1/x_2 = q_3$ produce a conflict. Many methods for solving nonlinear systems of equations use linearization techniques involving derivatives (see Section 3). For these methods to find solutions, further restrictions may be placed on $\mathbf{D}_g(\mathbf{x})$. For our purposes we will assume $\mathbf{D}_g(\mathbf{x})$ has full column rank k for each value of $\mathbf{x}$.

2.1. Deviance Measures

Since $n > k$, the constraints by themselves do not imply a unique solution $\mathbf{x}$, but instead a family of solutions. A reasonable criteria to select a member of this family is to choose the $\mathbf{x}$ "closest" to $\mathbf{y}$. This concept of closeness implies minimizing a scalar deviance between $\mathbf{x}$ and $\mathbf{y}$. We will generally restrict these deviances to be rather simple and interpretable. From the calibration literature (for example, see Deville and Särndal 1992), there are several deviance functions used. We highlight the three most popular (for example, see D'Arrigo and Skinner 2010): the quadratic deviance $\chi^2(\mathbf{x}|\mathbf{y})$, the Poisson deviance $l(\mathbf{x}|\mathbf{y})$, and the discrimination information $D(\mathbf{x}|\mathbf{y})$. Each of these measures falls within the framework developed. Practitioners may use their current preferred deviance measure and still take advantage of the results and ideas presented here. Alternatively, one can change the deviance measure while still maintaining the other structures described below, such as the weighting matrix and the form of the constraints.

We express these deviances in matrix formulation and provide a weighting structure ($\mathbf{W}$) which allows for user input from an analyst or another model (see Subsection 2.3). The matrix formulations for some of the deviance measures may seem unnecessary, but the key insights come from the matrix formulation of the *constraints*. Expressing both in terms of matrix operations makes them more directly compatible (Subsection 2.2). We assume the base $\mathbf{W}$ is symmetric and invertible (although often the case, $\mathbf{W}$ need not be positive definite). We define $\langle \mathbf{v} \rangle$ as a square diagonal matrix with vector v on the diagonal and 0s elsewhere. We use the notation $[\cdot]$ to denote elementwise operations in two ways: First we use $[\mathbf{ab} + \mathbf{c}]$ for vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ of the same dimension to produce a vector with the *ith* element equal to $a_i b_i + c_i$. Second we denote $f[\mathbf{v}]$ as yielding a vector with *ith* element equal to $f(v_i)$. In other words, $f[\mathbf{v}]$ applies the scalar function $f(\cdot)$ elementwise to each v_i.

The Quadratic Deviance

$$\chi^2(\mathbf{x}|\mathbf{y}) = (\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$$

Examples include the Pearson chi-squared distance ($\mathbf{W} = \langle \mathbf{y} \rangle^{-1}$) and the Least Squares distance ($\mathbf{W} = Var(\mathbf{y})^{-1}$). Use is often motivated by a regression-based approach (Fuller 2002).

The Poisson Deviance

$$l(\mathbf{x}|\mathbf{y}) = \mathbf{1}'\mathbf{W}_d\left[\mathbf{y}\log\left[\frac{\mathbf{y}}{\mathbf{x}}\right] - \mathbf{y} + \mathbf{x}\right]$$

where $\mathbf{W}_d$ is a diagonal matrix of full rank. Motivation comes from the deviance measure of a log-linear model comparing a restricted model of means to the saturated model (Agresti 2002). In this case, $\mathbf{y}$ is the data (or saturated model) and $\mathbf{x}$ is the restricted model for $\mathbf{y}$. We will show in Subsection 2.2 that the Poisson deviance leads to the *pseudo-empirical maximum likelihood estimator* of Chen and Sitter (1999).

The Discrimination Information

$$D(\mathbf{x}|\mathbf{y}) = l(\mathbf{y}|\mathbf{x}) = \mathbf{1}'\mathbf{W}_d\left[\mathbf{x}\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right] - \mathbf{x} + \mathbf{y}\right]$$

The name seems to come from the application of the principal of minimum discriminability to cell probabilities (see Ireland and Kullback 1968, who attribute this to Good and Kullback). So-called *raking* methods such as iterative proportional fitting (IPF; Deming and Stephan 1940) are readily available to minimize this deviance for specific settings.

Of the three, $\chi^2(\mathbf{x}|\mathbf{y})$ is the simplest to implement and will often lead to closed-form solutions (Subsection 3.1). However, when $\mathbf{y}$ are positive survey weights, some of the resulting $\mathbf{x}$ may be negative. $D(\mathbf{x}|\mathbf{y})$ and $l(\mathbf{x}|\mathbf{y})$ are often preferred in this context, because $\mathbf{x}$ will remain positive for both methods. We can see that $D(\mathbf{x}|\mathbf{y})$ and $l(\mathbf{x}|\mathbf{y})$ are closely related and easy to confuse. However, the estimating equations for each are clearly different (see Table 2), so the emphasis is often placed here rather than on the original measures. To further add to the confusion, when $\mathbf{y}$ is already close to satisfying the constraints ($\mathbf{Ay} \approx \mathbf{q}$ or $g(\mathbf{y}) \approx \mathbf{q}$), the three deviance criteria give very similar results, thus explaining the error in Deming and Stephan (1940) (see Section 4).

2.2. Solving for Linear and Nonlinear Constraints

Suppose we are given a vector $\mathbf{y}$ and wish to find the $\mathbf{x}$ satisfying a possibly nonlinear constraint $g(\mathbf{x}) = \mathbf{q}$ for some vector-valued function $g(\mathbf{x})$ with derivative matrix $\mathbf{D}_g(\mathbf{x})$ (the linear form $\mathbf{Ax} = \mathbf{q}$ is a special case with $\mathbf{D}_g(\mathbf{x}) = \mathbf{A}'$). Since such an $\mathbf{x}$ will generally not be unique, let $\mathbf{x}$ minimize the deviance $d(\mathbf{x}|\mathbf{y})$. Assume $d(\mathbf{x}|\mathbf{x}) = 0$ for all appropriate $\mathbf{x}$. However $d(\mathbf{x}|\mathbf{y})$ need not be symmetric $d(\mathbf{x}|\mathbf{y}) \neq d(\mathbf{y}|\mathbf{x})$. We assume the derivative

Table 2. Five common deviance measures (Deville and Särndal 1992) and the corresponding functions needed for estimation

Name	Deviance	$d^{(1)}$	$h[\mathbf{u}]$	$h^{(1)}[\mathbf{u}]$
Quadratic	$(\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$	$(\mathbf{x} - \mathbf{y})$	$\mathbf{y} + \mathbf{u}$	1
Discrimination	$\mathbf{1}'\mathbf{W}_d\left[\mathbf{x}\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right] - \mathbf{x} + \mathbf{y}\right]$	$\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right]$	$[\mathbf{y}\exp[\mathbf{u}]]$	$[\mathbf{y}\exp[\mathbf{u}]]$
Hellinger	$(\sqrt{\mathbf{x}} - \sqrt{\mathbf{y}})'\mathbf{W}_d(\sqrt{\mathbf{x}} - \sqrt{\mathbf{y}})$	$1 - \left[\frac{\mathbf{y}}{\mathbf{x}}\right]^{\frac{1}{2}}$	$[\mathbf{y}[1 - \mathbf{u}]^{-2}]$	$2[\mathbf{y}[1 - \mathbf{u}]^{-3}]$
Poisson	$\mathbf{1}'\mathbf{W}_d\left[\mathbf{y}\log\left[\frac{\mathbf{y}}{\mathbf{x}}\right] - \mathbf{y} + \mathbf{x}\right]$	$1 - \left[\frac{\mathbf{y}}{\mathbf{x}}\right]$	$[\mathbf{y}[1 - \mathbf{u}]^{-1}]$	$[\mathbf{y}[1 - \mathbf{u}]^{-2}]$
Alternative Quadratic	$(\mathbf{x} - \mathbf{y})'\mathbf{W}_d\langle\mathbf{x}\rangle^{-1}(\mathbf{x} - \mathbf{y})$	$1 - \left[\frac{\mathbf{y}}{\mathbf{x}}\right]^2$	$[\mathbf{y}[1 - \mathbf{u}]^{-\frac{1}{2}}]$	$\frac{1}{2}[\mathbf{y}[1 - \mathbf{u}]^{-\frac{3}{2}}]$

$\partial d(\mathbf{x}|\mathbf{y})/\partial \mathbf{x} = \mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y})$ is composed of well-defined elementwise invertible functions on the $\mathbf{x}$ vector. In other words, the *ith* element $d^{(1)}(\mathbf{x}|\mathbf{y})_i$ only contains information from $\mathbf{x}_i$ and $\mathbf{y}_i$ not $\mathbf{x}_j$ or $\mathbf{y}_j$. We also assume $d^{(1)}(\mathbf{x}, \mathbf{x}) = \mathbf{0}$. It's clear that $\chi^2(\mathbf{x}|\mathbf{y})$, $l(\mathbf{x}|\mathbf{y})$, and $D(\mathbf{x}|\mathbf{y})$ are each examples of $d(\mathbf{x}|\mathbf{y})$ (see Table 2 for these and two more from Deville and Särndal 1992). We take $\mathbf{W}$ to be a symmetric and invertible weight matrix.

In order to minimize $d(\mathbf{x}|\mathbf{y})$, subject to constraints $g(\mathbf{x}) = \mathbf{q}$, we use the method of Lagrange multipliers (see, for example Stewart 2011). When such a solution $\mathbf{x}$ exists, the derivatives $\mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y})$ are parallel to the columns in $\mathbf{D}_g(\mathbf{x})$, the derivatives of each of the k constraints. The $k \times 1$ vector $\boldsymbol{\lambda}$ scales for the differences in magnitude of these parallel vectors. Symbolically,

$$\mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}, \tag{1}$$

or equivalently,

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}.$$

Since $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{u}$ is an elementwise invertible operation on $\mathbf{x}$ producing the vector $\mathbf{u}$, the inverse function $h(\mathbf{u}) = \mathbf{x}$ exists and is also elementwise. Together with $g(\mathbf{x}) = \mathbf{q}$, we obtain the following estimating equations

$$\begin{aligned} \mathbf{x} &= h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}] \\ \mathbf{q} &= g(h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}]). \end{aligned} \tag{2}$$

Two properties become apparent from (1) and (2):

Lemma 1. *The solution* $\mathbf{x}$ *to (2) is invariant to the choice of the scalar* $\alpha \neq 0$ *in* $\mathbf{W}_{new} = \alpha\mathbf{W}_{old}$.

Example: If we are using $(\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$ or $100(\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$ as the deviance $d(\mathbf{x}|\mathbf{y})$, we will get the same solution $\mathbf{x}$.

Lemma 2. *The solution* $\mathbf{x}$ *to (2) is invariant to the rotation of constraints* $\mathbf{L}g(\mathbf{x}) = \mathbf{L}\mathbf{q}$ *for full rank square rotation matrix* L.

Example: In a two-way table, if we constrain all row and column totals, we have one redundant constraint. Ignoring any one row or column total gives the same solution $\mathbf{x}$.

Proof. Both properties come from $\boldsymbol{\lambda}$ being a dummy variable, an intermediate value used to solve for $\mathbf{x}$. This property implies an invariance to one-to-one transformations. In Equation (1), using $\alpha\mathbf{W}$ is equivalent to using $\boldsymbol{\eta} = \alpha^{-1}\boldsymbol{\lambda}$ as the multiplier. Likewise, rotating $\mathbf{L}g(\mathbf{x})$ will lead to the derivative $\mathbf{D}_g(\mathbf{x})\mathbf{L}'$, which is equivalent to using the rotated multiplier $\boldsymbol{\eta} = \mathbf{L}'\boldsymbol{\lambda}$. $\square$

Now consider partitioning $\mathbf{y}' = [\mathbf{y}'_{-s}, \mathbf{y}'_s]$ and $\mathbf{x}' = [\mathbf{x}'_{-s}, \mathbf{x}'_s]$ indexed by the set s of size n_s and its complementary set $-s$ of size n_{-s}. Define the selection operator $\boldsymbol{\delta}(s) = [\mathbf{0}_{n_s \times n_{-s}}, \mathbf{I}_{n_s}]'$ such that $\mathbf{y}_s = \boldsymbol{\delta}'\mathbf{y}$. The weight matrix $\mathbf{W}$ is also partitioned corresponding to s and $-s$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_a & \mathbf{W}_b \\ \mathbf{W}'_b & \mathbf{W}_c \end{bmatrix}$$

Set values in $\mathbf{W}$ corresponding to the set s equal to 0:

$$\mathbf{W}_0 = \begin{bmatrix} \mathbf{W}_a & \mathbf{0}_{n_{-s} \times n_s} \\ \mathbf{0}_{n_s \times n_{-s}} & \mathbf{0}_{n_s \times n_s} \end{bmatrix}.$$

Then the Moore-Penrose generalized inverse of $\mathbf{W}_0$ is

$$\mathbf{W}_0^- = \begin{bmatrix} \mathbf{W}_a^{-1} & \mathbf{0}_{n_{-s} \times n_s} \\ \mathbf{0}_{n_s \times n_{-s}} & \mathbf{0}_{n_s \times n_s} \end{bmatrix}.$$

Two more properties of the estimating equations (1) and (2) are now available:

Lemma 3. *For the estimating equations in (2) we add additional equality constraints of the form* $\mathbf{x}_s = \mathbf{y}_s$. *The following implementations give equivalent solutions for* $\mathbf{x}_{-s}$:

- *Augment the constraint targets* $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{y}_s']$ *and the corresponding equations* $g^{*\prime} = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.
- *Keep the original* $\mathbf{q}$ *and* $g(\mathbf{x})$ *and substitute* $\mathbf{W} = \mathbf{W}_0$ *and* $\mathbf{W}^{-1} = \mathbf{W}_0^-$.

Lemma 4. *For the estimating equations in (2) we add additional equality constraints of the form* $\mathbf{x}_s = \mathbf{z}_s$ *for arbitrary values* $\mathbf{z}_s \neq \mathbf{y}_s$. *If* $\mathbf{W}$ *is diagonal* $(\mathbf{W}_d)$ *or block-diagonal* $(\mathbf{W}_b = \mathbf{0}_{n_{-s} \times n_s})$, *the following implementations give equivalent solutions for* $\mathbf{x}_{-s}$:

- *Augment the constraint targets* $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}_s']$ *and the corresponding equations* $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.
- *Keep the original* $\mathbf{q}$ *and* $g(\mathbf{x})$ *and substitute* $\mathbf{y}_s = \mathbf{z}_s$, $\mathbf{W} = \mathbf{W}_0$, *and* $\mathbf{W}^{-1} = \mathbf{W}_0^-$.

Proof.   See Appendix A for details.   □

2.3.   User Interaction

The goal of this proposed framework is to provide an interface between an informed analyst (or metamodel) and an automated "optimal" procedure which minimizes a deviance measure as described above. The choice of deviance measure will likely be made based on the application area and current conventions (i.e., the discrimination information for *raking* problems). A default weight matrix $\mathbf{W}$ may be a function of estimated variances based on a sample design or a specified model. As we have mentioned in our example in Subsection 1.1, knowledge of the process may be more difficult to fully and directly incorporate into the initial estimation procedures, thus motivating the need for an analyst to make adjustments.

We may expect the user to have limited control over the original $\mathbf{y}$ and the necessary constraints $\mathbf{q}$, leaving only the $\mathbf{W}$ to be adjusted. However, the user is free to provide additional constraints by augmenting $\mathbf{A}$ and $g(\mathbf{x})$. From Lemmas 3 and 4, several of these augmentations can be implemented by changing $\mathbf{y}$ and $\mathbf{W}$, thus preventing an increase in the dimension of the estimating equations (2).

- The user wishes to protect some $\mathbf{y}_s$ from changing. Some values may be the result of previously published data and are therefore ineligible for adjustment.

- The user sets $\mathbf{x}_s = \mathbf{z}_s \neq \mathbf{y}_s$. The user may wish to replace some dubious values or force changes in a direction opposite of the default procedure.
- The user reduces the changes to $\mathbf{y}_s$ without fixing the $\mathbf{x}_s$ values. From Lemma 1, we know that choice of scalar $\alpha \neq 0$ in $\alpha \mathbf{W}$ has no impact on $\mathbf{x}$. However, multiplying subsets of $\mathbf{W}$ by α *will* affect $\mathbf{x}$:

$$\mathbf{W}_\alpha = \begin{bmatrix} \dfrac{1}{\alpha} \mathbf{W}_a & \dfrac{1}{\sqrt{\alpha}} \mathbf{W}_b \\[2ex] \dfrac{1}{\sqrt{\alpha}} \mathbf{W}'_b & \mathbf{W}_c \end{bmatrix}.$$

For $\alpha > 1$, the values of $\mathbf{x}_{-s}$ stray further from $\mathbf{y}_{-s}$, thus absorbing more change.

The sets $\{s\}$ and $\{-s\}$ must be chosen carefully when using $\mathbf{W}_0$ to avoid singularities. Since they are equivalent to adding more constraints to $\mathbf{q}$, we may inadvertently create a constraint on $\mathbf{x}$ which conflicts with $g(\mathbf{x}) = \mathbf{q}$. A finite choice of α, which provides weaker protection, can be used without this problem. Furthermore, $\mathbf{W}_0$ and α may be used together by establishing more than one partitioning set $\{s\}$.

This system provides a good compromise between an automated approach which ignores important expert knowledge for a specific subset $\mathbf{x}_s$, and a completely manual process which may use ad hoc methods to fill in the complementary $\mathbf{x}_{-s}$ values where knowledge is limited.

3. Implementation with Newton's Method

Given $\mathbf{x}$, we can use Newton's method to iteratively solve for the $\boldsymbol{\lambda}$ satisfying the second line of (2). We then update $\mathbf{x}$ and iterate the process until convergence. Denote $h^{(1)}(\mathbf{u}) = \partial h(\mathbf{u})/\partial \mathbf{u}$ as the *matrix* of derivatives. Since $h(\mathbf{u})$ is an elementwise function on $\mathbf{u}$, $h^{(1)}(\mathbf{u}) = \langle h^{(1)}[\mathbf{u}] \rangle$ where $h^{(1)}[\mathbf{u}]$ is a *vector* of elementwise derivatives of $\partial h(u_i)/\partial u_i$. Applying one chain rule for nested function gives:

$$\partial g(h[\mathbf{u}])/\partial \mathbf{u} = \langle h^{(1)}[\mathbf{u}] \rangle \mathbf{D}_g(h[\mathbf{u}]).$$

Then applying another chain rule for a change of variables:

$$\partial g(h[\mathbf{B}\boldsymbol{\lambda}])/\partial \boldsymbol{\lambda} = \mathbf{B}' \langle h^{(1)}[\mathbf{B}\boldsymbol{\lambda}] \rangle \mathbf{D}_g(h[\mathbf{B}\boldsymbol{\lambda}]),$$

where $\mathbf{B}$ is an arbitrary matrix. Let $\mathbf{B} = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})$. For a given $\mathbf{x}^i$, Newton's method becomes:

$$\begin{aligned} \boldsymbol{\lambda}_i^{j+1} = \boldsymbol{\lambda}_i^j &+ \left[\mathbf{D}'_g(\mathbf{x}^i)\mathbf{W}^{-1} \langle h^{(1)}[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j] \rangle \mathbf{D}_g\left(h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j]\right) \right]^{-1} \\ &\times \left(\mathbf{q} - g\left(h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j]\right) \right). \end{aligned} \tag{3}$$

We update $\mathbf{x}_i$ in an outer loop to satisfy the first line of (2):

$$\mathbf{x}^{i+1} = h\left[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right]. \tag{4}$$

After convergence, the estimate $\mathbf{x}$ will be the same regardless of rotation (Lemma 2). However, rotations of the constraints may lead to different intermediate values for

(3) and (4). By updating $\boldsymbol{\eta} = \mathbf{L}'\boldsymbol{\lambda}$ using $\mathbf{L}g(\cdot)$, $\mathbf{Lq}$, and $\mathbf{D}_g(\cdot)\mathbf{L}'$, we will get the same final solution, but different $\mathbf{x}^i$ and $\boldsymbol{\lambda}_i^j$ before convergence.

3.1. Linear Constraints

For linear constraints, $h(\mathbf{u})$ will eliminate the need to iterate (4) for every $d(\mathbf{x}|\mathbf{y})$. For the quadratic deviance $\chi^2(\mathbf{x}|\mathbf{y})$, the inner loop (3) is one step, thus leading to the closed form solution

$$\mathbf{x} = \mathbf{y} + \mathbf{W}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}')^{-1}(\mathbf{q} - \mathbf{A}\mathbf{y}). \tag{5}$$

This result is common in the econometrics and engineering literature (Green 2000; Pizzinga 2010), where the $\mathbf{y}$ are least squares estimates of regression coefficients and $\mathbf{W}$ is their covariance matrix. The linear case with diagonal $\mathbf{W}_d$ for the quadratic and discrimination information deviances is available in the survey literature discussed above.

3.2. Alternatives to Newton's Method

We have presented a Newton method above to provide a general approach that can utilize different deviance measures and can accommodate both linear and nonlinear constraints. Many alternatives to Newton methods exist for specific optimization problems. In the survey literature, alternatives for solving the estimating equations (2) tend to be specific to *one* deviance measure and *linear* constraints. For example, iterative proportional fitting (IPF) is a popular way to impose linear restrictions on the cells of a multiway table using the discrimination information deviance. Software for IPF include the R function "loglin" and the SAS subroutine "ipf". The function "apop_rake" in the Apophenia library (http://apophenia.info/) implements IPF in a low-level programming language.

We emphasize Newton's method for two main reasons. Firstly, Newton's method is applicable to more general classes of constraints, deviance functions, and data structures. It is not limited to linear constraints on multiway tables, but can apply simple (i.e., continuously differentiable) nonlinear constraints to any data set that can be represented as an array. Secondly, the procedures of Subsection 2.3 for incorporating user input via the modification of $\mathbf{W}$ and $g(\mathbf{x})$ are readily implemented with Newton's method. Choosing between methods may depend on software availability, the experience of the user, and the size and nature of the data set. However, the properties of the solutions (see the Lemmas above) and the use of a user framework come from the estimating equations (2) and therefore hold regardless of the manner in which a solution was obtained (IPF, Newton's method, stochastic search, etc.).

Within our proposed Newton method there are alternatives to using $h(\mathbf{u})$. Let $h_{\mathbf{x}}(\mathbf{u})$ be a function of $\mathbf{u}$ given $\mathbf{x}$ satisfying $h_{\mathbf{x}}(d^{(1)}(\mathbf{x})) = \mathbf{x}$. The inverse need not be true: $d^{(1)}(h_{\mathbf{x}}(\mathbf{u})) \neq \mathrm{u}$. Obviously $h(\mathbf{u})$ is a special case of $h_{\mathbf{x}}(\mathbf{u})$. For the Poisson deviance, we choose $h_{\mathbf{x}}(\mathbf{u}) = [\mathbf{xu} + \mathbf{y}]$ with $h_{\mathbf{x}}^{(1)}[\mathbf{u}] = \mathbf{x}$, which is *not* a function of $\mathbf{u}$. For this choice of $h_{\mathbf{x}}(\mathbf{u})$, for linear constraints, we need an outer loop (4), but not an inner (3) loop for $\boldsymbol{\lambda}$. See Appendix B for details. Since both the $h(\mathbf{u})$ and the $h_{\mathbf{x}}(\mathbf{u})$ approaches lead to the same solution at convergence, preference between the two methods may lie in interpretability of

the intermediate steps. For example, we prefer $h_{\mathbf{x}}(\mathbf{u})$ for the Poisson deviance, because the steps are the same as for the quadratic deviance, but with $\langle \mathbf{x} \rangle \mathbf{W}_d^{-1}$ replacing the $\mathbf{W}^{-1}$. Thus we can minimize the Poisson deviance by iteratively using the methods for the quadratic deviance. Using $h(\mathbf{u})$ for the Poisson deviance recreates the *pseudo-empirical maximum likelihood estimator* (Chen and Sitter 1999) since the solution $\mathbf{x}$ is invariant to choosing $\boldsymbol{\eta} = -\boldsymbol{\lambda}$. Chen et al. (2002) give an alternative iterative Newton method for linear constraints using this estimator.

4. Deming and Stephan (1940) Revisited

We revisit a classic example of *raking* by using our generalized techniques on the data set from Deming and Stephan (1940). The observed data (Table 3) are cell counts N_{ij} in a two-way table with margins $N_{i.}$ and $N_{.j}$ for rows $i \in 1, \ldots, 6$ and columns $j \in 1, \ldots, 4$. The constrained margins $M_{i.}$ and $M_{.j}$ are the targets q. The grand totals $N_{..} = M_{..}$ by coincidence and need not be true in general. The objective is to find cell counts M_{ij} that are closest to N_{ij} in terms of deviance, while satisfying the marginal constraints.

Although Deming and Stephan assert that their method of iterative proportional fitting (IPF) minimizes the least squares deviance $\chi^2(\mathbf{x}|\mathbf{y})$, IPF actually minimizes the discrimination deviance $D(\mathbf{x}|\mathbf{y})$ (Deville and Särndal 1992). We can obtain the estimates that minimize $\chi^2(\mathbf{x}|\mathbf{y})$ in one step (5) and use iteration (3) to obtain estimates minimizing $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$. As it turns out, the estimates are quite close across the three deviance measures.

Using the discrimination deviance, we wish to compare the original results to those from two hypothetical user actions (Figure 3). For the default choice, it seems that the row margins are dominant (rows are all $+$ or all $-$) and that most change occurs in the first column (darkest).

- The user specifies two cells ($M_{3,1} = 1{,}516$ and $M_{5,4} = 160$) and prevents these from changing. These are changes in the opposite direction from the default. Therefore the rest of the values in those rows and columns must take on more change (darker) and may switch direction ($-$ to $+$ or $+$ to $-$).
- The user down-weights columns 3 and 4 by a factor of $\alpha = 5$. This allows the values in these columns to absorb more change and thus provides a weak protection for

Table 3. Data from Deming and Stephan (1940). Margin targets increase (light) and decrease (dark)

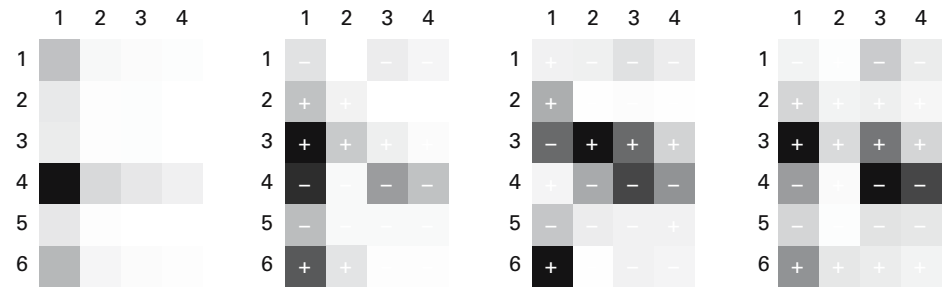| $i \backslash j$	1	2	3	4	$N_{i.}$	$M_{i.}$
1	3,623	781	557	313	5,274	5,252
2	1,570	395	251	155	2,371	2,395
3	1,553	419	264	116	2,352	2,432
4	10,538	2,455	1,706	1,160	15,859	15,766
5	1,681	353	171	154	2,359	2,330
6	3,882	857	544	339	5,622	5,662
$N_{.j}$	22,847	5,260	3,493	2,237	33,837	
$M_{.j}$	22,877	5,285	3,462	2,213		33,837

Fig. 3. Heat Maps using the Discrimination Measure: (left to right) Original Data, Default Changes, User Set Changes ($M_{3,1}$ and $M_{5,4}$), and Reweighted Changes ($M_{.3}$ and $M_{.4}$). White to black increases counts or size of change. Signs ($-/+$) show direction of change

columns 1 and 2. We notice that the general pattern of changes is similar to the default, but columns 3 and 4 are darker and columns 1 and 2 lighter compared to the original solution. This confirms that we have indeed shifted more change onto the last two columns.

4.1. Implementation

First we stack **y** by rows:

$$\mathbf{y} = [N_{i=1}, N_{i=2}, N_{i=3}, N_{i=4}, N_{i=5}, N_{i=6}]'.$$

Then we formulate **q**, remembering to remove one redundant constraint $M_{.1}$ (Lemma 2 assures us that any choice of row or column margin will do):

$$\mathbf{q} = [M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{5.}, M_{6.}, M_{.2}, M_{.3}, M_{.4}]'.$$

Next we construct **A**, which is simply a table of 1s and 0s. (Table 4 shows the transpose $\mathbf{A}'$). For example, the *2nd* column of **A**, (row of $\mathbf{A}'$) corresponds to $x_2 = M_{1,2}$. This cell is involved in the first constraint $M_{1.} = \sum_j M_{1,j}$ and the seventh constraint $M_{.2} = \sum_i M_{i,2}$, so the corresponding values in **A** have 1s. The rest of the entries for x_2 are 0s. We use $\mathbf{W} = \langle \mathbf{y} \rangle^{-1}$ for $\chi^2(\mathbf{x}|\mathbf{y})$ and $\mathbf{W}_d = \mathbf{I}_{24}$, the identity matrix, for $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$. For fixing values, we construct $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$ with zeros for setting $M_{3,1} = 1{,}516$ and $M_{5,4} = 160$. For down-weighting values, we pre- and postmultiply **W** by a diagonal matrix with 1 for columns 1 and 2 and $\sqrt{1/5}$ for columns 3 and 4.

5. Simulated Survey Example

We now provide more detail for the motivating example in Subsection 1.1, in which data occur in triplets of *numerator* (**n**), *denominator* (**d**), and the *ratio* (**r**) of the two. At each level of aggregation (individual, regional, national), we only need to focus on two of the three. It is often the case that we have already set (and published) triplets at a higher level of aggregation (national totals) and now wish to set triplets at lower levels constrained to be consistent when aggregated. For example, we would need the totals for **n** and **d** to sum to the higher level totals. In the context of the methods discussed above, we can do this in at least two ways:

Table 4. Value of $\mathbf{A}'$ for Deming and Stephan (1940) data

#	$x_\#$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$j=2$	$j=3$	$j=4$
1	$M_{(1,1)}$	1	0	0	0	0	0	0	0	0
2	$M_{(1,2)}$	1	0	0	0	0	0	1	0	0
3	$M_{(1,3)}$	1	0	0	0	0	0	0	1	0
4	$M_{(1,4)}$	1	0	0	0	0	0	0	0	1
5	$M_{(2,1)}$	0	1	0	0	0	0	0	0	0
6	$M_{(2,2)}$	0	1	0	0	0	0	1	0	0
7	$M_{(2,3)}$	0	1	0	0	0	0	0	1	0
8	$M_{(2,4)}$	0	1	0	0	0	0	0	0	1
9	$M_{(3,1)}$	0	0	1	0	0	0	0	0	0
10	$M_{(3,2)}$	0	0	1	0	0	0	1	0	0
11	$M_{(3,3)}$	0	0	1	0	0	0	0	1	0
12	$M_{(3,4)}$	0	0	1	0	0	0	0	0	1
13	$M_{(4,1)}$	0	0	0	1	0	0	0	0	0
14	$M_{(4,2)}$	0	0	0	1	0	0	1	0	0
15	$M_{(4,3)}$	0	0	0	1	0	0	0	1	0
16	$M_{(4,4)}$	0	0	0	1	0	0	0	0	1
17	$M_{(5,1)}$	0	0	0	0	1	0	0	0	0
18	$M_{(5,2)}$	0	0	0	0	1	0	1	0	0
19	$M_{(5,3)}$	0	0	0	0	1	0	0	1	0
20	$M_{(5,4)}$	0	0	0	0	1	0	0	0	1
21	$M_{(6,1)}$	0	0	0	0	0	1	0	0	0
22	$M_{(6,2)}$	0	0	0	0	0	1	1	0	0
23	$M_{(6,3)}$	0	0	0	0	0	1	0	1	0
24	$M_{(6,4)}$	0	0	0	0	0	1	0	0	1

- We can focus on adjusting $\mathbf{n}$ and $\mathbf{d}$ leading to linear constraints

$$\mathbf{q} = \left[\sum_i n_i, \sum_i d_i \right]' = \mathbf{Ax}.$$

- We can focus on adjusting $\mathbf{d}$ and $\mathbf{r}$ leading to nonlinear constraints

$$\mathbf{q} = \left[\sum_i r_i d_i, \sum_i d_i \right]' = g(\mathbf{x}).$$

We can motivate the first method based on simplicity. However, the second method appeals to us if there is more intuition for $\mathbf{r}$ than $\mathbf{n}$. It may also be the case that $\mathbf{r}$ is more independent of $\mathbf{d}$ than $\mathbf{n}$ is. For example, agricultural agencies publish total production ($\mathbf{n}$), harvested area ($\mathbf{d}$), and yield per area ($\mathbf{r}$) for major crops. Focusing on production $\mathbf{n}$ may be overemphasizing constraints on area $\mathbf{d}$. In addition, there is much scientific and commodity knowledge about the values for yield $\mathbf{r}$.

We consider a simulated set of triplets for $i = 1, \ldots, 20$ artificial regions which grow soybeans. Based on published values for the U.S. (www.nass.usda.gov), we choose a symmetric distribution of soybean yields ranging between 15–55 bu/acre and a skewed

distribution for harvested area over a 20-fold range with units in either the 100s (county) or 1,000s (state) of acres.

- Simulate 20 values for $d_i \sim \left[Unif\left(\frac{1}{1,000}, \frac{1}{50}\right) \right]^{-1}$.
- Simulate 20 values for $r_i \sim N(\mu = 35, \sigma = 10)$ independently of d_i.
- Calculate $n_i = d_i \times r_i$ for each value.

The resulting data were shown in Table 1. The target values for constraints were arbitrarily chosen such that the target total for **n** and **d** were 95% and 103% respectively of the observed totals.

5.1. A Hypothetical User Experience

A hypothetical user wants to constrain the triplets data from Table 1 with the option of imposing adjustments based on experience and judgment. The user has experience with *raking*, so decides to use the discrimination deviance. The nonlinear formulation is new to the user, so both it and the linear approach are run in parallel to compare the results. The actions of the user are summarized as a flowchart in Figure 4.

   The user begins with the default solutions from our motivating example above (Figure 1), but then realizes that regions 1, 3, 10, 12, and 14 are only sampled annually and
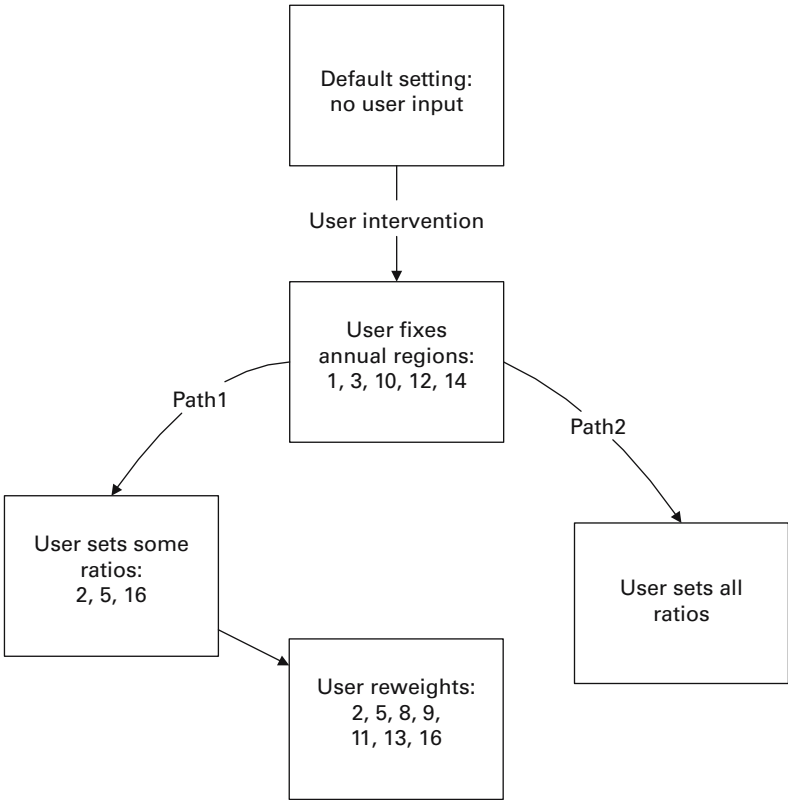

Fig. 4.   *Process flow of user decisions and estimates for the triplets data set*

have not been sampled in the current survey. Instead, the most recent valid values have been passed forward. These have already been published and are therefore not eligible to be changed. To protect these values, the user adds 0s into the corresponding entries of the weight matrix **W**. The procedure is run again and new values are produced.

Now the user looks at the yield ratios **r** more carefully and compares them to the survey estimates. Historically, the survey gives high quality estimates for this ratio. If possible, the user would like to keep these ratios fixed. The user decides to take two different paths and explore their impact (Figures 5 and 6):

- Path 1: The user sets ratios (rounded to the integer) for regions 2, 5, and 16. Several regions change more than the user can comfortably justify. Regions 2, 5, 8, 9, 11, 13, and 16 are reweighted ($\alpha = 5$) to absorb more change from the other regions. The heat maps confirm that adjustments are now more concentrated (darker) in these regions.
- Path 2: The user fixes all ratios and is surprised to see that the linear and nonlinear approaches give identical results. By setting all **r**, both **r** and **n** are eliminated from adjustment, producing two linear constraints on **d**. Regions with higher yield **r** have harvested area **d** decreased, whereas those with lower yield have harvested area
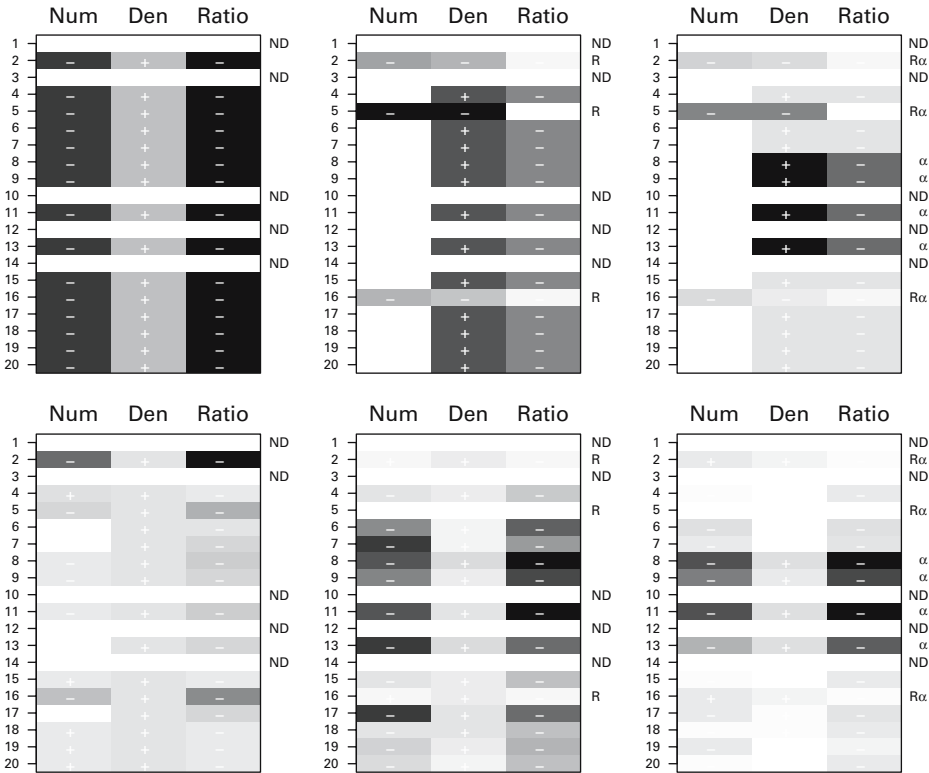

Fig. 5. *Heat Maps for Path 1 (using discrimination deviance) for linear (top row) and nonlinear (bottom row) approaches. User successively adds constraints (left to right): Fixing annual regions (ND), setting yield ratios (R), reweighting to redistribute (α)*
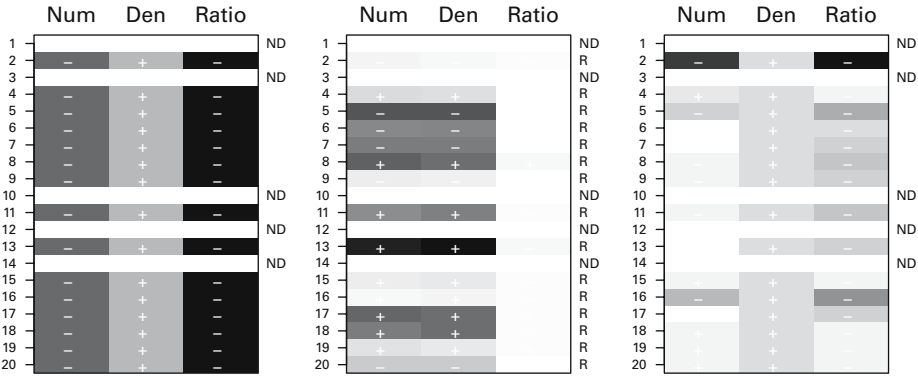
Fig. 6. Heat Maps for Path 2 (using discrimination deviance). Linear (left) and nonlinear (right) approaches converge when the user sets all yield ratios (center). Num and Den fixed (ND), Ratio set (R)

increased. Thus the overall production $\mathbf{n}$ has been decreased, but the overall harvested area $\mathbf{d}$ has been increased to simultaneously meet both aggregate targets.

5.2. Implementation Details

To implement the process above, we define $\mathbf{y}' = [\mathbf{n}'_\mathbf{y}, \mathbf{d}'_\mathbf{y}, \mathbf{r}'_\mathbf{y}]$ as the stacked set of unconstrained values. We seek the corresponding stacked constrained values $\mathbf{x}' = [\mathbf{n}'_\mathbf{x}, \mathbf{d}'_\mathbf{x}, \mathbf{r}'_\mathbf{x}]$ whose aggregate values satisfy the target $\mathbf{q}' = [120{,}237.35, 3{,}935.33, 30.55]$. For the linear approach, we use the numerator and denominator directly: $\mathbf{y}'_l = [\mathbf{n}'_\mathbf{y}, \mathbf{d}'_\mathbf{y}]$, $\mathbf{x}'_l = [\mathbf{n}'_\mathbf{x}, \mathbf{d}'_\mathbf{x}]$, and $\mathbf{q}'_l = [120{,}237.35, 3{,}935.33]$. Depending on the user's choices, the $\mathbf{A}$ matrix varies (Table 5). For the default settings, $\mathbf{A}$ is simply two rows of indicators, with 1 where an element of $\mathbf{x}_l$ is present in the sums $\{\sum n_i, \sum d_i\}$ and 0 otherwise. For Path 1, the user sets some r_i. Then for the total $\sum n_i$, the term $r_i d_i$ replaces some n_i. Thus the corresponding entries in $\mathbf{A}$ are 0 for n_i and r_i for d_i. Path 2 has a similar $\mathbf{A}$ matrix except with more r_i present. No adjustment to $\mathbf{A}$ is needed for jointly fixing the pairs $\{n_i, d_i\}$ for $i = 1, 3, 10, 12, 14$. For these cases, we construct $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$ with corresponding zeros. We suggest using $\mathbf{W} = \langle\mathbf{y}\rangle^{-1}$ for $\chi^2(\mathbf{x}|\mathbf{y})$ and $\mathbf{W}_d = \mathbf{I}_{40}$ for $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$.

For the nonlinear approach, we directly adjust denominator and ratio: $\mathbf{y}'_{nl} = [\mathbf{d}'_\mathbf{y}, \mathbf{r}'_\mathbf{y}]$, $\mathbf{x}'_{nl} = [\mathbf{d}'_\mathbf{x}, \mathbf{r}'_\mathbf{x}]$, but still use the total production and harvested area as the targets $\mathbf{q}'_{nl} = [120{,}237.35, 3{,}935.33]$. We define $g(\mathbf{x}) = [\mathbf{r}'_\mathbf{x}\mathbf{d}_\mathbf{x}, 1'\mathbf{d}_\mathbf{x}]'$. Then

$$\mathbf{D}_g(\mathbf{x}) = \begin{bmatrix} \mathbf{r}_\mathbf{x} & 1 \\ \mathbf{d}_\mathbf{x} & 0 \end{bmatrix}.$$

For the nonlinear case, fixing ratios r_i introduces more zeros into $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$. Setting values for r_i will change the initial $\mathbf{y}_{nl}$ (as opposed to changing $\mathbf{A}$ for the linear approach).

6. Conclusions and Future Work

In this work, we have provided an overview of constrained estimation and solutions for several common deviance measures based on first principles. While these tools are useful, our main goal was to use them to motivate a framework in which an analyst and a default

Table 5.   Values of A' for Triplets Example

#	$x_{\#}$	Default $\sum n_i$	Default $\sum d_i$	Path 1 $\sum n_i$	Path 1 $\sum d_i$	Path 2 $\sum n_i$	Path 2 $\sum d_i$
1	n_1	1	0	1	0	1	0
2	n_2	1	0	0	0	0	0
3	n_3	1	0	1	0	1	0
4	n_4	1	0	1	0	0	0
5	n_5	1	0	0	0	0	0
6	n_6	1	0	1	0	0	0
7	n_7	1	0	1	0	0	0
8	n_8	1	0	1	0	0	0
9	n_9	1	0	1	0	0	0
10	n_{10}	1	0	1	0	1	0
11	n_{11}	1	0	1	0	0	0
12	n_{12}	1	0	1	0	1	0
13	n_{13}	1	0	1	0	0	0
14	n_{14}	1	0	1	0	1	0
15	n_{15}	1	0	1	0	0	0
16	n_{16}	1	0	0	0	0	0
17	n_{17}	1	0	1	0	0	0
18	n_{18}	1	0	1	0	0	0
19	n_{19}	1	0	1	0	0	0
20	n_{20}	1	0	1	0	0	0
21	d_1	0	1	0	1	0	1
22	d_2	0	1	r_2	1	r_2	1
23	d_3	0	1	0	1	0	1
24	d_4	0	1	0	1	r_4	1
25	d_5	0	1	r_5	1	r_5	1
26	d_6	0	1	0	1	r_6	1
27	d_7	0	1	0	1	r_7	1
28	d_8	0	1	0	1	r_8	1
29	d_9	0	1	0	1	r_9	1
30	d_{10}	0	1	0	1	0	1
31	d_{11}	0	1	0	1	r_{11}	1
32	d_{12}	0	1	0	1	0	1
33	d_{13}	0	1	0	1	r_{13}	1
34	d_{14}	0	1	0	1	0	1
35	d_{15}	0	1	0	1	r_{15}	1
36	d_{16}	0	1	r_{16}	1	r_{16}	1
37	d_{17}	0	1	0	1	r_{17}	1
38	d_{18}	0	1	0	1	r_{18}	1
39	d_{19}	0	1	0	1	r_{19}	1
40	d_{20}	0	1	0	1	r_{20}	1

optimal procedure interact, allowing the user to input extra knowledge to create optimal "user-constrained" results. We demonstrated this framework on a classic *raking* example with linear constraints in the form of margins. We then examined two different approaches to a standard survey problem of constraining aggregate totals and ratios, one implying linear constraints and the other nonlinear ones. Overall, these methods provide a

framework from which to build an interface between automated model processes and expert knowledge via an analyst or metamodel.

We have deliberately avoided discussion of expectations and variances. It should be clear that $\mathbf{y}$ is often stochastic with an estimated distribution (perhaps just a mean and variance). However, $\mathbf{W}$ can be a function of $\mathbf{y}$ (as is often the case for the quadratic measure). More importantly, the user's choice of $\mathbf{q}$, α, and $\mathbf{z}_s$ (and whether or not to use the default settings) is undoubtedly related to both $\mathbf{y}$ and external information. Thus the distribution of $\mathbf{x}$ has connections to both $\mathbf{y}$ and the decision process of the analyst. Tools for finding asymptotic variance estimates when the $\mathbf{y}$ are sampling weights are already available in the literature for calibration (Deville and Särndal 1992; D'Arrigo and Skinner 2010) and would require minor modifications to apply to our setting. However, modeling the uncertainty associated with the decision process of the analyst might first involve implementing these methods and capturing and exploring their behavior. We feel that the framework here is sufficient to begin this process. Data-mining and decision science methods may then be able to construct larger metamodels which incorporate more of these sources of variability.

Appendix A.   Justification of $\mathbf{W}_0^-$

Instead of focusing on solving for $\boldsymbol{\lambda}$ and $\mathbf{x}$ (see Subsection 2.2), we assume this is possible and consider the question of whether to modify $\mathbf{q}$ or $\mathbf{W}$ to enforce additional equality constraints $\mathbf{x}_s = \mathbf{y}_s$ or $\mathbf{x}_s = \mathbf{z}_s$ where $\mathbf{x}' = [\mathbf{x}'_{-s}, \mathbf{x}'_s]$ and $\mathbf{y}' = [\mathbf{y}'_{-s}, \mathbf{y}'_s]$ are partitioned and $\mathbf{z}_s \neq \mathbf{y}_s$ is arbitrary.

One option is to augment the $\mathbf{q}$ vector: $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{y}'_s]$ (or $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}'_s]$). The corresponding $g(\mathbf{x})$ is augmented $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$. Then $D_g(\mathbf{x})$ is also augmented $D_g^*(\mathbf{x}) = [\mathbf{D}_g(\mathbf{x}), \boldsymbol{\delta}]$. We would then use $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$ (with $\boldsymbol{\lambda}^{*\prime} = [\boldsymbol{\lambda}', \boldsymbol{\eta}']$) and $\mathbf{q}^*$ to solve for $\mathbf{x}$.

Another option is to change the $\mathbf{W}$ or $\mathbf{W}^{-1}$ matrices. Since setting equalities for $\mathbf{x}_s$ should reduce the dimensions of the problem, introducing 0s into $\mathbf{W}$ may also work. We partition $\mathbf{W}$ accordingly and use $\mathbf{W}_0$ and $\mathbf{W}_0^-$ as defined in Subsection 2.3. Note that $\mathbf{W}_0^-$ and $\boldsymbol{\delta}$ are related by the following:

$$\mathbf{W}_0^- = \mathbf{W}^{-1} - \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}.$$

This can be verified using the block inverse formulas to confirm

$$\mathbf{W}_a^{-1} = \{\mathbf{W}^{-1}\}_a - \{\mathbf{W}^{-1}\}_b\{\{\mathbf{W}^{-1}\}_c\}^{-1}\{\mathbf{W}^{-1}\}_b',$$

where

$$\mathbf{W}^{-1} = \begin{bmatrix} \{\mathbf{W}^{-1}\}_a & \{\mathbf{W}^{-1}\}_b \\ \{\mathbf{W}^{-1}\}_b' & \{\mathbf{W}^{-1}\}_c \end{bmatrix}.$$

A.1.   Proof of Lemma 3

This scenario occurs when a user decides that some of the $\mathbf{y}_s$ need to be protected and are kept unchanged during the constraint process. We will show that the $\mathbf{W}_0^-$ and the $\mathbf{D}_g^*(\mathbf{x})$ methods lead to equivalent solutions for $\mathbf{x}_s = \mathbf{y}_s$.

Starting with the $\mathbf{D}_g^*(\mathbf{x})$ equations:

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$$

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta}$$

$$\boldsymbol{\delta}'d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{0}_{n_s} = \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta}$$

$$\boldsymbol{\eta} = -(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}.$$

Then plugging $\boldsymbol{\eta}$ back in:

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$$

$$= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} - \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$$

$$= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$$

This is the same as substituting $\mathbf{W}_0^-$ for $\mathbf{W}^{-1}$ in the default (no user input) setting.

Not only does the $\mathbf{W}_0^-$ approach give the same results as the $\mathbf{D}_g^*(\mathbf{x})$ approach, it also reduces dimensions instead of increasing them:

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$$

$$\begin{bmatrix} \{d^{(1)}(\mathbf{x}|\mathbf{y})\}_{-s} \\ \mathbf{0}_{n_s} \end{bmatrix} = \begin{bmatrix} \{\mathbf{W}_\alpha^{-1}\mathbf{D}_g(\mathbf{x})\}_{-s}\boldsymbol{\lambda} \\ \mathbf{0}_{n_s} \end{bmatrix}.$$

Thus we only need to keep track of n_{-s} equations and k constraints for the $\mathbf{W}_0^-$ approach instead of n equations and $k + n_s$ constraints with the $\mathbf{D}_g^*(\mathbf{x})$ approach.

A.2. Proof of Lemma 4

Now let us consider the case that $\mathbf{x}_s = \mathbf{z}_s$ for some arbitrary $\mathbf{z}_s \neq \mathbf{y}_s$. There are at least two ways to proceed:

- Create $\mathbf{y}^{*\prime} = [\mathbf{y}'_{-s}, \mathbf{z}'_s]$ and use the $\mathbf{W}_0^-$ approach as in the previous section.
- Keep $\mathbf{y}$ and set $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}'_s]$ with $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.

We begin with the second option and explore the conditions under which the two are equivalent. For convenience, define $d^{(1)}(\mathbf{x}_s|\mathbf{y}_s) = \{d^{(1)}(\mathbf{x}|\mathbf{y})\}_s$. Also note that $d^{(1)}(\mathbf{x}_s|\mathbf{z}_s) = \mathbf{0}_{n_s}$.

Starting with the $\mathbf{D}_g^*(\mathbf{x})$ equations:

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$$

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta}$$

$$\boldsymbol{\delta}'d^{(1)}(\mathbf{x}|\mathbf{y}) = d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) = \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta}$$

$$\boldsymbol{\eta} = (\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}[d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) - \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}].$$

Then

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$$

$$= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}[d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) - \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}]$$

$$= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}d^{(1)}(\mathbf{z}_s|\mathbf{y}_s)$$

$$= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \begin{bmatrix} \{\mathbf{W}^{-1}\}_b\{\{\mathbf{W}^{-1}\}_c\}^{-1} \\ \mathbf{I}_{n_s} \end{bmatrix} d^{(1)}(\mathbf{z}_s|\mathbf{y}_s).$$

When $\mathbf{W}$ is diagonal $\mathbf{W}_d$ (or block-diagonal with $\mathbf{W}_b = \mathbf{0}_{n_{-s}\times n_s}$): $\{\mathbf{W}^{-1}\}_b = \mathbf{0}_{n_{-s}\times n_s}$. Then $\{d^{(1)}(\mathbf{x}|\mathbf{y})\}_{-s} = \mathbf{W}_\alpha^{-1}\{\mathbf{D}_g(\mathbf{x})\}_{-s}\boldsymbol{\lambda}$. So solving $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$ for $\mathbf{x}_{-s}$ is equivalent to solving $d^{(1)}(\mathbf{x}|\mathbf{y}^*) = \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$. We would prefer the $\mathbf{y}^*$ method because it allows us to use $\mathbf{W}_0^-$ to reduce dimensions.

When $\mathbf{W}$ is more generally symmetric and invertible (as for the quadratic deviance), we may get two distinct estimates for $\mathbf{x}_{-s}$ from the $\mathbf{D}_g^*(\mathbf{x})$ and $\mathbf{W}_0^-$ approaches. Each approach gives an optimal solution to a set of constraints and slightly different deviance functions. The $\mathbf{W}_0^-$ approach ignores $d^{(1)}(\mathbf{z}_s|\mathbf{y}_s)$, the discrepancy between $\mathbf{y}_s$ and $\mathbf{z}_s$. Whereas the $\mathbf{D}_g^*(\mathbf{x})$ method uses the off-diagonal blocks of $\mathbf{W}$ to incorporate this term.

Appendix B.   Justification of $h_x(\mathbf{u})$ for Poisson Deviance

To obtain the $\mathbf{x}$ which minimizes $l(\mathbf{x}|\mathbf{y})$ subject to the constraint $g(\mathbf{x}) = \mathbf{q}$, we derive alternate estimation equations:

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$$

$$1 - \begin{bmatrix} \dfrac{\mathbf{y}}{\mathbf{x}} \end{bmatrix} = \mathbf{u}$$

$$\mathbf{x} - \mathbf{y} = \langle\mathbf{x}\rangle\mathbf{u}$$

$$\mathbf{x} = \mathbf{y} + \langle\mathbf{x}\rangle\mathbf{u}$$

$$\mathbf{q} = g(\mathbf{y} + \langle\mathbf{x}\rangle\mathbf{u}).$$

Then $h_\mathbf{x}(\mathbf{u}) = [\mathbf{y} + \mathbf{x}\mathbf{u}]$ with $h_\mathbf{x}^{(1)}(\mathbf{u}) = \langle\mathbf{x}\rangle$.

Substituting $h_\mathbf{x}(\mathbf{u})$ and $h_\mathbf{x}^{(1)}(\mathbf{u})$ into (3) and (4), we get an inner iteration

$$\boldsymbol{\lambda}_i^{j+1} = \boldsymbol{\lambda}_i^j + \left[\mathbf{D}_g'(\mathbf{x}^i)\mathbf{W}_d^{-1}\langle\mathbf{x}^i\rangle\mathbf{D}_g\left(\mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right)\right]^{-1}$$

$$\times \left(\mathbf{q} - g\left(\mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right)\right)$$

and an outer iteration

$$\mathbf{x}^{i+1} = \mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i.$$

We suggest $\mathbf{x}^0 = \mathbf{y}$ and $\boldsymbol{\lambda}_0^0 = 0$ as good initial values, with $\boldsymbol{\lambda}_i^0 = \boldsymbol{\lambda}_{i-1}$ from the previous iteration of $\mathbf{x}^i$.

For the linear case $g(\mathbf{x}) = \mathbf{Ax}$, the inner loop (3) is one step, eliminating $\boldsymbol{\lambda}$:

$$\mathbf{x}^{j+1} = \mathbf{y} + \langle \mathbf{x}^j \rangle \mathbf{W}_d^{-1} \mathbf{A}' \big(\mathbf{A} \mathbf{W}_d^{-1} \langle \mathbf{x}^j \rangle \mathbf{A}' \big)^{-1} (\mathbf{q} - \mathbf{Ay}).$$

We suggest starting with $\mathbf{x}^0 = \mathbf{y}$ since that will give an $\mathbf{x}^1$ which minimizes $\chi^2(\mathbf{x}|\mathbf{y})$ when $\mathbf{W} = \langle \mathbf{y} \rangle^{-1} \mathbf{W}_d$.

7. References

Agresti, A. (2002). Categorical Data Analysis, (2nd edition). New York: Wiley.

Chang, T. and Kott, P. S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 555–571.

Chen, J. and Sitter, R. (1999). A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys. Statistica Sinica, 9, 385–406.

Chen, J., Sitter, R., and Wu, C. (2002). Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys. Biometrika, 89, 230–237.

D'Arrigo, J. and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. Survey Methodology, 36, 181–192.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. Annals of Mathematical Statistics, 11, 427–444.

Deville, J. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Estevao, V., Hidiroglou, M., and Särndal, C.-E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, 181–204.

Fuller, W. (2002). Regression Estimation for Survey Samples. Survey Methodology, 28, 5–23.

Ghosh, M. (1992). Constrained Bayes Estimation with Applications. Journal of the American Statistical Association, 87, 533–540.

Green, W.H. (2000). Econometric Analysis, (4th edition). New Jersey: Prentice-Hall.

Ireland, C.T. and Kullback, S. (1968). Contingency Tables with Given Marginals. Biometrika, 55, 179–188.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology, 32, 133–142.

Nandram, B. and Sayit, H. (2011). A Bayesian Analysis of Small Area Probabilities Under a Constraint. Survey Methodology, 37, 137–152.

Pizzinga, A. (2010). Constrained Kalman Filtering: Additional Results. International Statistical Review, 78, 189–208.

Singh, A. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107–115.

Stewart, J. (2011). Multivariate Calculus, (7th edition). Belmont, California: Brooks Cole.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Wang, J., Fuller, W., and Qu, Y. (2008). Small Area Estimation Under a Restriction. Survey Methodology, 34, 29–36.