RESOLVING CONFIDENTIALITY AND DATA QUALITY ISSUES FOR TABULAR DATA

Lawrence H. Cox
National Center for Health Statistics, Centers for Disease Control and Prevention
Hyattsville, MD 20782 USA
LCOX@CDC.GOV

Keywords. Controlled tabular adjustment, linear programming, covariance

1 Introduction

Tabular data are ubiquitous and are a staple of official statistics. Data confidentiality was first investigated for tabular data (Fellegi 1972; Cox 1980). Tabular data are additive and thus naturally related to specialized systems of linear equations: **TX = 0**, where **X** represents the *tabular cells* and **T** the *tabular equations*, the entries of **T** are in the set {-1, 0, +1}, and each row of **T** contains one -1.

A recent methodology for *statistical disclosure limitation* in tabular data is *controlled tabular adjustment* (CTA). This development was motivated by computational complexity, analytical obstacles, and general user dissatisfaction with the prevailing methodology, *complementary cell suppression* (Cox 1980, 1995). Complementary suppression removes from publication all *sensitive cells*—cells that cannot be published due to confidentiality concerns—and in addition removes other, nonsensitive cells to ensure that values of sensitive cells cannot be reconstructed or closely estimated by manipulating linear tabular relationships. Drawbacks of cell suppression for statistical analysis include removal of otherwise useful information and difficulties analyzing tabular systems with cell values missing not-at-random. CTA replaces sensitive cell values with *safe values*, viz., values sufficiently far from the true value. Adjustments throw the tabular system out of kilter, so CTA adjusts some or all of the nonsensitive cells by small amounts to restore additivity. CTA is implemented using mathematical programming methods available in commercial software. Cox (2000) provides the first mathematical model for CTA in the literature.

The main question is the degree to which CTA distorts analytical outcomes--is analysis based on adjusted data in some sense equivalent to that based on original data. We demonstrate how using CTA and linear programming it is possible to preserve means, variances, covariances, correlations and the regression coefficient for univariate and multivariate data. Greater detail can be found in Cox and Kelly (2004) (univariate) and Cox, Kelly and Patil (2004) (multivariate). Our models are based on linear programming, easy to use and applicable to a range of problems.

Section 2 summarizes the original CTA methodology (Cox 2000; Dandekar and Cox 2002). Section 3 provides linear methods for preserving univariate means and variances of original data and for ensuring high correlation between original and adjusted data (Cox and Kelly 2004). Section 4 treats the multivariate case, providing linear programming formulations ensuring that covariances and correlations between two original variables exhibited in original data are preserved in adjusted data (Cox, Kelly and Patil 2004). Section 5 reports computational results. Section 6 provides concluding comments.

2 CTA methodology

CTA is applicable to tabular data in any form but for convenience we focus on magnitude data, where the greatest benefits are to be found. A simple paradigm for statistical disclosure in magnitude data is as follows. A tabulation cell, denoted *i*, comprises *k* respondents (e.g., retail clothing stores in a county) and their data (e.g., retails sales and employment data). The NSO assumes that any respondent is aware of the identity of the other respondents. The *cell value* is the total value of a statistic of interest (e.g., total retail sales), summed over the (nonnegative) *contributions* of each respondent in the cell to this statistic. Denote the cell value $v^{(i)}$ and the respondent contributions $v_j^{(i)}$, ordered from largest to smallest. It is possible for any respondent *J* to compute $v^{(i)} - v_J^{(i)}$ which yields an upper estimate of the contribution of any other respondent. This estimate is closest, in percentage terms, when J = 2 and j = 1. A standard disclosure rule, the *p-percent rule*, declares that the cell value represents *disclosure* if this estimate is closer than p-percent of the largest contribution. The sensitive cells are precisely those failing this condition.

The NSO may also assume that any respondent can use *public knowledge* to estimate the contribution of any other respondent to within q-percent (q > p, e.g., q = 50%). This additional information allows the second largest to estimate $v^{(i)} - v_1^{(i)} - v_2^{(i)}$, the sum of all contributions excluding itself and the largest, to within q-percent. This upper estimate provides the second largest a lower estimate of $v_1^{(i)}$. The *lower* and *upper protection limits* for the cell value equal, respectively, the minimum amount that must be subtracted from (added to) the cell value so that these lower (upper) estimates are at least p-percent away from the true value $v_1^{(i)}$. Numeric values below the lower or above the upper protection limit are *safe values* for the cell. A common NSO practice assumes that these protection limits are equal, to p_i. Complementary cell suppression suppresses all sensitive cells from publication, replacing sensitive values by variables in the tabular system **TX = 0**. Because, almost surely, one or more suppressed sensitive cell value can be estimated to within p-percent of it true value, it is necessary to suppress some nonsensitive cells until no sensitive estimates are within p-percent.

Controlled tabular adjustment replaces each sensitive value with a safe value. This is an improvement over complementary cell suppression as it replaces a suppression symbol by an actual value. However, safe values are not necessarily unbiased estimates of true values. To minimize bias, replace the true value with one of its protection limits, $v^{(i)} - p_i$ or $v^{(i)} + p_i$, viz., with either of the two safe values closest to the original value. Because these assignments throw the tabular system out of kilter, CTA adjusts nonsensitive values to restore additivity. Because choices to adjust each sensitive value down or up are binary, combined these steps define a MILP (Cox 2000). Using heuristics for the binary choices, the resulting *linear programming relaxation* is easily solved.

A (mixed integer) linear program will not assure that analytical properties of original and adjusted data are comparable. Cox and Dandekar (2003) address these issues in three ways. First, sensitive values are replaced by closest possible safe values. Second, bounds are imposed on changes to nonsensitive values to ensure individual adjustments are sufficiently small. Statistically sensible capacities would, e.g., be based on estimated measurement error for each cell e_i. Third, the linear program is optimized for an overall measure of data distortion such as sum of absolute adjustments or minimum sum of percent absolute adjustments, as follows.

Assume there are *n* tabulation cells of which the first *s* are sensitive, original data are represented by the nx1 vector **a**, adjusted data by $a + y^+ - y^-$; and $y = y^+ - y^-$. The MILP of Cox (2000) for minimizing sum of absolute adjustments is:

$$\min \sum_{i=1}^{n} (y_i^- + y_i^+) \quad \text{subject to:}$$

$$I_i \text{ binary, } i = 1, ..., s$$

$$T(y) = 0 \qquad (1)$$

$$y_i^- = p_i(1 - I_i), \quad y_i^+ = p_i I_i, i = 1, .., s$$

$$0 \le y_i^-, y_i^+ \le e_i \quad i = s+1, ..., n$$

The Cox and Dandekar (2003) constraints are useful. Unfortunately, choices for the optimizing measure are limited to linear functions. In the next two sections, we extend this paradigm in two separate directions, focusing on approaches to preserving mean, variance, correlation and regression between original and adjusted data.

Formulation (1) is a mixed integer linear program. The integer part can be solved by exact methods in small to medium sized problems or via heuristics to fix integer variables followed by a linear program relaxation (Cox and Kelly 2004). The remainder of this paper focuses on the problem of preserving data quality under CTA, and is not concerned with how the integer portion is being or has been solved.

3 Using CTA to preserve univariate statistics

We present linear programming formulations for preserving exactly or approximately mean, variance, correlation and regression slope between original and adjusted data.

Preserving mean values is straightforward. Any cell value $\mathbf{a_i}$ can be held fixed by forcing its corresponding adjustment variables y_i^+, y_i^- to zero, viz., set each variable's upper capacity to zero. Means are averages over sums. So, for example, to fix the grand mean, simply fix the grand total. To fix the mean of over any set of variables for which a corresponding variable has not been defined, incorporate a new constraint into the linear system: $\sum(y_i^+ - y_i^-) = 0$, where the sum is taken over the set of variables of interest. The corresponding MILP is:

$$\min c(y) \quad \text{subject to:}$$

$$I_i \text{ binary, } i = 1, ..., s$$

$$T(y) = 0 \qquad (2)$$

$$\sum_{i=1}^{s} (y_i^+ - y_i^-) = 0$$

$$p_i(1 - I_i) \le y_i^- \le q_i(1 - I_i), \quad p_i I_i \le y_i^+ \le q_i I_i$$

$$i = 1, ..., s$$

$$0 \le y_i^-, y_i^+ \le e_i \qquad i = s+1, ..., n$$

$c(y)$ is used to keep adjustments close to their lower limit, e.g., $c(y) = \sum y^+ + y^-$.

For variance, any subset of cells of size *t* with $\bar{y} = 0$,

$$Var(a + y) = (1/t)(\sum((a_i + y_i - (\bar{a} + \bar{y}))^2)$$

$$= Var(a) + (2/t)\sum(a_i - \bar{a})\, y_i + Var(y)$$

Define $L(y) = Cov(a, y)/Var(a)$.

As $\bar{y} = 0$:

$$L(y) = (1/(tVar(a)))\sum_{i=1}^{t}(a_i - \bar{a})\, y_i, \text{ so}$$

$$V(a + y)/V(a) = 2L(y) + (1 + V(y)/V(a))$$

and

$$|V(a + y)/V(a) - 1| = |2L(y) + (V(y)/V(a))|$$

It suffices to minimize $|L(y)|$;as follows:

a) incorporate two new linear constraints into the system (2):

$$w \geq L(y), \quad w \geq -L(y) \qquad (3)$$

b) minimize *w*

Regarding correlation, the objective is to achieve high positive correlation between original and adjusted values. We seek Corr (**a, a + y**) = 1, exactly or approximately. As $\bar{y} = 0$,

$$Corr(a,a+y) = Cov(a,a+y) / \sqrt{Var(a)Var(a+y)}$$

$$= (1 + L(y)) / \sqrt{\mathrm{Var(a + y)/ Var(a)}}$$

With Var(**y**)/Var(**a**) typically small, the denominator should be close to one, and $\min |L(y)|$ subject to (2) should do well in preserving correlation. Note that denominator equal to one is equivalent to preserving variance, which as we have seen also is accomplished via $\min |L(y)|$.

Finally, we seek to preserve ordinary least squares regression $Y = \beta_1 X + \beta_0$ of adjusted data Y = **a + y** on original data X = **a**, viz., we want β_1 near one and β_0 near zero.

$$\beta_1 = Cov(a + y, a) / \mathrm{Var(a)} = 1 + L(y)$$

$$\beta_0 = (\bar{a} + \bar{y}) - \beta_1 \bar{a}$$

As $\bar{y} = 0$, then $\beta_0 = 0$, $\beta_1 = 1$ whenever L(y) = 0 is feasible. Again, this corresponds to min $|L(y)|$ subject to constraints of (2), viz., (3).

4 Using CTA to preserve multivariate statistics

In place of a single data set of data **a** organized in tabular form, viz., **Ta = 0**, to which adjustments **y** are to be made for confidentiality purposes, henceforth we assume there are multiple data sets, all organized within a common tabular structure **T**. For concreteness, we focus on bivariate case. Original data are denoted **a** and **b** and corresponding adjustments to original values are denoted by variables **y** and **z**. In the multivariate situation, preserving covariance and variance is of key importance. Namely, if we can preserve mean values and the variance-covariance matrix of original data, then we have preserved essential properties of original data, particularly for the case of linear statistical models. We also would like to preserve simple linear regression of original data **b** on original data **a** in the adjusted data. These are the objectives of this section.

4.1 Preserving the variance-covariance matrix

The separate copies of model (3) of the preceding section preserve univariate variances Var (**a**) and Var (**b**). To preserve Cov (**a, b**), we require:

$$\mathrm{Cov}(\mathbf{a}, \mathbf{b}) = \mathrm{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) = \mathrm{Cov}(\mathbf{a}, \mathbf{b}) + \mathrm{Cov}(\mathbf{a}, \mathbf{z}) + \mathrm{Cov}(\mathbf{b}, \mathbf{y}) + \mathrm{Cov}(\mathbf{y}, \mathbf{z})$$

Consequently, we seek a precise or approximate solution to:

$$\min |\mathrm{Cov}(\mathbf{a}, \mathbf{z}) + \mathrm{Cov}(\mathbf{b}, \mathbf{y}) + \mathrm{Cov}(\mathbf{y}, \mathbf{z})|, \text{ subject to (3)} \qquad (4)$$

One linear approach to solving (4) is to perform successive alternating linear optimizations, viz., solve (2) for $\mathbf{y} = \mathbf{y_0}$, substitute $\mathbf{y_0}$ into (4) and solve for $\mathbf{z} = \mathbf{z_0}$ and continue in this fashion until an acceptable solution is reached.

4.2 Preserving the simple linear regression coefficient

Our objective is to preserve the estimated regression coefficient under simple linear regression of **b** on **a**. We do not address here related issues of preserving the standard error of the estimate and goodness-of-fit. We seek exactly or approximately:

Cov(**a, b**)/V(**a**)=Cov(**a + y, b + z**)/V(**a + y**)

Var(a+y)/Var(a)=Cov(a+y,b+z)/Cov(a, b)
=1+C(a,z)/C(a,b)+C(b,y)/C(a,b)+C(y,z)/C(a,b)

2L(y)+V(y)/V(a)=(C(a,z)+C(b,y)+C(y,z))/C(a,b

To preserve univariate properties L (y) = 0 exactly or approximately. To preserve bivariate covariance Cov (y, z) = 0 exactly or approximately. So, if in addition we seek to preserve the regression coefficient, then we must satisfy the linear program:

min |(Cov (**a, z**) + Cov (**b, y**)) / Cov (**a, b**)|,
subject to (4) (5)

In implementation, the objective is represented as a near-zero constraint on the absolute value.

4.3 Preserving correlations

The objective here is to ensure that correlations between variables computed on adjusted data are close in value to correlations based on original data, viz., that, exactly or approximately Corr (**a, b**) = Corr (**a + y, b + z**). After some algebra, preserving correlation is equivalent to satisfying, exactly or approximately:

$$\sqrt{\frac{Var(a+y)}{Var(a)}}\sqrt{\frac{Var(b+z)}{Var(b)}} = \frac{Cov(a+y,b+z)}{Cov(a,b)}$$

Methods already included in (5) for preserving both univariate variances and the covariance in many cases will preserve correlation. Otherwise, it may be helpful to employ iteration aimed at controlling the right hand product.

5 Results of computational simulations

We tested our methods on the three two-dimensional tables to analyze the performance of the proposed linear formulations on both univariate and bivariate statistical measures. Three tables were taken from a 4x9x9 three-dimensional table. This table contained actual magnitude data and disclosure was defined by (1 contributor, 70%) dominance rule, viz, a cell is sensitive if the largest contribution exceeds 70% of the cell value. This results in protection levels $p_i = (v_1^i)/0.7 - v^i$. Tables A, B, C contain 6, 5, 4 sensitive cells respectively. Upper bounds (capacities) for adjustments to sensitive and non-sensitive cells were set at 20-percent of cell value.

First, we used the MILP formulation to compute exact solutions for the three instances AB, AC, BC. The MILP formulation used the mean and variance preserving constraints and a covariance change minimization objective, aimed at preserving both univariate and bivariate measures.

Table 1 reports performance on the covariance, correlation, regression coefficient, A data variance, and B data variance measures. Table values are percent change in the corresponding statistics. Means were preserved.

Case	Cov Chg.	Cor Chg	Coef Chg.	Var A Chg
AB	3.15	1.09	5.94	-3.22
AC	1.13	2.63	1.14	-2.43
BC	3.6	6.12	6.7	-3.6
Avg.	2.62	3.28	4.59	-3.08

Case	Var B Chg	Original Corr. Coeff.
AB	6.2	0.77
AC	0.1	0.40
BC	-1.89	0.49
Avg.	1.47	0.55

Table 1: Performance of linear formulations (in percent change)

The *ordering heuristic* (Dandekar and Cox 2002) sorts sensitive cells in and assigns their directions in an alternating fashion. It is intended to find good solutions with respect to absolute cell deviation in a computationally efficient manner.

We studied its performance on preserving statistical measures by comparing its performance to that of the exact MILP method and to an optimal (nonlinear) solution. Table 2 reports this comparison. Given the variation in the quality of the solutions on the covariance and variance measures, the performance of the ordering heuristic was good. In fact, the ordering heuristic improved performance on absolute cell deviation. This finding is consistent with the literature, which demonstrates the better performance of the ordering heuristic on cell deviation.

Solution Method	Cov. Chg.	Corr. Chg.	Var A Chg.	Var. B Chg.
Exact	3.15	1.09	5.94	-3.22
Ordering Heuristic	5.34	2.19	4.49	-4.56
Ordering vs. optimal	69	100	-24	41

Solution Method	Absol. Cell Dev.
Exact	8.81e+7
Ordering Heuristic	8.78e+7
Ordering vs. optimal	-0.34

Table 2: Performance of ordering heuristic vs. exact and nonlinear optimal (% chg)

6 Concluding comments

Developments over the past two decades have resulted in a variety of methods for statistical disclosure limitation in tabular data. Among these, controlled tabular adjustment yields the most useable data product, thereby supporting broad opportunities to analyze released data. The question is then how well adjusted data preserve analytical outcomes of original data. Cox and Kelly (2003) addressed this issue in the univariate case. This paper has reported effective linear formulations for preserving key statistics in univariate and multivariate cases.

REFERENCES

Cox, L.H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association* **75**, 377-385.

_____ (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association* **90**, 1153-1162.

_____ (2000). Discussion. *ICES II: The Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions*. Alexandria, VA: American Statistical Association, 905-907.

_____and R.A. Dandekar. (2003). A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. *Proceedings of the 2002 FCSM Statistical Policy Seminar*. Washington, DC: U.S. Office of Management and Budget (in press).

_____ and J.P. Kelly (2004). Balancing data quality and confidentiality for tabular data. **Work Session on Statistical Data Confidentiality, Luxembourg, 7 to 9 April 2003**, Monographs of Official Statistics, Luxembourg: European Communities, 11-23.

_____, J.P. Kelly and R. Patil (2004). Balancing quality and confidentiality for multivariate tabular data. **Lecture Notes in Computer Science 3050**, New York: Springer Verlag, 87-98.

Dandekar, R.A. and L.H. Cox (2002). Synthetic tabular data--an alternative to complementary cell suppression (manuscript).

Fellegi, I.P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association* **67**, 7-18.

Fischetti, M. and J.J. Salazar-Gonzalez (2000). Models and algorithms for optimizing cell suppression in tabular data with linear constraints. *Journal of the American Statistical Association* **95**, 916-928.