Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

What shall we do with the ratios?

Prepared Sarah Giessing, German Federal Statistical Office, Germany

What shall we do with the ratios?

Sarah Giessing[*]

[*] Statistisches Bundesamt, 65180 Wiesbaden, Germany, Sarah.Giessing@destatis.de

Abstract: Traditionally, many statistical agencies protect magnitude tabular establishment data by cell suppression. Typical risk concepts, rules and techniques for cell suppression apply to data of an additive nature, i.e. sums of a quantitative variable. For tables presenting means, ratios or other indicators, it is often considered enough to suppress a cell only, if it relates to just a single unit. Exceptions from this simple rule are made, if a publication also presents tabulations of the enumerator or denominator variable (in case of a ratio), or the cell frequencies (in case of a mean). In those cases, it is obvious that cell suppressions must be consistent to avoid disclosure risks.

This paper presents an idea for how to deal with indicator that are the ratio of two sums, in a context where sums are protected not through cell suppression, but through a perturbative method like stochastic noise.

1 Introduction

Literature offers numerous ideas how to protect statistical data by methods that perturb the data rather than attempting to avoid disclosure risks through partial suppression or reductions of detail. Some are tailored to protect data presented in aggregated format, i.e. tabular data which is the typical data release format for statistical agencies. In this paper we focus on perturbative post-tabular methods, masking the final output data, for example through (random) rounding or noise addition (see f.i. (Fraser and Wooton, 2005), (Giessing, 2011), or non-stochastic manipulation of the data (Castro and Gonzales, 2009)). All those methods typically apply to data of an additive nature, computed as sums of a quantitative (in the case of counts data: dichotomous) variable.

When discussing a statistical disclosure control concept based on a perturbative approach with data managers of a statistical institute or with data users, typically a question like “What happens to time series data and ratios?” arises. Extending the rounding/noise strategy of (Giessing, 2012), the present paper proposes a strategy to handle such data.

The paper is organized in six sections. Sec. 2 briefly recalls the rounding/noise strategy of (Giessing, 2011) which is then extended in sec. 3 to the case of ratios. Some illustration and results derived from an application to real data from German tourism statistics are presented in sec. 4. In section 5 we look at the disclosure risks. The paper finishes with a summary section.

2 Flexible Rounding Based on Post-tabularStochastic Noise

This section briefly recalls the main concept of the methodology proposed in (Giessing, 2011, 2012). That methodology applies to data that are sums of a quantitative variable.

The masking method proposed in (Giessing, 2011) requires a certain minimum deviation between true and masked cell value for sensitive cells. E.g., that the masked cell value is non-sensitive according to the sensitivity rule employed. This can be achieved by masking the largest contribution y_1 of a table cell with original value T^{orig} and replacing it by $y_1{}^{post(T)}$ when computing the perturbed cell value T^{post} : To control the strength of the perturbation, choose parameters μ_0 and σ_0^2. When a cell c is sensitive, multiply the largest contribution y_1 by r_c , where r_c := (1 ± (μ_0+ abs(z_c))) (z_c drawn from a N(0, σ_0^2) distribution[1]). When cell c is non-sensitive, multiply y_1 by (1 ± (abs(z_c))). When the method should correspond to the p% sensitivity rule, choose μ_0:= 2*p/100. Then the noisy value T^{post} of a sensitive true value T^{orig} becomes non-sensitive, i.e. $|T^{post} - y_1 - y_2| \geq p*y_1$, ($y_1$ and y_2 denoting the two largest contributions to T^{orig}) as proven in (Giessing, 2012).
(Giessing, 2012) explains also that (for fixed T^{post} and y_1) the confidence interval
(2.1) [T^{post} -$y_1(\mu_0+\sigma_0\zeta_\gamma)$; T^{post} +$y_1(\mu_0+ \sigma_0\, \zeta_\gamma)$] for T^{orig} has probability $2\gamma - 1$ (ζ_γ denote the γ-quantile of the standard normal distribution). (Giessing, 2012) proposes then to turn this confidence interval into a rounding interval: Compute its width
(2.2) $2y_1(\mu_0+\sigma_0 u_\gamma)$, select the power of 10 (10, 100, 1000, etc) “closest” to this width as rounding base b, and publish T^{post} rounded to the next multiple of b.

3 A Flexible Rounding Strategy for Ratios based on Stochastic Noise

This section elaborates an idea presented in (Chipperfield and Yu, 2011) for a simple special case. (Chipperfield and Yu, 2011) address the standard case for estimating regression coefficients in a regression model. Estimating a ratio can be regarded as a special case: we estimate parameter β_1 of a simple two-variable regression model that assumes the regression line to pass through the origin, e.g. β_0=0. We consider two variables y and x, with n observations for both of them. Following the denotation used in (Chipperfield and Yu, 2011) with some adaptation to our simple special case, this means we consider fitting a regression model with parameter β (e.g. the only parameter of the two-variable regression through the origin) using estimating function $H(\beta)$:
$H(\beta) = \sum_{i=1}^{n}\{y_i - f_i(\beta)\}$, where $f_i(\beta) = E(y_i|x_i) = \beta x_i$. The solution to $H(\beta) = 0$ gives the estimate $\hat{\beta}$.

[1] Note, abs(z_c) are then distributed according to a normal distribution truncated at zero.

In our simple case, solving $\sum_{i=1}^{n}\{y_i - \beta x_i\} = 0$ yields $\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, i.e. the ratio $\hat{\beta} = Y/X$, with enumerator $Y = \sum_{i=1}^{n} y_i$ and denominator $X = \sum_{i=1}^{n} x_i$.

(Chipperfield and Yu, 2011) propose to perturb the estimate as follows: Instead of solving $H(\beta) = 0$ and releasing $\hat{\beta}$, they suggest to solve $H(\beta) = E^*$ and to release the resulting estimator $\hat{\beta}^*$. E^* is the perturbation introduced for disclosure limitation, $E^* = u \cdot e$ where e denote the "maximum effect a record may have on the ratio", computed as $e = \max_i\{|y_i - f_i(\beta)|\}$ and u noise drawn from a suitable random distribution.

In our simple case, we have $e = \max_i\{|y_i - \hat{\beta}x_i|\} = \max_i\left\{\left|y_i - \frac{Y}{X}x_i\right|\right\}$.

(Chipperfield and Yu, 2011) propose a uniform distribution on the range (-1,1) for noise generation. The present paper, however, follows the approach of (Giessing, 2011, 2012) to draw the noise from a normal N(0,σ_0^2) distribution.

Solving $H(\beta) = E^*$ means then that we solve $\sum_{i=1}^{n}\{y_i - \beta x_i\} = u \cdot e$. This yields $\hat{\beta}^* = \frac{(\sum_{i=1}^{n} y_i) - u \cdot e}{\sum_{i=1}^{n} x_i} = \frac{Y}{X} - \frac{u \cdot e}{X} = \hat{\beta} - u\frac{e}{X}$ with noise u drawn from a N(0, σ_β^2) distribution. Then the confidence interval

(3.1) $\left[\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}; \hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right]$ for the original ratio $\hat{\beta}$ has probability $2\gamma - 1$ (ξ_γ denote the γ-quantile of the standard normal distribution).

The strategy suggested here is to fix a suitable parameter σ_β and release $\hat{\beta}^*$ along with a confidence interval covering the true ratio with a 'sensible' probability. On the other hand, in the case of (tables of) ratios published by National Statistical Institutes which are widely used and not only by trained statisticians, it might be better to turn the confidence interval into a rounding interval as suggested in (Giessing, 2012): E.g. compute the width
(3.2) $2\sigma_\beta \xi_\gamma \frac{e}{X}$ of the confidence interval (4.1), select the power of 10 (10, 100, 1000, etc) "closest" to this width as rounding base b, and publish $\hat{\beta}^*$ rounded to the next multiple of b.

The question is now, if this method is 'safe' in a scenario, where the agency also releases rounded, noisy estimates for enumerator and denominator? Or does it thwart the disclosure control applied to the enumerator or denominator data? We will consider this question in section 5, after illustrating the method with some toy data in the following section.

4 Illustrative Example and Test Result

This section first illustrates the method proposed using a few lines of toy example data, and then reports some preliminary results obtained with a real data set from German Tourism statistics.

4.1 Illustrative Example

Imagine two variables Y and X. Our toy dataset consists of data for three non-sensitive cells. Table 1 displays the instance and the estimates users are left to derive on their own, if no special estimates like those of sec. 3 are offered. Table 2 presents the methodology of section 3, applied to the same data.

For all three cells of the instance, we assume a value for the denominator variable X=1000 with largest contribution x_1=300. The three lines of table 1 refer to those three cells. They show the cell value of the enumerator Y with its largest contribution (y_1) and the (true) ratio *Y/X*. The next columns display the noisy cell values Y* and X* computed according to (2.1). In all three cases u =1*0.05 is the absolute of the noise "drawn" from $\mathrm{N}(0, \sigma_0^2)$ with σ_0^2=0,05 and the deviation sense is negative. Table 1 also displays the ratios obtained as ratios of the perturbed data Y*/X*, confidence interval bounds according to (2.2) for enumerator and denominator (lb_Y, ub_Y, lb_X, ub_X), each obtained at level ξ_Y=1, and finally the width of a confidence interval for the true ratio, obtained as $ub_Y/\,lb_X - lb_Y/ub_X$.

Y	y_1	*Y/X*	*Y**	*X**	*Y*/X**	ub_{Y^*}	lb_{Y^*}	ub_{X^*}	lb_{X^*}	$\frac{ub_{Y^*}}{lb_{X^*}} - \frac{lb_{Y^*}}{ub_{X^*}}$
20	7	0.0200	19.65	985	0.0199	20	19.30	1030	1000	0.00126
200	70	0.2000	196.5	985	0.1995	200	193.00	1030	1000	0.01262
2000	700	2.00	1965	985	1.99	2000	1930.00	1030	1000	0.12621

Table 1 Illustrative example: the case of no special estimates

Table 2 extends the instance by data made up for e , e.g. the "maximum effect a record may have on the ratio". The table consists of three blocks. Each block basically refers to the same instance, e.g. the same three cells, but with different values for the $\mathrm{N}(0, \sigma_\beta^2)$ noise (σ_β^2=0,05). Those three table 2 variants of the instance refer to negative deviation senses too. For every line of the instance, table 2 displays the estimate $\hat{\beta}^*$ and its confidence interval width obtained according to (3.2) as $2\sigma_\beta \xi_Y \frac{e}{X}$, at $\xi_Y = 1$. This width is then used to round the data, e.g. to eventually reduce the level of precision at which the data is displayed:

For our three cells, assume the initial intention is to display the first two cells with d=4 places after the decimal separator (equivalent to a format in % with 2 decimal places), and the third cell with d=2 places after the decimal separator. For the rounding step, we compute $10^d \hat{\beta}^*$ and $b(= 10^{round(log_{10}(2\sigma_\beta \frac{e}{X} 10^d))})$. The final column shows round($\hat{\beta}^*$), e.g. $10^d \hat{\beta}^*$ rounded to the next multiple of 10^b and divided afterwards by 10^d. Digits to be replaced by a zero due to the rounding are replaced by an asterisk here.

e	$\hat{\boldsymbol{\beta}}^*$	$10^d\hat{\boldsymbol{\beta}}^*$	$2\sigma_\beta\frac{e}{X}$	***b***	***round($\hat{\boldsymbol{\beta}}^*$)***
***u*=1*0.05=0.05**					
3	0.01985	198.5	0.0003	1	0.0199
30	0.19850	1985	0.003	10	0.199*
300	1.98500	19850	0.03	100	1.99**
u =0.5*0.05=0.025					
3	0.01993	199.25	0.0003	1	0.0199
30	0.19925	1992.5	0.003	10	0.199*
300	1.99250	19925	0.03	100	1.99**
u =2*0.05=0.1					
3	0.01970	197	0.0003	1	0.0197
30	0.19700	1970	0.003	10	0.197*
300	1.97000	19700	0.03	100	1.97**

Table 2 Illustrative example: methodology of sec. 3.

Comparing the confidence intervals of the two tables (last col. of table 1 vs. 4th col. of table 2), we find that the methodology of sec. 3 yields better results. The point estimates ***Y*/X**** for the ratio in table 1 are slightly closer to the true value (compared to those displayed in col. $\hat{\boldsymbol{\beta}}^*$ in the first three rows of table 2). But note, by changing the deviation sense for the denominator to the opposite of that of the enumerator, we would get the following, considerably poorer estimates: 0.01936, 0.19360 and 1.93596.

4.2 Test Results

The methodology has been implemented for two test tables from German tourism statistics, each of them with results for about 400 table cells (most of them at municipality level). The ratio of the first test table is a change rate between two periods, taking values between 0.1 and 7. The ratio of the second test table is the capacity utilisation, e.g. the rate of nights spent as ratio to the number of beds offered by the establishments. It takes values between 0.9 % and 80 %. For both tables we observe promising results. With the data of the first table, for more than 97% of the table cells the relative deviation between true and perturbed ratio is below 2%. In the second table there are about 97% of table cells with a relative deviation below 0.2%.

With one exception (observed at an extremely large value of the ratio), in the first table all confidence intervals obtained for the $\hat{\beta}^*$estimate are smaller compared to those a user should assume for the “simple” estimate Y^*/X^* (e.g. $\left[\frac{lb_{Y^*}}{ub_{X^*}};\frac{ub_{Y^*}}{lb_{X^*}}\right]$ in the denotation of 4.1.). For the second table (of “small” ratios below 1), for about 7% of the cells the confidence interval of the simple estimate is the smaller one. So it seems the method tends to perform the better, the larger the ratios tend to be, as might be expected.

5 Addressing the Disclosure Risk Issue

If, denominator and enumerator of a ratio, X or Y, are protected by a perturbative method, the ratio Y/X must not be published in a way that yields disclosure risk for the true, unperturbed values of enumerator or denominator – even if both are non-sensitive aggregates. Otherwise, this might create secondary disclosure risks for sensitive cells in tabulations of the enumerator or denominator variable.

This section studies the following two disclosure risk scenarios: A user/intruder computes an interval for the enumerator by using the interval released for the ratio and the interval released for its denominator. The second scenario is the ‘reverse’ case: A user/intruder computes an interval for the denominator by using the interval released for the ratio and the interval released for its enumerator. We consider the release of the rounded ratio ‘thwarting’ the protection of enumerator or denominator, when the interval obtained by the intruder in this way tends to shrink compared to the respective rounding/confidence interval released by the agency. In 5.1 and 5.2 we show that this will not happen, if enumerator and denominator are non-sensitive cells, and if the agency uses suitable σ_β parameters for drawing the noise for the perturbation of the ratio.

For sake of simplicity, we assume all data to be positive and we will look at the problems in terms of the confidence intervals, ignoring the final rounding step of the procedures of sections 2 and 3.

5.1 Disclosure Risk for the Enumerator

With probability $2\gamma - 1$ the true ratio $\hat{\beta}$ is contained in the confidence interval (3.1), and X and Y are contained in intervals computed according to (2.1). For a fixed γ, these interval bounds are released and hence known to the user/intruder[2]. So we have

(5.1) $\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X} \leq \frac{Y}{X} \leq \hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}$

and (after change of denotation, replacing e.g. T by X, T^{post} by X^*, y_1 by x_1), and using non-sensitivity of X (hence μ_0=0)

(5.2) $X^* - \sigma_X \xi_\gamma x_1 \leq X \leq X^* + \sigma_\beta \xi_\gamma x_1$,

and (after another change of denotation, replacing e.g. T by Y, T^{post} by Y^*), and using non-sensitivity of Y (hence μ_0=0)

(5.3) $Y^* - \sigma_0 \xi_\gamma y_1 \leq Y \leq Y^* + \sigma_0 \xi_\gamma y_1$.

[2] As mentioned above, we ignore for now the rounding step of the procedure, assuming the exact confidence intervals to be released.

Obviously (5.1) is equivalent to $X\hat{\beta}^* - \sigma_\beta \xi_\gamma e \leq Y \leq X\hat{\beta}^* + \sigma_\beta \xi_\gamma e$. Using (5.2) it follows then $(X^* - \sigma_0 \xi_\gamma x_1)\hat{\beta}^* - \sigma_\beta \xi_\gamma e \leq Y \leq (X^* + \sigma_0 \xi_\gamma x_1)\hat{\beta}^* + \sigma_\beta \xi_\gamma e$, with a probability of at least $(2\gamma - 1)^2$ (or maybe larger, if the perturbations of *X* and $\frac{Y}{X}$ are not independent). The interval given by these bounds is the interval we regard to be obtained by the intruder in our scenario.

In the appendix (A.1), we show that it is always possible to choose σ_β (e.g. require: $\sigma_\beta \geq \frac{\sigma_0}{(1-\xi_\gamma \sigma_0)}$) such that the width of that interval, $(X^* + \sigma_0 \xi_\gamma x_1)\hat{\beta}^* + \sigma_\beta \xi_\gamma e - [(X^* - \sigma_0 \xi_\gamma x_1)\hat{\beta}^* - \sigma_\beta \xi_\gamma e]$ is at least the width of the interval given by (5.3) , i.e. $2\sigma_0 \xi_\gamma y_1$.

This means that the new interval for *Y* that can be obtained after release of the interval for $\frac{Y}{X}$ does not tend to be smaller than the interval directly released for *Y*. Note, because the estimates $\hat{\beta}^*$ and $Y^*/X^*$are usually different, this does not mean that the new interval always covers the interval directly released for *Y* – the intervals may overlap only partially and occasionally the intersection may be small. On the other hand, both intervals are merely confidence intervals with a probability unknown to the user, so this kind of residual disclosure risk should be considered tolerable.

5.2 Disclosure Risk for the Enumerator

Like in 5.2, we have that with probability 2 γ – 1 the true ratio $\hat{\beta}$ is contained in the confidence interval (3.1), and *X* and *Y* are contained in intervals computed according to (2.1). For a fixed γ , these interval bounds are released and hence known to the user/intruder.

Obviously (5.1) implies the following inequalities for X: $\frac{Y}{\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}} \leq X \leq \frac{Y}{\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}}$ which is equivalent to $\frac{Y\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{Y}{X} e}{\left(\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}\right)} \leq X \leq \frac{Y\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{Y}{X} e}{\left(\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}\right)}$. Using (5.3) it follows then $\frac{(Y^* - \sigma_0 \xi_\gamma x_1)\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{Y}{X} e}{\left(\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}\right)} \leq X \leq \frac{(Y^* + \sigma_0 \xi_\gamma x_1)\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{Y}{X} e}{\left(\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}\right)}$.

Like in 5.1, the interval given by these bounds is the new interval for *X* we regard to be obtained by the intruder in our scenario. In the appendix, A.2, we show that for larger levels of confidence it is possible to choose σ_β (for a start, require σ_β to exceed $\frac{\sigma_0}{(1-2\xi_\gamma \sigma_0)}$)) such that its width,

$\frac{(Y^*+\sigma_0\xi_\gamma x_1)\widehat{\beta}^*+\sigma_\beta\xi_\gamma\frac{Y}{X}e}{\left(\widehat{\beta}^*+\sigma_\beta\xi_\gamma\frac{e}{X}\right)\left(\widehat{\beta}^*-\sigma_\beta\xi_\gamma\frac{e}{X}\right)} - \frac{(Y^*-\sigma_0\xi_\gamma x_1)\widehat{\beta}^*-\sigma_\beta\xi_\gamma\frac{Y}{X}e}{\left(\widehat{\beta}^*+\sigma_\beta\xi_\gamma\frac{e}{X}\right)\left(\widehat{\beta}^*-\sigma_\beta\xi_\gamma\frac{e}{X}\right)}$, exceeds in most cases the width of the interval given by (5.2) , e.g. $2\sigma_0\xi_\gamma x_1$.

Like in 5.1, this means that new interval for X that can be obtained after release of the interval for $\frac{Y}{X}$, tends not to be smaller as the interval directly released for X.

6 Summary and Final Comments

In the context of the post-tabular noise based on the disclosure control concept of (Giessing, 2012) and building on an idea of (Chipperfield and Yu, 2011), this paper has proposed a flexible rounding strategy for ratios also based on stochastic noise. Encouraging first test results are reported, observed with real life data from German Tourism Statistics. However, testing of the methodology is still work in progress, especially to prove the concepts suggested for establishing suitable parameters for the method. Another important issue for future work is to test, or eventually extend the method for special, important types of ratio data, like means, i.e. a magnitude variable divided by a count.

The main contribution of the paper is a theoretical proof that the method proposed here is consistent with the approach of (Giessing, 2011) in the following sense: confidence intervals released according to the concept for the ratio data do not tend to inflict a substantial disclosure risk for respective enumerator and denominator data, if these are non-sensitive. Note that, otherwise, disclosure risk for such non-sensitive data might in turn inflict a (secondary) disclosure risk for cells with sensitive enumerator or denominator cell values because of additive relations between table cells.

Acknowledgement

This work has been supported by the FP7-INFRASTRUCTURES-2010-1 project “DwB-Data without Boundaries”, number 262608. This support is appreciated.

References

Castro, J., Gonzalez, J.A., and Baena, D. (2009). *User's and programmer's manual of the RCTA package*, Technical Report DR 2009-01, Dept. of Statistics and Operations Research, Universitat Politecnica de Catalunya.

Chipperfield, J., Yu, F. (2011). *Protecting Confidentiality in a Remote Analysis Server for Tabulation and Analysis of Data*, paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Tarragona, 2-4 December 2009) available at
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/52_Australia.pdf

Fraser, B., Wooton, J. (2006). *A proposed method for confidentialising tabular output to protect against differencing*, in Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302

Giessing, S. (2011). Post-tabular Stochastic Noise to Protect Skewed Business Data, paper presented at the *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Tarragona, 2-4 December 2009)* available at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/47_Giessing.pdf

Giessing, S. (2012). *Flexible Rounding Based on Consistent Post-tabular Stochastic Noise.* In J. Domingo-Ferrer and I. Tinnirello (Eds.), Privacy in Statistical Databases, 22-34. New York: Springer-Verlag. LNCS 7556

Appendix

A.1 In the following, we prove that it is always possible to choose σ_β such that

$(A.1.1)\left(X^* + \sigma_0\xi_\gamma x_1\right)\hat{\beta}^* + \sigma_\beta\xi_\gamma e - \left[\left(X^* - \sigma_0\xi_\gamma x_1\right)\hat{\beta}^* - \sigma_\beta\xi_\gamma e\right] \geq 2\sigma_0\xi_\gamma y_1$.

Obviously, $(A.1.1)$ is equal to

$2\xi_\gamma\left(\sigma_0 x_1\hat{\beta}^* + \sigma_\beta e\right) \geq 2\sigma_0\xi_\gamma y_1$ and hence to $\sigma_0 x_1\hat{\beta}^* + \sigma_\beta e \geq \sigma_0 y_1$. After subtracting $\sigma_0\frac{Y}{X}x_1$ from both sides of this inequality we see that it is equal to $\sigma_0 x_1\left(\hat{\beta}^* - \frac{Y}{X}\right) + \sigma_\beta e \geq \sigma_0\left(y_1 - \frac{Y}{X}x_1\right)$. Because of (3.1), the distance between true and perturbed ratio, $\hat{\beta}^* - \frac{Y}{X}$, is at most $\sigma_\beta\xi_\gamma\frac{e}{X}$. Hence, that the left hand side of the inequality is at least $\sigma_\beta e\left(1 - \xi_\gamma\frac{\sigma_0 x_1}{X}\right)$. This means that it suffices to prove $\sigma_\beta e\left(1 - \xi_\gamma\frac{\sigma_0 x_1}{X}\right) \geq \sigma_0\left(y_1 - \frac{Y}{X}x_1\right)$. By its definition (see sec. 3) $e \geq y_1 - \frac{Y}{X}x_1$. So it is enough to prove $\sigma_\beta e\left(1 - \xi_\gamma\frac{\sigma_0 x_1}{X}\right) \geq \sigma_0 e$. Obviously, this will hold, if $\sigma_\beta \geq \frac{\sigma_0}{\left(1-\xi_\gamma\frac{\sigma_0 x_1}{X}\right)}$. As $\frac{x_1}{X} \leq 1$ it is enough to require $\sigma_\beta \geq \frac{\sigma_0}{(1-\xi_\gamma\sigma_0)}$.

A.2 In this section, we work out how to choose σ_β such that

$(A.2.1)\ \ \left(Y^* + \sigma_0\xi_\gamma y_1\right)\hat{\beta}^* + \sigma_\beta\xi_\gamma\frac{Y}{X}e - \left(\left(Y^* - \sigma_0\xi_\gamma y_1\right)\hat{\beta}^* - \sigma_\beta\xi_\gamma\frac{Y}{X}e\right) \geq 2\sigma_0\xi_\gamma x_1\left(\hat{\beta}^* + \sigma_\beta\xi_\gamma\frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta\xi_\gamma\frac{e}{X}\right)$.

Recall the definition of $\hat{\beta}^*$ from sec. 3, e.g. $\hat{\beta}^* = \frac{Y}{X} - u\frac{e}{X}$. We substitute $-u = r\sigma_\beta\xi_\gamma$, requiring $0 \leq r \leq 1$. Then straightforward arithmetic (left to the reader) shows that $\left(\hat{\beta}^* + \sigma_\beta \xi_\gamma \frac{e}{X}\right)\left(\hat{\beta}^* - \sigma_\beta \xi_\gamma \frac{e}{X}\right)$ can be written as $\frac{Y}{X}\left(\frac{Y}{X} + \sigma_\beta \xi_\gamma \frac{e}{X}[2r - \varepsilon_1]\right)$ where $\varepsilon_1 =: \sigma_\beta \xi_\gamma \frac{e}{Y}(1 - r^2)$, $0 \le \varepsilon_1 \le \sigma_\beta \xi_\gamma \frac{e}{Y}$.

(A.2.1) can then be written as

(A.2.2) $2\xi_\gamma\left(\sigma_0 y_1 \hat{\beta}^* + \sigma_\beta \frac{Y}{X} e\right) \ge 2\sigma_0 \xi_\gamma x_1 \frac{Y}{X}\left(\frac{Y}{X} + \sigma_\beta \xi_\gamma \frac{e}{X}[2r - \varepsilon_1]\right)$. After dividing this inequality by (-2ξ_γ) and addition of $\sigma_0 \frac{Y}{X} y_1$ to both sides, we see that it is equal to

$\sigma_0 y_1\left(\frac{Y}{X} - \hat{\beta}^*\right) - \sigma_\beta \frac{Y}{X} e \le \sigma_0 \frac{Y}{X} y_1 - \sigma_0 x_1 \frac{Y}{X}\left(\frac{Y}{X} + \sigma_\beta \xi_\gamma \frac{e}{X}[2r - \varepsilon_1]\right)$ which is equivalent to

(A.2.3) $\sigma_0 y_1\left(\frac{Y}{X} - \hat{\beta}^*\right) - \sigma_\beta \frac{Y}{X} e \le \sigma_0 \frac{Y}{X}\left(\frac{e}{e}\left(y_1 - x_1 \frac{Y}{X}\right) - e\sigma_\beta \xi_\gamma \frac{x_1}{X}[2r - \varepsilon_1]\right)$.

Like above, we substitute $-u = r\sigma_\beta \xi_\gamma$ in the definition of $\hat{\beta}^* = \frac{Y}{X} - u\frac{e}{X}$. Then the difference between true and perturbed ratio, $\frac{Y}{X} - \hat{\beta}^*$, is $-r\sigma_\beta \xi_\gamma \frac{e}{X} = -r\sigma_\beta \xi_\gamma \frac{e}{Y}\frac{Y}{X}$ and hence we find for the left hand side of the inequality: $\sigma_0 y_1\left(\frac{Y}{X} - \hat{\beta}^*\right) - \sigma_\beta \frac{Y}{X} e = \sigma_0 y_1\left(-r\sigma_\beta \xi_\gamma \frac{e}{Y}\frac{Y}{X}\right) - \sigma_\beta \frac{Y}{X} e = -\sigma_\beta \frac{Y}{X} e\left(r\xi_\gamma \frac{\sigma_0 y_1}{Y} + 1\right)$. The term in brackets can be written as $1 + \varepsilon_2$, where $-1 \ll \varepsilon_2 \ll 1$, because usually parameters are chosen such that $\sigma_0 \xi_\gamma \ll 1$ and of course $\frac{y_1}{Y} \le 1$.

Dividing now both sides of (A.2.3) by -$e\frac{Y}{X}$ we see that it is equivalent to

$\sigma_\beta(1 + \varepsilon_2) \ge \sigma_0\left(\sigma_\beta \xi_\gamma \frac{x_1}{X}[2r - \varepsilon_1] - \frac{\left(y_1 - x_1\frac{Y}{X}\right)}{e}\right)$ which is the same as

(A.2.4) $\sigma_\beta \ge \frac{\sigma_0}{(1+\varepsilon_2)}\left(\sigma_\beta \xi_\gamma \frac{x_1}{X}[2r - \varepsilon_1] - \frac{\left(y_1 - x_1\frac{Y}{X}\right)}{e}\right)$

Because of the definition of e (in sec. 3) $\left|\frac{\left(y_1 - x_1\frac{Y}{X}\right)}{e}\right| \le 1$. On the other hand, common choices of parameters yield $\sigma_\beta \xi_\gamma \ll 1$. As $\frac{x_1}{X} \le 1$, the absolute of the left side of the difference in the brackets will be considerably smaller than its right side. Thus, obviously, (A.2.4) imposes the most critical requirement on σ_β, if $y_1 - x_1\frac{Y}{X}$ is negative and r is positive. An initial choice might be $\sigma_\beta =: \frac{\sigma_0}{1 - 2\xi_\gamma \sigma_0}$. It should then be confirmed, if this choice fits the data, e.g. if (A.2.4) holds "generally". Otherwise, slightly increase σ_β until (A.2.4) is satisfied for all cells (of a sufficiently large test data set).