The Effect of Data Swapping Procedures on Regression Estimates – Evidence from a Simulation Study*

Tamás Bartus
PhD, associate professor
Corvinus University
of Budapest
E-mail: tamas.bartus@uni-corvinus.hu

Data swapping is a technique of statistical disclosure control. Although it has several desirable properties, it biases covariances and linear regression coefficients.

The study examines by simulation the extent of biases caused by various data swapping techniques, using anonymized Labour Force Survey (LFS) data of the Hungarian Central Statistical Office (HCSO). It is found that the relative bias associated with data swapping may be held under 10%, and the bias can be minimized by selecting the pairs to be swapped randomly and by manipulating several explanatory variables simultaneously. Interpretation of the results is based on the theory of measurement errors.

KEYWORDS:
Microdata protection.
Data swapping.
Statistical disclosure control.

* The paper is a shortened version of the study written for „Making labour market forecasts and foreseeing structural changes" (TÁMOP-2.3.2-09/1-2009-0001) priority project and published by the *Hungarian Statistical Review* (Vol. 91. No. 5. pp. 465–497.) in Hungarian. The author thanks *Gergely Daróczi* for his excellent research assistance and *Zsombor Cseres-Gergely* for providing data and presenting suggestions. The helpful comments of *Katalin Szép*, *Zoltán Vereczkei*, *László Antal*, *Máté Dobány* and *Beáta Nagy* on the earlier version of the paper are also appreciated.

In recent decades, several disclosure control techniques have been developed to protect the identity of individuals and organizations that provide data for statistical offices and surveys (*Hundepool et al.* [2010]). Most of the publications on disclosure control are concerned with procedures to protect identities effectively and to minimize disclosure risk. Although indicators measuring information loss are available (*Domingo-Ferrer–Torra* [2001], *Hundepool et al.* [2010]), we have only scant knowledge about the implications of data protection for the reliability of statistical estimates. It is well known that disclosure control techniques affect means, standard deviations and even covariances. Hence, disclosure control comes at the real cost of biasing estimates and thereby constraining the interest of researchers (*Boudreau* [2005]).

This study examines the impact of data swapping on covariance and regression estimates. Data swapping was chosen because it has two desirable properties. First, it can be applied to both continuous and categorial variables. Second, data swapping does not affect the means and standard deviations of manipulated variables. In contrast, alternative techniques like micro-aggregation, noise addition can be applied to either categorial or continuous variables, and these alternatives do not leave standard deviations unaffected.

We have scant knowledge about the impact of data swapping on regression coefficients. Even if there are studies on estimating covariances from protected data (*Kim* [1990], *Gouweleeuw et al*. [1998]), it is not known whether these results apply to multiple regression. So far, simulation has been used to examine the bias of regression coefficients only for micro-aggregation (*Liu–Little* [2003]; *Lenz et al.* [2006]; *Schmid* [2006]; *Schmid–Schneeweiss* [2005], [2008], [2009]; *Schmid–Schneeweiss–Kuchenhoff* [2007]).

In this paper, simulation is used to assess the bias of multiple regression coefficients that arises when researchers analyse protected microdata. Multiple linear regression coefficients are obtained by multiplying the inverse of the variance, covariance matrix of explanatory variables with covariances between dependent and independent variables. The matrix algebra makes it difficult to assess the extent to which regression estimates are biased if some of the explanatory variables are protected with data swapping. The analytical intractability in the context of data protection is similar to that of measurmenet error: if any of the explanatory variables is measured with error, the coefficients of all predictors become biased, the magnitude of which is difficult to predict (*Fuller* [1987]). Simulation is a general method of studying problems that are difficult to solve analytically.

In the first chapter of the paper, the effect of data swapping on covariance and regression estimates is discussed. The second and third chapters present the description of and the results from a simulation study, while the final chapter reaches conclusion.

1. The effect of data swapping on regression estimates

In this section, we discuss the effect of data swapping on covariance and regression estimates. Our main objective is to examine the implications of data swapping for biases in estimated multiple regression coefficients. Since multiple regression coefficients depend on covariances (and variances), we begin the exposition with the effect of data swapping on covariance estimates.

1.1. Bias when estimating covariances

Data swapping exchanges values of sensitive variables among individual records (*Dalenius–Reiss* [1982]). The procedure and its outcome can be formally defined as follows. Let $\delta_{xij} = 1,$ if the i^{th} value of variable *x* is replaced with the j^{th} value and vice versa; otherwise, $\delta_{xij} = 0.$ The outcome of the procedure is the anonimized variable x^a that is defined as

$$x_i^a = \left(1 - \delta_{xij}\right) x_i + \delta_{xij} x_j ,$$

$$x_j^a = \left(1 - \delta_{xij}\right) x_j + \delta_{xij} x_i .$$

(See *Boudreau* [2005].) Although the definition makes no explicit reference to the disclosure risks of individuals *i* and *j*, it is trivial that either *i* or *j* is an individual with high disclosure risk. This is due to the fact that the objective of the procedure is to protect the identity of an otherwise easily identifiable individual. The selection procedure of individuals is also left open in the former definition. For instance, the exchange partners might be selected randomly, as it is the case during post-randomization (*Kooimann–Willenborg–Gouweleeuw* [1997], *Gouweleeuw et al.* [1998]).

Data swapping keeps means and variances intact; however, it modifies the joint distribution of variables. *Dalenius* and *Reiss* [1982] suggested that data swapping should be an iterative procedure that ends when the original joint distribution is restored. In practice, there is no guarantee that the procedure will be sucessfully terminated; therefore, current practice favours techniques that aim to approximate the original joint distribution as closely as possible (*Reiss* [1984], *Shlomo–Tudor–Groom* [2010]).

The fact that data swapping affects the joint distribution of variables has two important consequences. First, weighted estimates are biased since data swapping does not apply to weight variables. The general formulas are quite complicated (*Boudreau* [2005]) but the intuition is quite simple. After exchanging the values x_j and x_k, the difference in the weighted means between anonimized and original variables is

$$\frac{w_j\left(x_k - x_j\right) + w_k\left(x_j - x_k\right)}{\sum w_i}. \qquad /1/$$

The second consequence is the bias of (unweighted) covariances. Since data swapping is usually applied to categorical variables, the covariance between the indicator variable x and another variable y is considered. Suppose that data swapping pertains to pn pairs of observations, where n is the sample size and $0 < p < 0.5$. Exchanging the values of x has the same consequence as exchanging the values of y. Let y_{01} denote the values of y in the subsample in which the zero values of x were changed to ones. Similarly, y_{10} denotes the values of y in the subsample in which x was changed from 1 to 0. It is easy to show that the bias in covariance estimates introduced by data swapping is

$$Cov\left(x^0, y\right) - Cov\left(x, y\right) = -p\left[\overline{y}_{10} - \overline{y}_{01}\right]. \qquad /2/$$

If the expected value of y is higher in the group $x = 1$, then $\overline{y}_{10} > \overline{y}_{01}$ and the right side of the equation /2/ is negative. In the opposite case, the right side is positive. In short, data swapping attenuates the absolute value of the original covariance. The degree of this bias increases with the relative frequency of data swaps.

Since the covariance between x and an either continuous or dummy y is the product of the variance of x and the difference $\overline{y}_1 - \overline{y}_0$, the covariance between the anonimized indicator variable x^a and y can be expressed as follows:

$$Cov\left(x^a, y\right) = Cov\left(x, y\right)\left[1 - \frac{p}{Var\left(x\right)}\frac{\overline{y}_{10} - \overline{y}_{01}}{\overline{y}_1 - \overline{y}_0}\right]. \qquad /3/$$

It is useful to denote the expression within the brackets as

$$Q_x\left(y\right)=1-\frac{p}{Var\left(x\right)}\frac{\overline{y}_{10}-\overline{y}_{01}}{\overline{y}_1-\overline{y}_0}\ . \qquad /4/$$

If exchange partners are selected randomly so that $\overline{y}_{10}-\overline{y}_{01}=\overline{y}_1-\overline{y}_0$, $Q_x\left(y\right)$ would simplify into

$$Q_x\left(y\right)=1-\frac{p}{Var\left(x\right)}\ . \qquad /5/$$

Formula /5/ can be easily interpreted: the larger the number of pairs affected by data swapping, the larger the bias in covariance estimates. The extent of bias depends on the variance of the variable to be protected. If the parameter *p* is known, "true" covariance can be estimated using the formula

$$\widehat{Cov\left(x,y\right)}=\frac{Cov\left(x^a,y\right)}{Q_x\left(y\right)}\ . \qquad /6/$$

1.2. Bias when estimating bivariate regression coefficients

Consider the linear regression model with a single explanatory variable *x*. The estimated coefficient of *x* is the ratio of the covariance between the dependent variable *y* and *x* to the variance of *x*. Equation /6/ implies that the estimated coefficient as computed using the protected dataset equals the "true" coefficient (computed from the original dataset) multiplied by $Q_x\left(y\right)$. This is similar to the well-known result for the attenuation bias arising from measurement errors. Suppose that x^a is not the anonymized variable but a variable measured with error; that is, x^a is the sum of the original variable *x* and the random measurement error *u*. Indeed, one of the methods of data protection, noise addition minimizes disclosure risks by adding a random error (*Shlomo* [2010]). The link between the bias arising from anonimization and that arising from measurement error is explicit. Random measurement error of the covariate is known to attenuate the regression coefficient (*Fuller* [1987]):

$$\hat{\beta}=\frac{Cov\left(y,x^a\right)}{Var\left(x^a\right)}=\frac{Cov\left(x,y\right)}{Var\left(x\right)+Var\left(u\right)}=\beta\frac{Var\left(x\right)}{Var\left(x\right)+Var\left(u\right)}=\beta R_x\,, \qquad /7/$$

where R_x is the coefficient of reliability.

Given the similarities of equations /6/ and /7/, Q is labeled "the relative coefficient of reliability". The term relative implies that Q assumes the presence of another variable with which the covariance is calculated. The use of the term coefficient of reliability is justified by the fact that both random measurement errors and data swapping attenuate regression coefficients. Imagine that data swapping is a procedure that adds a random measurement error u to the original variable. Given the identity $R = Q$, the variance of the imaginary measurement error can be defined as:

$$Var(u) = \frac{pVar(x)}{Var(x) - p}.$$

To ensure that this variance is nonnegative, the inequality $p \le \bar{x} - (\bar{x})^2$ should hold.

1.3. Bias when estimating multiple regression coefficients

The effect of data swapping on multiple regression coefficients is difficult to predict. Multiple linear regression coefficients are obtained by multiplying the inverse of the variance-covariance matrix of explanatory variables with the covariances between dependent and independent variables. The matrix algebra makes it difficult to assess the extent to which regression estimates are biased if some of the explanatory variables are protected with data swapping.

Consider the simple multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon .$$

In the original (not anonymized) dataset, the coefficients are calculated as

$$\beta_1 = \frac{\dfrac{Cov(yx_1)}{Var(x_1)} - \dfrac{Cov(yx_2)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - \dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}},$$

$$\beta_2 = \frac{\dfrac{Cov(yx_2)}{Var(x_2)} - \dfrac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - \dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}.$$

Suppose that x_1 is protected by data swapping. Given the results in the previous subsections, the coefficients should be calculated as follows:

$$\hat{\beta}_1 = \frac{Q_1(y)\dfrac{Cov(yx_2)}{Var(x_2)} - Q_1(x_2)\dfrac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - Q_1(x_2)\dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}, \qquad /8/$$

$$\hat{\beta}_2 = \frac{\dfrac{Cov(yx_2)}{Var(x_2)} - Q_1(y)Q_1(x_2)\dfrac{Cov(yx_1)Cov(x_1x_2)}{Var(x_1)Var(x_2)}}{1 - Q_1(x_2)\dfrac{Cov^2(x_1x_2)}{Var(x_1)Var(x_2)}}.$$

Data swapping affects both coefficients. The sign and magnitude of bias are difficult to predict because the bias depends not only on Q but also on other variances and covariances. This difficulty is similar to that of predicting the attenuation of coefficients when one of the predictors is measured with error (*Fuller* [1987]). For this reason, the sign and magnitude of bias is examined using simulation.

1.4. Stratified and restricted sampling methods

Suppose that in the sample $x = 1$, some of the cases are at the risk of disclosure. These cases constitute what we are labelling "risk set". Data swapping requires the selection of cases from the subsample $x = 0$. The subsample of selected cases can be called "donor set". Equation /3/ implies that the attenuation bias of the covariance between any y and x depends on the difference in the means of y in the risk and donor sets.

From a statistical point of view, a good data swapping procedure minimizes difference in the means of y in the risk and donor sets. An obvious sampling method, which is close to post-randomization, is the random selection of cases from both sets. The donor set chosen should be relatively similar to the risk set.

Similarity can be achieved both indirectly and directly. The indirect method consists of stratifying the sample using additional variables and then selecting exchange partners from the same strata. If the stratifying variable(s) correlate(s) with the y variable, the difference in the means of y in the risk and the donor sets should be smaller than in the case of random selection. This procedure is often called "targeted data swapping" (*Shlomo–Tudor–Groom* [2010]). However, we prefer to use the term "stratified data swapping".

Restricted data swapping is a direct method of achieving similarity. Consider the following example. A survey data includes information on educational attainment and the type of settlement where the respondent lives. Suppose that most highly educated people live in urban areas and only a few people live in villages. For this reason, the variables education and type of settlement together enable intruders to identify highly educated villagers. To protect the anonymity of these individuals, highly educated villagers are changed to villagers with lower education and urban residents with lower education are changed to urban residents with higher education. Note that lower education here means the three educational attainment levels that are lower than college degree: upper secondary education (which is completed with A-level examination and allows graduates to apply for college and university education), lower secondary education (which does not allow the opportunity of passing the A-level exam), and primary education. Out of these educational categories, the upper secondary level resembles most to higher education. For this reason, the attenuation bias can be minimized by exchanging higher education with upper secondary education. In other words, the donor set is restricted to a certain subsample of the potential donors. Regardless of stratifying the sample or not, the restricted procedure selects exchange partners that are similar with regard to the variable partially responsible for high disclosure risk.

2. The simulation study

Our study simulates the situation of a researcher who is interested in estimating the effect of education and settlement type on wages, using linear regression. He/she would like to use data from a large-scale survey that were collected by the HCSO and include the variables of interest as well as control variables such as age and gender. There is no access to the original, unanomyzed dataset because the combinations of education and type of settlement would enable intruders to identify individuals. Therefore the statistical office provides an anonymized dataset for researchers. Our objective is to examine the extent of bias that arises from using the protected dataset instead of the original one.

Data from the HCSO Labour Force Survey conducted in the first quarter of 2011 are used. The dataset is anonymized following procedures unknown to the public (at the time of conducting the study, we had no permission to access the original dataset). Therefore we pretend as if it were the original one. The dataset includes socio-economic information on 47 162 individuals, at the time of the survey, 23 783 people were employed among them. Only data on these employed individuals are used in the study. Education is measured by means of three indicator variables: lower secondary education (*EDU*$_2$), upper secondary education (*EDU*$_3$) and college degree (*EDU*$_4$). Type of settlement (*SETTLEMENT*) is a categorical variable with four categories: 1 = Budapest, the capital city, 2 = city with county rights, 3 = smaller town, and 4 = village. Gender is an indicator variable which equals one if the respondent is male.

The dependent variable, our imaginary researcher wishes to use, is the natural log of wages. Since the Labour Force Survey includes no information on wages, the log wage variable is created as follows:

$$\begin{aligned} log\ \text{wage} = 9{,}67 + 0{,}1EDU_2 + 0{,}2EDU_3 + 0{,}6EDU_4 - 0{,}1\left(SETTLEMENT - 1\right) + \\ +0{,}5AGE - 0{,}0002AGE^2 + 0{,}2GENDER + e, \end{aligned} \qquad /9/$$

where e is a standard normal random variable. The coefficients are taken from *Kertesi–Köllő* [2002]. Since the standard deviation of the residual is unity, the coefficient of determination (R^2) is about 25%.

Labour Force Survey data together with simulated log wage data are considered the original dataset that allows intruders to disclose the identity of some of the respondents. Several data swapping procedures are applied to create a protected dataset. The bias of estimates is assessed by comparing the estimates computed from protected data with those computed from original data.

We apply 16 distinct methods of data swapping. All of them assume that a certain proportion of villagers with college level are at the risk of disclosure. The procedures differ in three dimensions.

> *1*. There are four methods of selecting variables to protect the anonymity of highly educated villagers: *a)* only the education variable is manipulated; *b)* only the settlement variable is manipulated; *c)* both education and settlement variables are manipulated; *d)* either education or settlement variable is manipulated, and the selection of the variable to be manipulated is random, with equal probability.
>
> *2*. The donor set is selected either with or without stratification. In the first case, strata are formed by combining the two gender categories with five (16–25, 26–35, 36–45, 46–55, and 56–65) age group categories. Hence, the number of strata is 10.
>
> *3*. The selection of donors is either restricted or not restricted. For the educational variable, the donor set includes people with upper secondary education. For the place of residence variable, the donor set includes the residents of smaller towns.

For each of these procedures, we assume that p percent of highly educated villagers can be disclosed. During simulations, p takes the values 10%, 25% and 50%. According to the Labour Force Survey data, about 6% of the employed fall in the category of highly educated villagers. For each combination of the 16 procedures and the values of p, there are 1000 replications of the study. Each replication consists of generating wage data according to equation /9/, resulting in an "original dataset", protecting data with data swapping.

During each replication, we compute estimates both from the original and the protected datasets and compare them in terms of relative bias. The relative bias of a particular statistics s is given by

$$\text{relative bias of } s = 100\frac{\sum_{r=1}^{R} s_r - RS}{RS}, \qquad /10/$$

where R is the number of replications, S_r is the value of the statistics computed from the protected sample in replication r and S is the parameter computed from the original sample. For instance, equation /5/ implies that the relative bias of covariance estimates arising from data swapping is

$$\frac{\hat{\beta}-\beta}{\beta} = -\frac{p}{Var(x)}. \qquad /11/$$

In multiple regressions, there is no similar simple formula due to the complexity of equation /8/.

3. Results

Relative bias of covariance estimates. Since regression coefficients depend on covariances (and variances), it is useful to begin with examining the effect of data swapping on covariance estimates. Table 1 displays the relative biases of the estimates of covariance between simulated log wage and higher education.

The results support the expectation that relative bias is directly proportional to the rate of cases affected by data swapping. Bias can be kept at a low level if data swapping affects 10% or 25% of the cases with high disclosure risk. The relative biases are about 2.5 times larger in the column $p = 25\%$ than in the column

$p = 10\%$ and the bias is two times larger in the column $p = 50\%$ than in the column $p = 25\%$.

In our study, we use four different sampling methods for selecting donors: random selection, restricted random selection, stratified selection, and restricted and stratified selection. Out of the four methods, random selection of exchang partners minimizes bias. Surprisingly, neither stratification nor targeting improves the precision of the estimates computed from protected datasets. In other words, the donor set should be selected completely at random. Of course, it is also true for datasets that are similar to that of the labour force survey used in this study.

Bias of univariate regression coefficients. Our imaginary researcher wishes to estimate differences in log earnings by educational attainment and place of residence. He/she is especially interested in estimating the wage advantage of college graduation and the wage disadvantage associated with rural residence. He/she uses upper secondary education and residence in small towns as respective reference categories.

Biases of the univariate regression coefficients of higher education and village residence are shown in Tables 2 and 3. Again, bias is proportional to the parameter *p*. Regardless of the size of *p*, it is always possible to find a procedure which keeps the relative bias of the regression coefficients under 10%. Bias is the smallest if *1.* data swapping is shared between the variables higher education and rural residence, and *2*. exchange partners are selected randomly. The first result is not surprising since the distribution of the data manipulation burden among two variables is similar to the reduction of the proportion of data swaps. However, it is surprising that neither stratification nor targeting performs better than random selection.

Table 1

Relative biases of the estimates of covariance between log wage and college degree by various data swapping procedures

Method	$p = 10\%$		$p = 25\%$		$p = 50\%$	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Only education is manipulated						
random selection	–1.600	0.906	–4.000	1.365	–8.192	1.799
restricted random selection	–1.924	1.016	–4.764	1.469	–9.622	1.902
stratified selection	–2.142	0.770	–5.413	1.159	–11.043	1.596
restricted and stratified selection	–2.089	0.889	–4.991	1.341	–10.188	1.703
Only settlement type is manipulated	0	0	0	0	0	0

(Continued on the next page.)

(Continuation.)

Method	$p = 10\%$		$p = 25\%$		$p = 50\%$	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Both education and settlement type are manipulated						
random selection	–1.598	0.892	–4.035	1.385	–8.227	1.831
restricted random selection	–1.916	1.022	–4.748	1.521	–9.476	1.863
stratified selection	–2.138	0.764	–5.407	1.210	–10.970	1.610
restricted and stratified selection	–2.081	0.871	–5.045	1.299	–10.168	1.580
Either education or settlement type is manipulated						
random selection	–0.829	0.647	–1.963	1.026	–3.979	1.343
restricted random selection	–0.941	0.726	–2.445	1.092	–4.775	1.528
stratified selection	–1.060	0.562	–2.693	0.866	–5.354	1.204
restricted and stratified selection	–1.051	0.617	–2.488	0.942	–5.130	1.273

Note. The true value of the covariance is 0.08. The anonymization of the settlement type cannot bias the covariance.

Table 2

Relative biases of the univariate regression estimate of higher education compared with upper secondary education by various data swapping procedures

Method	$p = 10\%$		$p = 25\%$		$p = 50\%$	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Only education is manipulated						
random selection	–1.686	0.926	–4.174	1.483	–8.306	1.931
restricted random selection	–2.590	1.318	–6.378	2.018	–12.720	2.559
stratified selection	–2.158	0.849	–5.377	1.278	–10.844	1.737
restricted and stratified selection	–2.632	1.146	–6.797	1.748	–13.522	2.099
Only settlement type is manipulated	0	0	0	0	0	0
Both education and settlement type are manipulated						
random selection	–1.628	0.960	–4.203	1.483	–8.384	1.941
restricted random selection	–2.558	1.327	–6.315	1.965	–12.601	2.600
stratified selection	–2.176	0.852	–5.322	1.310	–10.864	1.707
restricted and stratified selection	–2.647	1.174	–6.832	1.786	–13.642	2.217

(Continued on the next page.)

(Continuation.)

Method	$p = 10\%$		$p = 25\%$		$p = 50\%$	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Either education or settlement type is manipulated						
random selection	–0.837	0.705	–2.087	1.068	–4.161	1.501
restricted random selection	–1.322	0.968	–3.139	1.477	–6.252	1.999
stratified selection	–1.095	0.593	–2.685	0.971	–5.313	1.289
restricted and stratified selection	–1.298	0.827	–3.450	1.279	–6.859	1.761

Note. The estimated regression coefficient of higher education, as computed from the original dataset, is 0.464. Anonymizing the settlement type cannot bias this coefficient.

Table 3

Relative biases of the univariate regression estimate of living in a smaller town compared with living in a village, by various data swapping procedures

Method	$p = 10\%$		$p = 25\%$		$p = 50\%$	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Only education is manipulated	0	0	0	0	0	0
Only settlement type is manipulated						
random selection	0.903	1.667	2.182	2.625	4.295	3.497
restricted random selection	2.911	2.597	7.172	3.831	14.065	5.015
stratified selection	1.689	1.450	4.043	2.303	8.120	3.099
restricted and stratified selection	4.576	2.325	11.639	3.593	22.922	4.496
Both education and settlement type are manipulated						
random selection	0.781	1.636	2.224	2.587	4.215	3.416
restricted random selection	2.894	2.624	7.257	3.902	14.088	5.124
stratified selection	1.611	1.484	4.046	2.336	8.012	3.112
restricted and stratified selection	4.533	2.380	11.527	3.530	22.851	4.613
Either education or settlement type is manipulated						
random selection	0.427	1.137	1.083	1.873	2.019	2.568
restricted random selection	1.412	1.894	3.763	2.946	7.154	3.971
stratified selection	0.821	1.036	1.997	1.663	4.112	2.326
restricted and stratified selection	2.267	1.671	5.770	2.573	11.460	3.547

Note. The estimated regression coefficient of smaller towns, as computed from the original dataset, is 0.161. The manipulation of education cannot bias this coefficient.

Table 4

Relative biases of the multiple regression estimate of higher education compared with upper secondary education, by various data swapping procedures

Method	*p* = 10%		*p* = 25%		*p* = 50%	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Only education is manipulated						
random selection	–3.081	0.993	–7.647	1.592	–15.190	2.075
restricted random selection	–4.117	1.351	–10.052	2.026	–19.796	2.685
stratified selection	–3.021	1.011	–7.504	1.524	–15.098	2.100
restricted and stratified selection	–3.922	1.350	–9.969	2.068	–19.642	2.552
Only settlement type is manipulated						
random selection	–0.440	0.105	–1.072	0.185	–2.038	0.301
restricted random selection	–0.587	0.078	–1.350	0.174	–2.308	0.367
stratified selection	–0.442	0.113	–1.064	0.192	–2.023	0.287
restricted and stratified selection	–0.580	0.080	–1.336	0.181	–2.271	0.369
Both education and settlement type are manipulated						
random selection	–3.049	1.042	–7.740	1.648	–15.489	2.103
restricted random selection	–3.800	1.379	–9.432	0.200	–18.642	2.664
stratified selection	–3.052	1.028	–7.506	1.622	–15.257	2.111
restricted and stratified selection	–3.661	1.402	–9.371	2.147	–18.686	2.686
Either education or settlement type is manipulated						
random selection	–1.752	0.730	–4.350	1.134	–8.576	1.571
restricted random selection	–2.398	0.978	–5.688	1.546	–10.975	2.068
stratified selection	–1.747	0.705	–4.270	1.157	–8.391	1.559
restricted and stratified selection	–2.233	0.970	–5.703	1.519	–11.090	2.099

Note. The true value of the coefficient of higher education is 0.4.

Table 5

Relative biases of the multiple regression estimate of living in a larger town compared with living in a village, by various data swapping procedures

Method	*p* = 10%		*p* = 25%		*p* = 50%	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Only education is manipulated						
random selection	–4.611	0.694	–11.002	1.131	–20.423	1.649
restricted random selection	–5.836	0.575	–13.539	0.984	–23.925	1.574
stratified selection	–4.506	0.713	–10.828	1.099	–20.194	1.649
restricted and stratified selection	–5.906	0.566	–13.593	1.003	–24.024	1.487
Only settlement type is manipulated						
random selection	–2.490	2.665	–6.229	4.145	–12.746	5.701
restricted random selection	–4.637	4.119	–11.528	6.049	–23.129	8.183
stratified selection	–2.437	2.554	–6.480	4.205	–13.083	5.489
restricted and stratified selection	–4.916	4.146	–11.635	6.479	–23.443	8.330
Both education and settlement type are manipulated						
random selection	–1.161	2.814	–1.646	4.513	–1.132	6.004
restricted random selection	3.965	4.359	10.386	6.481	21.544	8.476
stratified selection	–1.120	2.841	–2.081	4.556	–1.939	5.951
restricted and stratified selection	3.556	4.400	9.884	6.501	21.085	8.516
Either education or settlement type is manipulated						
random selection	–3.571	1.917	–8.617	3.074	–16.178	4.235
restricted random selection	–5.249	2.959	–12.026	4.755	–22.414	6.460
stratified selection	–3.488	1.921	–8.455	3.043	–15.947	4.325
restricted and stratified selection	–5.344	3.068	–12.420	4.770	–22.749	6.456

Note. The true value of the coefficient of higher education is 0.089.

Bias of multiple regression coefficients. The imaginary researcher now moves on to estimate the coeffients in the equation /9/ using multiple linear regression. His/her interest is centered on estimating the returns of higher education compared with upper secondary education and those of living in a smaller town compared with living in a village. The relative biases arising from data swapping are shown in Tables 4 and 5.

When examining the bias of univariate regression coefficients, it was found that *1*. random selection of exchange partners minimizes the bias, *2*. data swapping is shared between the variables higher education and rural residence, and *3*. relative bias of the regression coefficients is under 10%. In the context of multiple linear regression, only the first of these results holds. The second result is only true for the estimated wage advantage of college graduates. In contrast, the bias associated with the coefficient of living in a smaller town is kept at the lowest level if both education and settlement are protected. Finally, the 10% limit is often exceeded, thus the bias of multiple regression coefficients is larger than that of the univariate ones.

4. Conclusions

Data swapping is a technique of statistical disclosure control with several desirable properties. It can be applied to both continuous and categorial variables; and it does not affect the means and standard deviations of manipulated variables. However, data swapping biases covariances and linear regression coefficients. In multiple regression models, the extent of bias is difficult to predict. In this paper, we present the results of a simulation study in order to assess the effects of various techniques of data swapping on regression coefficients. Using a modified version of the HCSO Labour Force Survey, the relative biases of the estimates of the returns to higher education and the wage disadvantages associated with rural residence were studied.

It was found that the random selection of exchange partners minimizes the bias of regression coefficients. The random selection of observations resembles closely the post-randomization method of disclosure control. In fact, in the context of a simulation study, data swapping and post-randomization are the same since the values to be exchanged are picked up randomly. This result implies that sophisticated procedures including stratification and targeting might not pay off if the precision of estimates is considered. It was also found that it is useful to manipulate several variables at the same time since it might help to reduce the risk of disclosure.

In most of the cases, relative bias can be kept under 10%. Is it negligible or significant? The bias arising from protecting microdata can be considered as a form of non-sampling error (*Sarndal–Swensson–Wretman* [1992], *Biemer–Lyberg* [2003]). The link between disclosure control and measurement error is natural. For instance, noise addition is all about introducing measurement error in order to minimize the disclosure risk (*Shlomo* [2010]). The paper shows that the bias arising from data swapping is similar to the attenuation bias arising from measurement errors. In the context of univariate regression, the only minor difference is that measurement errors increase the variance of explanatory variables, while data swapping decreases the covariance between dependent and explanatory variables. Nevertheless, both data swapping and random measurement errors of regressors attenuate regression coefficients (*Fuller* [1987]). In short, data swapping is a measurement error if its consequences are considered. It is believed that the bias arising from data swapping is smaller than that arising from measurement errors during data collection.

Future research should address our finding that the bias arising from data swapping is relatively small, especially if exchange partners are selected randomly. In this study, data of a large and representative social survey were used; for other datasets, however, different results may be obtained. Future research should assess systematically the statistical implications of the procedures that rely on stratification and targeting.

References

BIEMER, P. P. – LYBERG, L. E. [2003]: *Introduction to Survey Quality*. John Wiley & Sons. Hoboken.

BOUDREAU, J.-R. [2005]: *Data Swapping is not the Panacea.* Proceedings of Statistics Canada's Symposium "Methodological Challenges for Future Information Needs". 25–28 October. Statistics Canada. Ottawa.

DALENIUS, T. – REISS, S. P. [1982]: Data Swapping. A Technique for Disclosure Control. *Journal of Statistical Planning and Inference.* Vol. 6. No. 1. pp. 73–85.

DOMINGO-FERRER, J. – TORRA, V. [2001]: Disclosure Control Methods and Information Loss for Microdata. In: *Doyle, P. – Lane, J. – Theeuwes, J. – Zayatz, L.* (eds.): *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North–Holland. Amsterdam. pp. 93–112.

FULLER, W. A. [1987]: *Measurement Error Models*. John Wiley & Sons. New York.

GOUWELEEUW, J. – KOOIMAN, P. – WILLENBORG, L. C. R. J. – DE WOLF, P.-P. [1998]: The Post Randomization Method for Protecting Microdata. *Qüestiió.* Vol. 22. No. 1. pp. 145–156.

HUNDEPOOL, A. – DOMINGO-FERRER, J. – FRANCONI, J. – GIESSING, S. – LENZ, R. – NAYLOR, J. – SCHULTE NORDHOLT, E. – SERI, G. – DE WOLF, P.-P. [2010]: *Handbook on Statistical Disclosure Control*. ESSNet SDC. http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

KERTESI, G. – KÖLLŐ, J. [2002]: Economic Transformation and the Revaluation of Human Capital – Hungary, 1986–1999. In: *de Grip, A. – Van Loo, J. – Mayhew, K.* (eds.): *The Economics of Skills Obsolescence*: *Theoretical Innovations and Empirical Applications*. Elsevier. Amsterdam. pp. 235–273.

KIM, J. J. [1990]: *Subpopulation Estimation for the Masked Data*. Proceedings of the American Statistical Association Section on Survey Research Methods. Alexandria. pp. 303–308.

KOOIMAN, P. – WILLENBORG, L. C. R. J. – GOUWELEEUW, J. M. [1997]: *PRAM: A Method for Disclosure Limitation of Microdata*. Research paper 9705. Statistics Netherlands. Voorburg, Heerlen.

LENZ, R. – ROSEMANN, M. – VORGRIMLER, D. – STURM, R. [2006]: *Anonymising Business Micro Data – Results of a German Project.* Statistisches Bundesamt. Berlin.

LIU, F. – LITTLE, R. J. A. [2003]: *SMIKe vs. Data Swapping and PRAM for Statistical Disclosure Control in Microdata: A Simulated Study*. Proceedings of the American Statistical Association Survey Research Methods Section. http://www.amstat.org/sections/srms/proceedings/

REISS, S. P. [1984]: Practical Data Swapping: The First Steps. *ACM Transactions on Database Systems*. Vol. 9. No. 1. pp. 20–37.

SARNDAL, C. E. – SWENSSON, B. – WRETMAN, J. [1992]: *Model Assisted Survey Sampling*. Springer. New York.

SCHMID, M. [2006]: Estimation of a Linear Model under Microaggregation by Individual Ranking. *Allgemeines Statistisches Archiv*. Vol. 90. No. 3. pp. 419–438.

SCHMID, M. – SCHNEEWEISS, H. [2005]: The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study. In: *Pohlmeier, W. – Ronning, G. – Wagner, J.* (eds): *Econometrics of Anonymized Micro Data*. Jahrbücher für Nationalökonomie und Statistik. No. 225. Lucius and Lucius. Stuttgart. pp. 529–543.

SCHMID, M. – SCHNEEWEISS, H. – KUCHENHOFF, H. [2007]: Estimation of a Linear Regression under Microaggregation with the Response Variable as a Sorting Variable. *Statistica Neerlandica*. Vol. 61. No. 4. pp. 407–431.

SCHMID, M. – SCHNEEWEISS, H. [2008]: *Estimation of a Linear Model in Transformed Variables under Microaggregation by Individual Ranking*. University of Munich. Munich.

SCHMID, M. – SCHNEEWEISS, H. [2009]: The Effect of Microaggregation by Individual Ranking on the Estimation of Moments. *Journal of Econometrics*. Vol. 153. No. 2. pp. 174–182.

SHLOMO, N. [2010]: *Measurement Error and Statistical Disclosure Control*. S3RI Methodology Working Papers, M10/05. Southampton Statistical Sciences Research Institute, University of Southampton. Southampton.

SHLOMO, N. – TUDOR, C. – GROOM, P. [2010]: *Data Swapping for Protecting Census Tables*. S3RI Methodology Working Papers, M10/06. Southampton Statistical Sciences Research Institute, University of Southampton. Southampton.