Discussion Paper No. 03-55

Microdata Disclosure by Resampling – Empirical Findings for Business Survey Data

Sandra Gottschalk

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Non–technical Summary

In recent years empirical researchers' demand for releasing business survey data has dynamically increased. Statistical offices and offical or private research institutes are asked to pass on their data for scientific use to the scientific community. German law provides that individual datafiles are only allowed to pass on for scientific use and if disclosure limitation is guaranteed in effect. In case of personal data, confidentiality is easier to achieve than it is in that of firm data. Methods to avoid re-identification of individual enterprises have to be developed, which are simultaneously suitable to preserve as much information as possible.

Traditional methods to avoid disclosure often destroy the structure of data, i.e., information loss is potentially high. Therefore, I discuss an alternative technique of creating anonymized datasets. The procedure creates datasets - the resample - which should have the same characteristics as the original survey data and hinder re–identification as they only consist of synthetic values. Some applications of this method with (a) simulated data and (b) innovation survey data, the Mannheim Innovation Panel (MIP), are presented in comparison to a common method of disclosure control, disturbance with multiplicative error.

The experiments show that univariate distributions can be better reproduced by unweighted resampling. Linear regression results can be reproduced quite well if the resampling procedure implements directional information of the data in form of the correlation structure. If multiplicative disturbance is controlled for in the estimation approach, parameter estimates will also remain. With regard to disclosure avoidance anonymized data with multiplicative perturbed variables better performs on the average. Even though resamples consist of synthetic values, confidentiality problems remain.

Discussion Paper No. 03-55

Microdata Disclosure by Resampling – Empirical Findings for Business Survey Data

Sandra Gottschalk

Microdata Disclosure by Resampling - Empirical Findings for Business Survey Data

by
Sandra Gottschalk

Centre for European Economic Research (ZEW)

September 1, 2003

Abstract:

A problem statistical offices and research institutes are faced with by releasing micro-data is the preservation of confidentiality. Traditional methods to avoid disclosure often destroy the structure of data, i.e., information loss is potentially high. In this paper I discuss an alternative technique of creating scientific-use-files, which reproduce the characteristics of the original data quite well. It is based on Fienberg's (1997 and 1994) [5], [6] idea to estimate and resample from the empirical multivariate cumulative distribution function of the data in order to get synthetic data. The procedure creates datasets - the resample - which have the same characteristics as the original survey data. In this paper I present some applications of this method with (a) simulated data and (b) innovation survey data, the Mannheim Innovation Panel (MIP), and compare resampling with a common method of disclosure control, i.e. disturbance with multiplicative error, concerning confidentiality on the one hand and the appropriateness of the disturbed data for different kinds of analyses on the other. The results show that univariate distributions can be better reproduced by unweighted resampling. Parameter estimates can be reproduced quite well if (a) the resampling procedure implements the correlation structure of the original data as a scale and (b) the data is multiplicative perturbed and a correction term is used. On average, anonymized data with multiplicative perturbed values better protect against re–identification as the various resampling methods used.

Keywords: resampling, multiplicative data perturbation, Monte Carlo studies, business survey data

JEL Classification: C13, C15, C81

1 Introduction

Empirical research in economic and social science requires information about households and firms, which is collected by statistical offices and public or private research institutes in form of microdata. As computer capability and availability of statistical software increased in recent years, empirical analyses and thus demand for microdata have been advancing dynamically. German law provides that microdata from government statistics are allowed to be passed on for scientific use only and if disclosure limitation is in effect[1] guaranteed. The same holds for survey data assembled by private or public research institutes, if confidentiality is promised to the respondents. Hence, a problem statistical offices and research institutes are faced with by releasing micro-data is the preservation of confidentiality. Even business survey data are at risk because disclosure is more likely than for personal data as additional information are easier obtainable and population size is smaller (see e.g. Brand, 2000 [2]). Traditional methods to avoid disclosure often destroy the structure of data, i.e., information loss is potentially high and the potential for empirical analyses decreases (see Rosemann, 2003 [15]).

In this paper I discuss an alternative technique of creating scientific-use-files[2], resampling, which generates a synthetic microdata file with nearly the same characteristics as the original survey data. It is based on Fienberg's (1997 und 1994) [5], [6] idea to estimate and resample from the empirical multivariate cumulative distribution function of data. As elements of the resample are only replicates and do not necessarily correspond to any individuals in the original sample survey, an identification of the true values should not be possible. Nevertheless one cannot rule out the possibility of disclosure, as synthetic datasets could be very similar to real characteristics of observations. Especially, extreme values are at risk.

The paper is structured as follows: in section 2 I describe the idea of resampling and an easily constructed algorithm to create synthetic data, attributed to Devroye and Györfi (1985) [4] and Silverman (1986) [18]. Subsequently, applications with simulated data (Section 3) and business innovation survey data (Section 4, see also Appendix B) point out the properties of resamples. Confidentiality and applicability are examined. In a second step I compare resampling in particular with a common method of disclosure control, i.e. disturbance with multiplicative error (see e.g. Hwang, 1986 [10]), concerning confidentiality on the one hand and the appropriateness of the disturbed data for different kinds of analyses on the other.

2 The Idea of Resampling

In order to generate synthetic data with the same characteristics as an original survey data file, one has to estimate the density function of the variables in the data and then sample from it. In practice, it will be impossible to achieve this with complete accuracy, as full information about the true density of data is not available. One could apply a parametric approach in assuming a theoretical density function with unknown parameters, such as the

[1] Disclosure should not be possible without unusually high costs and waste of time and energy.

[2] In contrast to public-use-files, which should be totally anonymized, i.e. disclosure is not possible under no circumstances. Scientific-use-files guarantee only disclosure limitation in effect (see above), therefore such files are still exploitable for scientific use.

normal distribution. The parameters would have to be estimated with the data, e.g. means and variances. Then sampling from a theoretical distribution function can be quite easily done. However, in reality survey data will rarely follow a specific theoretical distribution.

Fienberg (1997) [5] proposes non-parametric and semi-parametric estimation methods, such as kernel density estimators or a Bayesian approach (see also Fienberg et al., 1996 [7]). The estimated cumulative distribution function will differ to some degree from the real distribution of the data. Most of the survey data, official statistics as well as surveys from research institutes, undercover the real population. Therefore, the sample distribution is merely an estimation of reality. Another source of bias is introduced by measurement errors. Furthermore, techniques to estimate multivariate cumulative distribution functions have only been used for low-dimensional data yet. Even three–dimensional relations are difficult to describe and only if the sample size is large enough (e.g. the software package XploRe, which is described in Härdle et al., 1991 [9]). One possibility to improve the estimation is to use a Bayesian method: one has to estimate the empirical distribution function and generate the full posterior distribution (dependent distribution). This approach takes into account regression-like relationships within the sample. "It provides a way of formalising the process of learning from data to update beliefs in accord with recent notions of knowledge synthesis" (Congdon, 2001, 1 [3]). In the following, a sample is drawn from the posterior distribution. Fienberg et al. (1996) [7] propose to sample from it using Rubin's multiple imputation technique (Rubin, 1993 [16] , 1987 [17]), which includes Bootstrap sampling.

Devroye and Györfi (1985) [4] and Silverman (1986) [18], who deal with nonparametric density estimation and sampling from density estimates, show how to draw from density functions without need to estimate it explicitly. The procedure can be used to create samples that have the underlying characteristics and structure of the real data. However, spurious details that have arisen from random effects are oppressed. The algorithm for the univariate case is described in the following: suppose a continuous variable $X = X_1, X_2, ..., X_n$ (n observations) and a kernel density function K with bandwidth h. The bandwidth specifies the halfwidth of the kernel, the width of the density window around each point.

1. Draw observations X_Z of the data file X with replacement.
2. Compute k to have probability density function K.
3. Generate $Z = X_Z + hk$.

The kernel function can be simulated from an Epanechnikov kernel, for example[3]:

$$K(x) \quad = \quad \frac{3}{4}(1 - x^2) \quad \text{for} \quad |x| \leq 1 \tag{1}$$

A simple procedure to simulate from the rescaled Epanechnikov kernel is given by Devroye and Györfi (1985)[4]:

1. Compute three univariate random numbers ZV_1, ZV_2, ZV_3 within $[-1, 1]$.

[3] One can also think of the normal density.

2. Generate $k = ZV_2$, if $|ZV_3| \geq |ZV_2|$ and $|ZV_3| \geq |ZV_1|$, otherwise $k = ZV_3$.

The procedure resamples with replacement[4] from the data and disturbs the information in such a manner that the distribution of each variable is retained. The sample size of Z has to be large enough to approximate the distribution of the original data X.

In the literature the choice of the smoothing parameter, the bandwidth, of the kernel has been discussed frequently (see e.g. Parzen, 1962 [14], Tapia and Thompson, 1978 [20]). The appropriate choice for the smoothing parameter will always be influenced by the purpose for which the density estimate is to be used. An optimal bandwidth minimizes the mean square error between the real and the estimated kernel density. It unfortunately depends on the (unknown) density. A meaningful approach is to choose a bandwidth with reference to a standard family of densities. Hence, Silverman (1986) [18] obtains a bandwidth which minimizes the mean integrated square error, if the data were Gaussian and a Gaussian kernel were used, and therefore is not optimal in any global sense:

$$\begin{aligned} h &= \frac{0.9m}{n^{1/5}} \qquad (2) \\ m &= \min\left[\sqrt{\sigma_x^2}, \frac{quantile(p75)_x - quantile(p25)_x}{1.349}\right]. \end{aligned}$$

According to the confidentiality problem, the choice of bandwidth h of the used kernel function is rather difficult because it influences the goodness of fit to the original distribution on the one hand and the probability of disclosure on the other. A narrow bandwidth causes a better approximation of distributions but raises the probability of re-identification as the resampled values - though synthetic - could be very similar to the original. One should also consider if a data intruder might be interested in disturbed values.

A higher–dimensional version of the algorithm can be constructed by using directional information in the data, such as the co–variance–matrix of X. Therefore, the multivariate distribution can nearly be performed. Hence, Devroye and Györfi (1985) [4] modify the third step of the algorithm for I-dimensions:

$$\begin{aligned} Z &= X_Z + h\kappa A \qquad (3) \\ VCV^{-1} &= A'A \\ \kappa &= [k_1 \quad k_2 \quad \ldots \quad k_I]. \end{aligned}$$

VCV is the co–variance–matrix of the original variables $X_1, X_2, ..., X_I$, which is used as weight of the different kernels κ. To get first and second moment properties the same as those of the data the procedure can be transformed (Silverman, 1986 [18]):

$$Z = \bar{X} + (X_Z - \bar{X} + hk)/(1 + h^2\sigma_k^2/\sigma_x^2)^{1/2}, \qquad (4)$$

where $\bar{X}$ is the sample mean of X and σ_x^2 and σ_k^2 the variances of X respectively k. These corrections prevent an overestimation of variances. In the multivariate case Silverman (1986) [18] proposes to scale the kernel to have the same variance matrix as the data.

[4] Resampling with replacement increases confidentiality as some of the initial observations appear several times in the anonymized data. Hence, extreme values rarely arise.

Devroye and Györfi (1985) [4] provide some modified versions of the above algorithms for simulating from density estimations of various kinds, e.g. for variables which concentrate their masses on an intervall, e.g. positive numbers.

The procedure presented above is only applicable for continuous variables. As most of the survey data contain discrete variables, too, one has to find additional masking methods while confidentiality problems concerning these exist. Especially, regional information and classifications of economic sectors could be meaningfully used for re–identification of individual firms. Another problem arise when using resamples for analyses. One has to consider that several calculations are not possible. Non–linear functions of variables cannot be estimated by their resampled pendants, because $f(g(x)) \neq g(f(x))$. For example, the estimated R&D-intensity, the ratio of resampled R&D-expenditure (Z_1) and sales (Z_2), is biased in general:

$$\begin{aligned} E\left[\frac{Z_1}{Z_2}\right] &= E\left[\frac{X_{Z_1} + h_1 k_1}{X_{Z_2} + h_2 k_2}\right] \\ &\neq E\left[\frac{X_{Z_1}}{X_{Z_2}}\right] \end{aligned}$$

But data providers are able to compute several combinations or functions of variables, which data users need for their analyses, before the resampling procedure and add it to the resample.

3 Simulations

To demonstrate the effects of the resampling procedure Monte Carlo simulations are very useful as regularities can be revealed (see e.g. Robert and Casella, 2002 [13]). Here, the simulated data contains 2000 observations, respectively, and the procedure is repeated 100 times.[5] Six variables are simulated by drawing from four theoretical distribution functions: 1. normal, 2. the logarithm of 1, 3. exponential and 4. chi-square distribution. The fifth variable is a linear combination of the first and fourth variable and number six (Y) is a linear combination of the 1. (X_1), 3. (X_2), 4. (X_3) and an error term (u). Hence, a linear regression model is constructed (see below).

In order to find an optimal way of constructing a resample, eight different kinds of resamples are constructed, which differ concerning the bandwidth of the kernel and the usage of the co–variance- or correlationmatrix of the unmasked data as weights. The interaction between confidentiality and data quality can be examined.

1. A1: the resample is constructed as described above but step three is replaced by equation 4 to better reproduce the variances of the variables. Different bandwidths according to 2 are used for each variable. It is expected that the bandwidth leads to a good approximation of the kernel density but therefore to a higher disclosure risk as well. As the chosen bandwidth is only optimal for variables with normal distributions, the error will increase in the case of that variables in the datasets which follow different distributions.

[5] Earlier experiments have shown that results will not remarkably change anymore when the number of repetitions is raised.

2. A2: the bandwidth of A1 is multiplied by factor 1.5. When increasing the window width of the kernel the expected values of the resample get a higher variability. Though distortion strengthen. Hence, the density estimation will not be as accurate as in case A1, but confidentiality will increase.
3. B1: is the same as A1, additionally the kernels are weighted with the co–variance–matrix as shown in equation 4.
4. B2: Is the same as A2, additionally the kernels are weighted with the co–variance–matrix.
5. C1: A1 is computed but the kernel is scaled to have the same co–variance matrix as the data (see Appendix A).
6. C2: C1 is generated with the same bandwidth as in A2.
7. D1: is similar to C1 though using the correlation matrix. The idea is to reproduce the correlation structure of the data (see Appendix A). Hence, improvement of regression results is expected.
8. D2: D1 is generated with the same bandwidth as in A2.

For comparison, an anonymized version of the data is constructed by multiplying each variable with random numbers from univarate distribution functions within the intervall [0.5;1.5]. Hence, means remain the same, but variances and co–variances become biased.

To get a measure of how much confidentiality is provided by the masking techniques one should consider the situation which a potential data intruder is faced with: The intruder has additional information about individual firms, the so called additional database, at his disposal and these data contain several variables which are also included in the anonymized microdata file (common variables), which he wants to disclose. With the number of common variables disclosure risk increases (see Spruill, 1983 [19]). To get an upper bound of re–identification risk the anonymized microdata in these simulations are matched with the original data. Confidentiality can then be defined as follows (see e.g. Spruill, 1983 [19]):

1. For all elements in the original data, find the observation in the anonymized file that minimizes the sum of absolute deviations for all common variables (I chose three variables). The elements in the anonymized file can be matched several times to the original observations. Therefore, some observations of the resamples can possibly not be assigned.
2. If the observation that is found in 1. is the same as the one the masked file is based upon and only differs 20% from the original[6], a link is made.
3. The confidentiality criteria is then defined as the percentage of observations for which such a link cannot be made.

[6] If the values differ more than 20%, I presume that confidentiality is still satisfied as uncertainty of a re–identification is too high.

This proceeding should be distinguished from an estimation of the re–identification probability, which - in addition to the applied anonymization techniques - should take into account (a) the probability that observations of the anonymized microdata file are also involved in the additional database used for disclosure, and (b) the possibility of measurement errors in both data files (see e.g. Brand, 2000 [2] for a discussion of the re–identification risk of business survey data).

Table 1: Monte Carlo Simulations - Confidentiality Measure (CM)

	Resample A1	Resample A2	Resample B1	Resample B2	
CM	73.4%	81.4%	73.4%	80.6%	
	Resample C1	Resample C2	Resample D1	Resample D2	Multiplicative Error
CM	96.6%	98.3%	88.3%	93.5%	97.2%

In Table 1 confidentiality measures of the different kinds of resamples and a dataset of masked variables with multiplicative errors are shown. As expected, confidentiality increases if bandwidths of the kernels are multiplied. The possibility of re-identification is less than 30% regarding each method. The resampling procedure C2 and data perturbation with multiplicative errors perform the best with shares of identified observations of only 2 to 3%.

Kernel density estimates of the raw versus anonymized variables are shown in Appendix E to visualize the effects of the different anonymization procedures. When an Epanechnikov kernel is used, the normal distribution is reproduced quite well by each resampling method. In comparison, the perturbed variable norm_mp has a smaller variance than the original. Resampling procedures C and D, where the kernels are scaled to have the same co–variance or correlation structure as the original data, failed to retain highly skewed distributions even if the bandwidths were multiplied. In contrast, methods A and B fit fairly precisely. Altogether, multiplicative perturbations satisfactorily reproduce univariate distributions.

Tables 2 and 3 show the quantitative extent of information loss the different perturbation procedures add to the data. Deterioration is measured by the average absolute deviation from the original measuring unit, respectively (see Appendix B). Means are better reproduced by multiplicative perturbation but the method distorts the variance of variables by around 12% - resampling on the average by 7%. The co–variance–matrix is biased if the variables are multiplied with errors (15%). Resampling seriously distorts co–variances (about 35-42% bias), even if the kernel–matrix is scaled to have the co–variance–matrix of the original data. Correlations and rank–correlations of masked data remarkably do not differ more than 3-4% from the original in most cases, except C1 and C2. Resampling D1 and D2 best performs the correlation structure. Even a wider window in computing the kernel-matrix does not additionally destroy multivariate relationships and multivariate distributions are only little more biased when using a wider kernel-window. In general, the application of multiplied bandwidths does not worsen performance in a notable way. Deterioration of variances even decreases.

Table 2: Monte Carlo Simulations - Average Absolute Deviation in %

Method	Means	Variances	Correlations	Rank-Correlations	Co–variances
Resample A1	4.98	7.45	2.94	3.17	36.28
Resample A2	4.99	7.42	3.11	3.65	36.50
Resample B1	4.99	7.42	3.03	3.33	36.23
Resample B2	5.00	7.38	3.39	4.06	36.92
Resample C1	4.72	6.32	4.12	6.89	41.18
Resample C2	4.60	5.58	4.77	8.12	42.05
Resample D1	4.84	7.36	2.85	3.30	35.81
Resample D2	4.79	7.21	2.85	3.80	35.40
Multiplicative Error	1.28	12.43	3.58	2.79	14.23

Table 3 gives an impression of the effects on econometric parameter estimations produced by the different kinds of anonymization. The original linear model is constructed as follows:

$$Y = 0.7 + 0.5X_1 + X_2 + 0.2X_3 + u, \quad u \sim N(0, 1). \tag{5}$$

In the simulations, the model–parameters are alternately estimated with OLS-method using perturbed versions of the orignal variables Y and X_1, X_2, X_3. The expectation is that estimations with resamples lead to unbiased parameter estimates if multivariate relationships are retained. In case of multiplicative perturbation, model parameters cannot be consistently estimated. Hwang (1986) [10] shows how to correct biased estimates to get consistent results if the distribution of the multiplicative random number is known. The co–variance–matrix of the errors E_i ($i = 1, 2, ..., I$, where I is the number of quantitative variables) can be computed. As they are independently distributed, all co–variances are zero and $Var(E_i) = \frac{1}{12}$ ($i = 1, 2, ..., I$). A consistent estimator with a perturbed data matrix Z and an endogenous variable Y_z with weight $\mathcal{U} = diag[E(E_i^2)]$, $E(E_i^2) = Var(E_i) + E(E_i)^2$, can easily be constructed[7] [8]:

$$\hat{\beta} = [(Z'Z) \div \mathcal{U}]^{-1} Z'Y_z. \tag{6}$$

Table 3 shows the results of the Monte Carlo simulations for the different types of resamples and data with multiplicative perturbation. The last column contains estimated values with correction term $\mathcal{U}$. Coefficients that do not significantly differ by more than 5% from the original parameters (see above) are marked with*[9].

[7] See an application with simulated data and with the Mannheim Innovation Panel in manufacturing of the year 1997 in Gottschalk, 2002 [8].

[8] $A \div B$ is the Hadamard division of the matrizes A und B, i.e. elementwise division.

[9] A t-test (100 observations) is computed to reveal significant deviations.

Table 3: Monte Carlo Simulations - OLS-Regression Results

Variable	Resample A1	Resample A2	Resample B1	Resample B2
X_1	0.499*	0.493	0.483	0.457
(t-stat.)	(19.98)	(17.71)	(17.16)	(13.54)
X_2	0.987	0.974	0.971	0.941
(t-stat.)	(65.87)	(58.33)	(57.55)	(46.53)
X_3	0.201*	0.198*	0.195*	0.185
(t-stat.)	(5.63)	(4.99)	(4.86)	(3.84)
Const.	0.727	0.759	0.779	0.873
(t-stat.)	(10.13)	(9.50)	(9.66)	(10.82)
R^2	0.90	0.88	0.88	0.83

Variable	Resample C1	Resample C2	Resample D1	Resample D2	Multiplicative Error uncorrected	Multiplicative Error corrected
X_1	0.487	0.483	0.503*	0.503*	0.452	0.501*
(t-stat.)	(20.42)	(20.01)	(22.58)	(22.53)	(14.72)	(15.26)
X_2	0.998*	0.998*	0.997*	0.997*	0.900	0.997*
(t-stat.)	(69.21)	(68.34)	(74.40)	(74.15)	(48.72)	(50.53)
X_3	0.177	0.165	0.203*	0.201*	0.176	0.197*
(t-stat.)	(5.19)	(4.79)	(6.36)	(6.31)	(4.11)	(4.27)
Const.	0.741	0.757	0.702*	0.704*	0.969	0.710*
(t-stat.)	(10.82)	(10.91)	(10.95)	(10.95)	(10.73)	(7.53)
R^2	0.91	0.91	0.92	0.92	0.72	0.80

In all cases, coefficients remain significantly unequal to zero and signs do not change. Resampling procedures D1 and D2 as well as the corrected estimates in the case of multiplicative perturbation produce the best results as all coefficients are significantly equal to the real values. When regression analysis is carried out without a correction term, parameter estimates with multiplicated perturbed data slightly differ from the original, goodness of fit R^2 clearly decreases as well. Weighted or scaled resamples with co–variance–matrix (B– and C–versions) produce some remarkably differing regression coefficients, e.g. the constant in B and the parameter of X_3 in C. A larger bandwidth in the second versions, respectively, worsen parameter estimations. For example, the coefficient of X_1 is significant unequal to 0.5, in contrast to version A1.

To be concise, one can say that univariate distributions can be best reproduced by resamples with unweighted kernels, whereas multivariate structures can be better retained in using directional information in form of the correlation structure of the data when constructing resamples. Multiplicative perturbation re–performs univariate and multivariate distribution parameters quite well. However, in linear regression analysis a correction term should be implemented to retain accurate parameter estimates. Yet, it should be noted that in comparison, the last method has a lower re–identification risk.

4 Empirical Application - An Example

In a second step, anonymized data are constructed by using real data - the Mannheim Innovation Panel (MIP)[10] in the manufacturing sector from 1999 (see also Appendix C). Five quantitative variables are chosen: "sales", "number of employees", "research and development (R&D) expenditure per sales" (R&D-intensity), "innovation expenditure" and the "number of high qualified personnel". These variables are censured to the left, i.e. they have only positive values. The resampling procedure here used does not consider these restrictions. Therefore, a few of the synthetic observations in the resample have negative signs. Tests have shown that this fact does not matter for a lot of descriptive and regression analyses. However, variables with a notable number of values that are zero - e.g. R&D expenditure - are difficult to reproduce due to the fact that the share of zeros cannot be maintained. Some modifications of the resampling procedure are necessary. This will be the subject of further work.

Table 4: MIP - Confidentiality Measure (CM)

	Resample A1	Resample A2	Resample B1	Resample B2	
CM	61.5%	67.4%	53.6%	53.7%	
	Resample C1	Resample C2	Resample D1	Resample D2	Multiplicative Error
CM	99.6%	99.7%	68.8%	74.2%	84.6%

Table 4 lists the share of links that could not be made with the original data. "Sales" and "number of employees" are chosen as link variables.[11] The procedure additionally divides the data in industry classes (two–digit NACE-level) and East and West German firms. Hence, a link can only be made within a strata. In contrast to the Monte Carlo simulations above, the anonymized datasets on average involve higher re–identification risks for the individual firms. Where resampling D best protects the data with a confidentiality level of nearly 100%, resampling B-versions have re–identification risks of nearly 50%.

[10] The scientific-use-files of the MIP are freely available for purely non-commercial basic research. External users are informed about anonymization techniques. The applied anonymization methods are described in Gottschalk, 2002 [8]. In the scientific-use-files of the MIP, firm specific random numbers are used and only "sales" and "number of employees" are perturbed with multiplicative error, which is an univariate random number between [0.5;1.5]. Hence, productivity (sales per number of employees) remains constant. Here, each continuous variable is multiplied by different random numbers and the scenarios in this paper are not completely transferable to the scientific-use-files of the MIP.

[11] In a realistic scenario, one would also presume to have "sales" and "number of employees" as common variables in anonymized microdata and additional data.

Multiplicative data perturbation preserves confidentiality on a high level (85%) compared to the average of resampling procedures.

Informational loss due to different anonymization methods is measured as the sum of variables' "sales", "number of employees" and "innovation expediture"[12] average absolute deviation from original statistics (as described in Appendix B). The results are presented in Table 5. In contrast to the Monte Carlo simulations, resampling - except of versions C - performs very well in reproducing univariate and multivariate distribution parameters, whereas similar errors occur in each statistic when applying multiplicative perturbation to the data. These findings seem to be an indication of dependence between the used disturbance technique and the data generating process. One should consider that different results could occur regarding various datasets.

Table 5: MIP - Average Absolute Deviation in %

Method	Means	Variances	Correlations	Rank-Correlations	Co–variances
Resample A1	0.02	0.00	0.00	2.98	0.00
Resample A2	0.03	0.00	0.00	5.12	0.01
Resample B1	0.00	0.00	0.00	0.68	0.00
Resample B2	0.00	0.00	0.00	0.67	0.00
Resample C1	7.14	20.72	39.06	18.98	59.00
Resample C2	8.27	20.83	34.06	19.24	54.29
Resample D1	0.05	0.02	0.00	5.52	0.01
Resample D2	0.08	0.03	0.00	7.01	0.03
Multiplicative Error	6.38	19.95	3.63	2.06	11.15

Table 6 in Appendix D presents the results of parameter estimations (OLS) of an exemplary linear model, explaining "R&D-intensity", defined as "R&D expenditure per sales", of innovative firms[13]. The independent variables are "ln(firm size)" (logarithm of the "number of employees") and its square ("ln(firm size)^2"), a measure of market concentration within an industry (2-digit NACE level), the Herfindahl index ("herfindahl")[14], a dummy variable indicating an expected positive growth rate of sales ("demand"), an indicator of East German firms ("East"), the share of highly qualified personnel within the enterprise ("qualified pers."), a binary variable, which takes the value one if the firm introduced at least one innovation, which is essentially based on new developments at scientific

[12] These variables are used because they are highly correlated.

[13] Innovators are firms that have successfully completed at least one innovative project within a three-year period.

[14] I calculate the Herfindahl index from estimated market shares of firms in the Mannheim Enterprise Panel (MUP), which includes 12,000 firms (see Almus et al., 2000 [1], for more detailed information on the MUP).

institutions (“science”), and twelve sector dummies (two-digit NACE level: 10,15,17-36) to control for heterogeneity. For the sake of simplicity, the latter are not listed in the table. All variable combinations are constructed before the anonymization processes. This is necessary because variables within the resample cannot be combined meaningfully (see above). Densities of functions of variables in resamples strikingly differ from the original ones. For comparability, shares of sales are also computed before masking data with multiplicative error. As mentioned before, only quantitative values are perturbed regarding both resampling and multiplicative perturbation. All indicator variables and the Herfindahl index (a firm level variable) are maintained.

Regression results do not strikingly differ from the original values in the cases of resamples D1 and D2. Only the effects of concentration and the share of qualified personal are under- and overestimated, respectively. This result confirms the conclusions of the Monte Carlo simulations. The A-versions of resampling also perform quite well. Significance levels and signs remain in most cases and values do not differ remarkably. Only the coefficients of the Herfindahl index are not significant anymore. Resamples B and C do not produce satisfactory results as already seen in the Monte Carlo study. Multiplicative errors underestimate the effect of employment–level, even the sign of variable-parameter ln(Firm Size)^2 is reversed. This mistake can be eliminated by the correction term, but the coefficient of ln(Firm size) greatly decreases compared to the original estimation result. In comparison to the simulation disturbance is even more severe regarding linear regression analyses. As mentioned above, one should consider that the extent of deterioration is a cumulated effect of the applied anonymization technique but also of special data characteristics and model specifications. The effect of anonymization on regression results maybe multiplied if model specifications are wrong or essential explaining variables are missing.

5 Resume

The Monte Carlo studies and application with real data, the Mannheim Innovation Panel, show the effects of resampling in comparison to multiplicative data perturbation on different kinds of analyses. Univariate distributions can be best re–performed by resampling with independent kernels. When using the correlation structure of the original data as scales by constructing the kernel–matrix, resampling best retains linear regression results but univariate distributions - especially of skewed variables - are biased. Multiplicative perturbation reproduces descriptive statistics quite well and in linear regression analysis a correction term could be implemented in order to retain accurate parameter. Dependent on the kind of anaylses - univariate or multivariate, the optimal resampling techniques varies. For econometric calculations versions D are most suitable as shown in the simulations and the empirical example. But univariate distributions fairly deviate in some cases, as shown with the help of kernel densitiy estimations in Appendix E. In comparison, multiplicative perturbation as well as resampling versions A performs better due to the reproduction of univariate and multivariate distribution parameters and regression results can be retained fairly.

Though resamples consist of synthetic values, confidentiality problems remain. On average, anonymized data with multiplicative perturbed values performs better. But confidentiality increases when the kernels are correlated. Though data quality and data protection is not a contradiction in any case. When involving the correlation structure of the data by constructing the various kernels, confidentiality is relatively high and multivariate relations are produced quite well. As long as data characteristics are satisfactorily be retained, the enlargement of the kernel bandwidth can be used to improve confidentiality. The optimal point on the trade–off between confidentiality and data quality has to be discovered in further work. Confidentiality measures and estimations of re–identification risks should be computed in realistic scenarios where additional databases are used for a match with perturbed microdata sets to finally assess the various anonymization techniques.

It remains to examine effects on non–linear and semi–parametric model estimations as well as on different kinds of model specifications with the MIP. The latter is quite important as potential mis–specifications may influence regression analysis and deterioration of regression results due to anonymization may not be independent of wrong model specifications.

References

[1] Almus, M., D. Engel and S. Prantl (2000), The ZEW Foundation Panels and the Mannheim Enterprise Panel (MUP) of the Centre for European Economic Research (ZEW), Schmollers Jahrbuch 120, 301-308.

[2] Brand, R. (2000), Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, Nürnberg.

[3] Congdon, P. (2001), Bayesian Statistical Modelling, Wiley Series in Probability and Statistics, New York.

[4] Devroye, L. and L. Györfi (1985), Nonparametric Density Estimation, New York.

[5] Fienberg, S.E. (1997), Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, Technical Report No. 161, Carnegie Mellon University, Pittsburgh.

[6] Fienberg, S.E. (1994), A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality, Technical Report No. 611, Carnegie Mellon University, Pittsburgh.

[7] Fienberg, S.E., R.J. Steele und U. Makov (1996), Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models, Proceedings of US Bureau of the Census 1996 Annual Research Conference, US Bureau of the Census, Washington DC, 87-105.

[8] Gottschalk, S. (2002), Anonymisierung von Unternehmensdaten - Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels, ZEW Discussion Paper 02-23.

[9] Härdle, W., S. Klinke und M. Müller (1991), XploRe - Learning Guide, Berlin.

[10] Hwang, J.T. (1986), Multiplicative Errors-in-Variables Models with Application to Recent Data Released by U.S. Department of Energy, Journal of the American Statistical Association, 81, 395, 680-688.

[11] Janz, N., G. Ebling, S. Gottschalk und H. Niggemann (2001), The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW), Schmollers Jahrbuch 121, Journal of Applied Social Science Studies, 123-129.

[12] OECD (1997), Oslo Manual: Proposed Guidelines for Collecting and Interpreting Technological Innovation Data, Paris.

[13] Robert, C.P. and G. Casella (2002), Monte Carlo Statistical Methods, New York.

[14] Parzen, E. (1962), On Estimation of a Probability Density Function and Mode, Annals of Mathematical Statistics 32, 1065-1076.

[15] Rosemann, M. (2003), Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik, IAW-Diskussionspapier 9.

[16] Rubin, D. (1993), Discussion - Statistical Disclosure Limitation, Journal of Official Statistics, 9, 461-468.

[17] Rubin, D. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, USA.

[18] Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability 26, London.

[19] Spruill, N.L. (1983), The Confidentiality and Analytic Usefulness of Masked Business Microdata, American Statistical Association, Proceedings of the Section on Survey Research Methods 1983, Washington, D.C., 602-610.

[20] Tapia, R.A. and J.R. Thompson (1978), Nonparametric Probability Demsity Estimation, Baltimore, John Hopkins University Press.

A Transformation of the Kernel

1. Reform the kernels, generated from equation 1, to be zero correlated:

$$\begin{aligned}\tilde{\kappa} &= \kappa A_k \\ VCV_k^{-1} &= A_k' A_k,\end{aligned}$$

   where VCV is the co–variance–matrix of the kernels.

2. Transform the kernels to the desired co–variances (C1,C2) or correlations (D1,D2):

$$\begin{aligned}\hat{\kappa}_j &= \tilde{\kappa} A_j, \quad j = cov \vee j = corr \\ VCV &= A_{cov}' A_{cov} \\ C &= A_{corr}' A_{corr},\end{aligned}$$

   where VCV and C are the co–variance– and the correlation–matrix of the original variables $X_1, X_2, ..., X_I$.

B Measures of Information Loss

For all variables $X = (X_1, X_2, ..., X_I)$ of the original data and $Z = (Z_1, Z_2, ..., Z_I)$ of the perturbed data the following measures are computed to estimate information loss due to the different anonymization procedures:

- Average relative absolute deviation of means:

$$\frac{1}{I}\sum_{i=1}^{I}\frac{|\bar{Z}_i - \bar{X}_i|}{|\bar{X}_i|}$$

- Average relative absolute deviation of variances:

$$\frac{1}{I}\sum_{i=1}^{I}\frac{|\sigma_{z_i}^2 - \sigma_{x_i}^2|}{|\sigma_{x_i}^2|}$$

- Average relative absolute deviation of co–variances:

$$\frac{1}{\frac{1}{2}I(I-1)}\sum_{i=1}^{I-1}\sum_{j>i}^{I}\frac{|\sigma_{z_{ij}} - \sigma_{x_{ij}}|}{|\sigma_{x_{ij}}|}$$

- Average relative absolute deviation of correlations:

$$\frac{1}{\frac{1}{2}I(I-1)}\sum_{i=1}^{I-1}\sum_{j>i}^{I}\frac{|\rho_{z_{ij}} - \rho_{x_{ij}}|}{|\rho_{x_{ij}}|}$$

- Average relative absolute deviation of rank–correlations (Spearman rank correlation coefficient):

$$\frac{1}{\frac{1}{2}I(I-1)}\sum_{i=1}^{I-1}\sum_{j>i}^{I}\frac{|s_{z_{ij}} - s_{x_{ij}}|}{|s_{x_{ij}}|}$$

C The Mannheim Innovation Panel

The Mannheim Innovation Panel (MIP) was assigned by the German government to conduct an innovation survey representative of the German economy leading to internationally comparable data on the innovation behaviour of German firms. It started in 1993 as a voluntary mail survey and is constructed as a panel with yearly waves. The population of the MIP covers legally independent German firms in the sectors mining and manufacturing. In 1995 the survey on innovation activities of distributive and business related service sector firms (Mannheim Innovation Panel - Services, MIP-S) was additionally initiated. Up to 2002 the MIP and MIP-S have been running ten times in co-operation with infas Institute for Applied Social Science. The MIP is strongly based on the recommendations on innovation surveys manifested in the Oslo-Manual of the OECD and Eurostat (OECD, 1997 [12]). It provides basic information on product and process innovations, innovation activities and components of innovation expenditure related to these activities (see Janz et al., 2001 [11]). The data are available in an anonymized version (scientific-use-file) to external users for non-commercial basic research. Currently, more than 30 researchers utilize the scientific-use-files.

D Estimation Results

Table 6: R&D-intensity - OLS-Regression Results

Variable	Original	Resample	Resample	Resample	Resample	Resample	Resample	Resample	Resample	Multiplicative Error	
		A1	A2	B1	B2	C1	C2	D1	D2	uncorrected	corr.
ln(Firm size)	-0.024	-0.018	-0.015	-0.006	-0.002	-0.004	-0.004	-0.022	-0.021	-0.004	-0.077
(t-stat.)	(-3.73)	(-3.22)	(-2.95)	(-1.52)	(-0.59)	(-2.28)	(-2.25)	(-3.56)	(-3.48)	(-3.27)	(-1.91)
ln(Firm Size)^2	0.002	0.001	0.001	0.000	-0.000	0.000	0.000	0.002	0.002	-0.000	0.006
(t-stat.)	(3.24)	(2.69)	(2.35)	(0.67)	(-0.50)	(0.79)	(1.04)	(3.04)	(2.95)	(1.28)	(1.57)
Concentration	0.082	0.029	0.029	0.047	0.052	0.027	0.027	0.022	0.020	0.096	0.063
(t-stat.)	(3.00)	(1.03)	(1.05)	(1.64)	(1.80)	(0.98)	(0.97)	(0.79)	(0.71)	(3.52)	(2.18)
Demand	0.008	0.004	0.004	0.008	0.008	0.004	0.004	0.005	0.005	0.005	0.012
(t-stat.)	(1.61)	(0.96)	(0.94)	(1.61)	(1.76)	(0.88)	(0.89)	(0.99)	(1.00)	(1.11)	(1.29)
East	0.016	0.018	0.018	0.020	0.020	0.017	0.018	0.018	0.019	0.013	0.013
(t-stat.)	(3.15)	(3.62)	(3.66)	(3.93)	(3.81)	(3.46)	(3.49)	(3.63)	(3.69)	(2.59)	(2.13)
Qualified Pers.	0.134	0.187	0.187	0.070	0.002	0.194	0.194	0.184	0.183	0.137	0.089
(t-stat.)	(7.81)	(11.10)	(11.12)	(4.02)	(0.10)	(11.55)	(11.56)	(10.91)	(10.91)	(8.81)	(5.44)
Science	0.016	0.016	0.016	0.008	0.007	0.018	0.018	0.016	0.015	0.017	0.019
(t-stat.)	(2.61)	(2.62)	(2.61)	(1.20)	(1.04)	(2.91)	(2.89)	(2.61)	(2.51)	(2.79)	(1.74)
Constant	0.063	0.060	0.051	0.025	0.016	0.023	0.022	0.069	0.065	0.011	0.210
(t-stat.)	(2.93)	(2.91)	(2.62)	(1.45)	(0.97)	(1.46)	(1.42)	(3.23)	(3.13)	(0.83)	(1.42)
R^2	0.23	0.27	0.27	0.13	0.08	0.27	0.27	0.28	0.27	0.22	0.29
Obs.	885	920	920	920	920	920	920	920	920	885	885

E Kernel Density Estimates

Figure 1: Normal Distribution

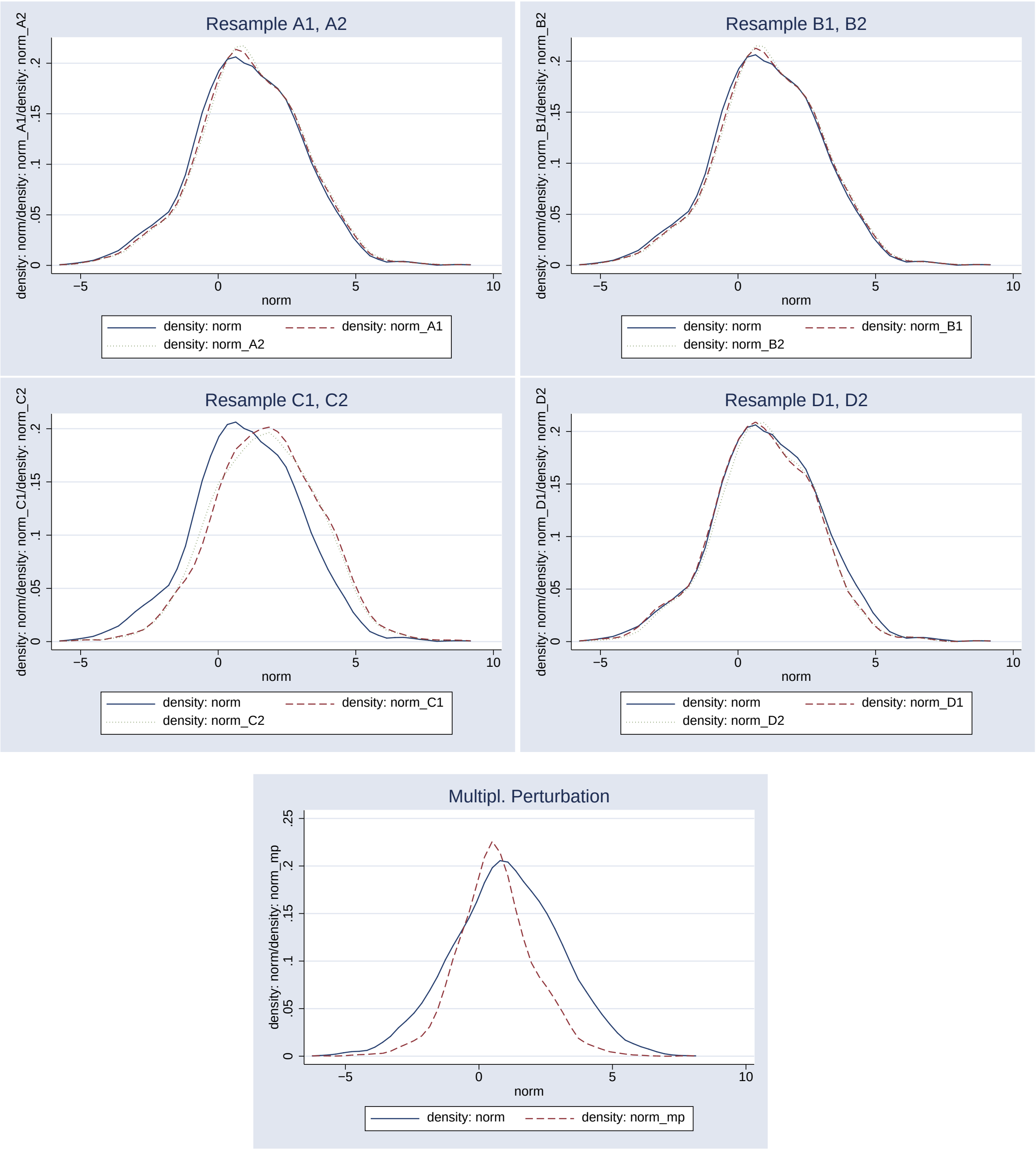


Figure 2: Exponential Distribution

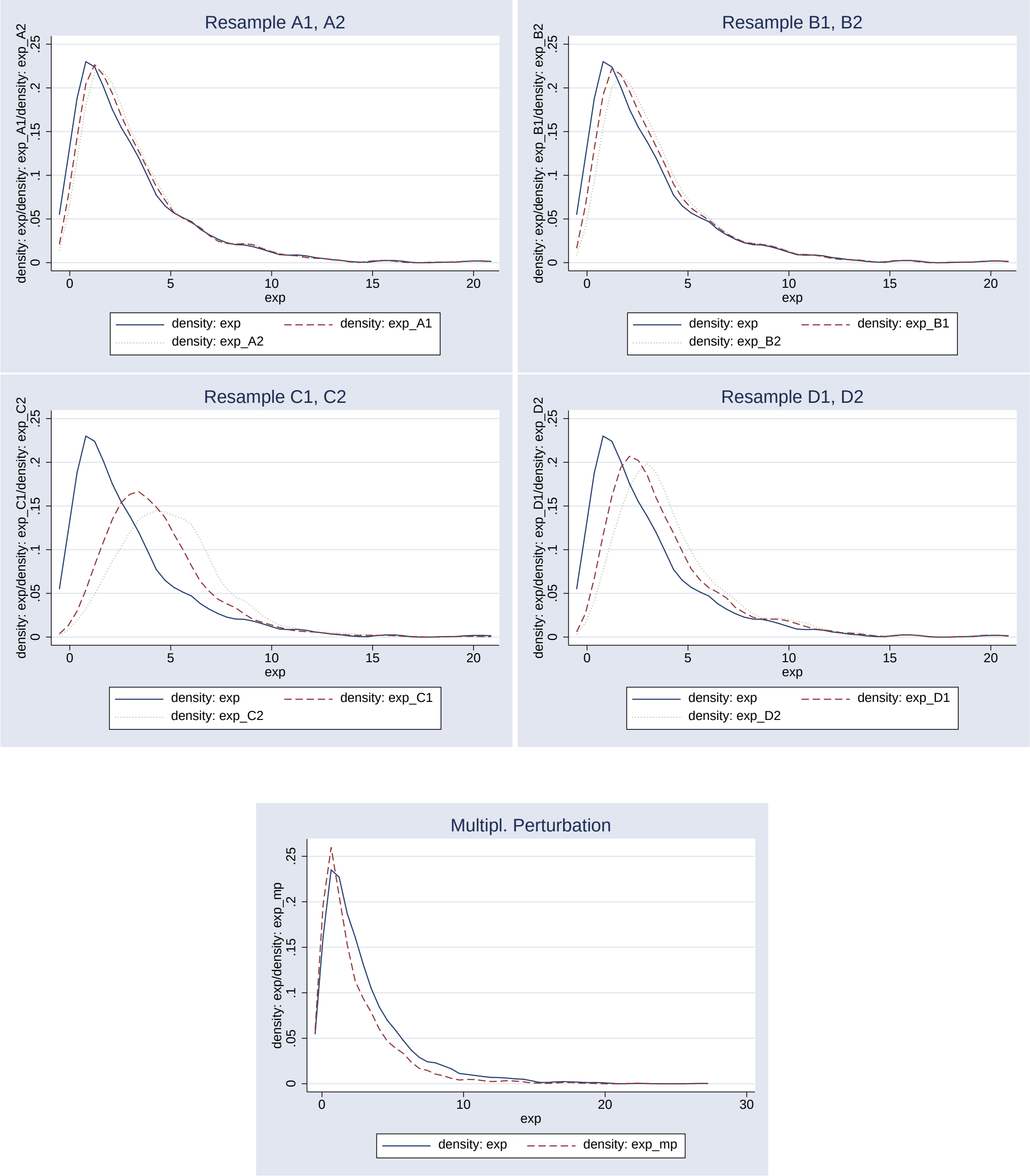


Figure 3: Chi-square Distribution

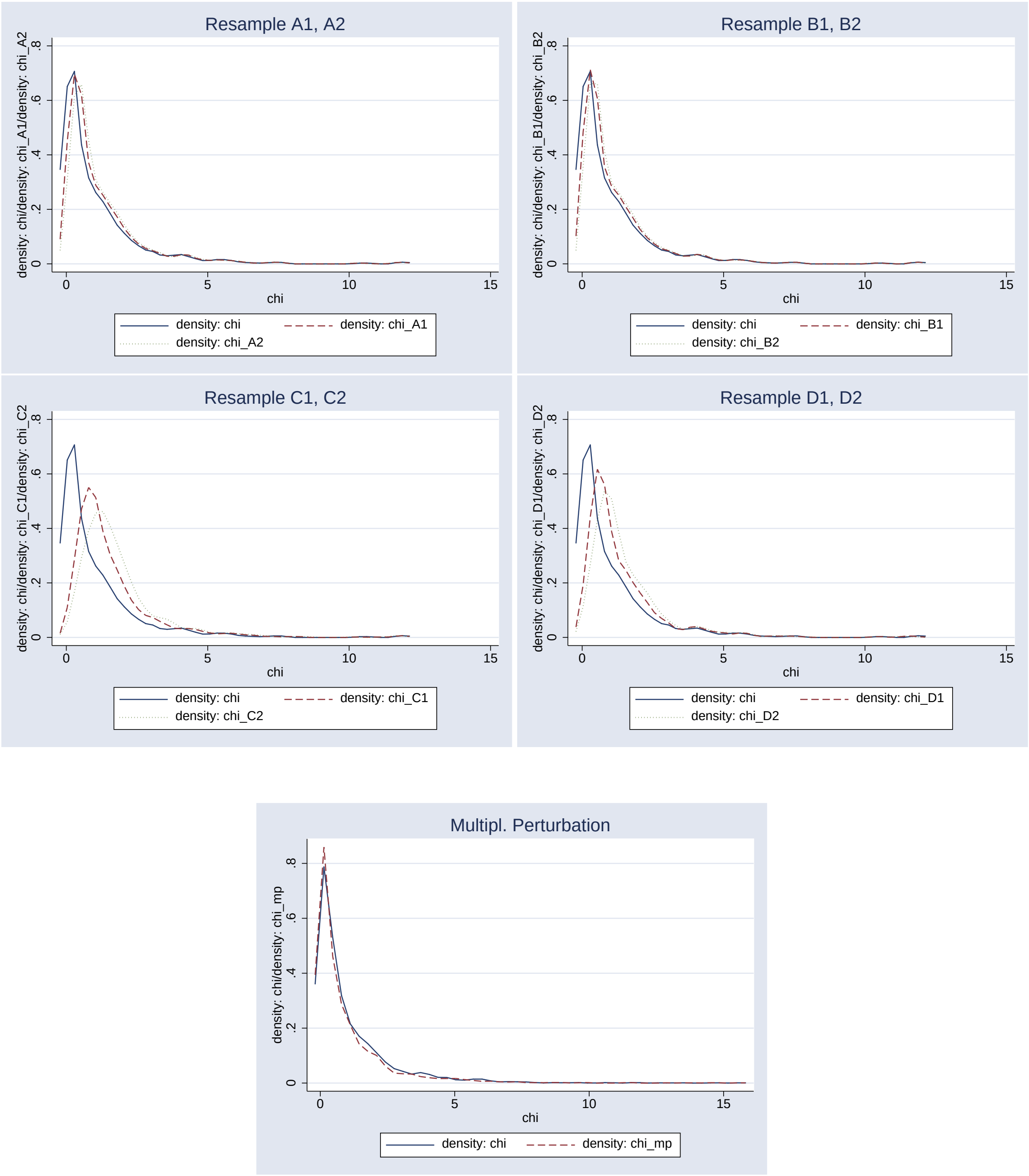


Figure 4: Logarithm of Variable 1

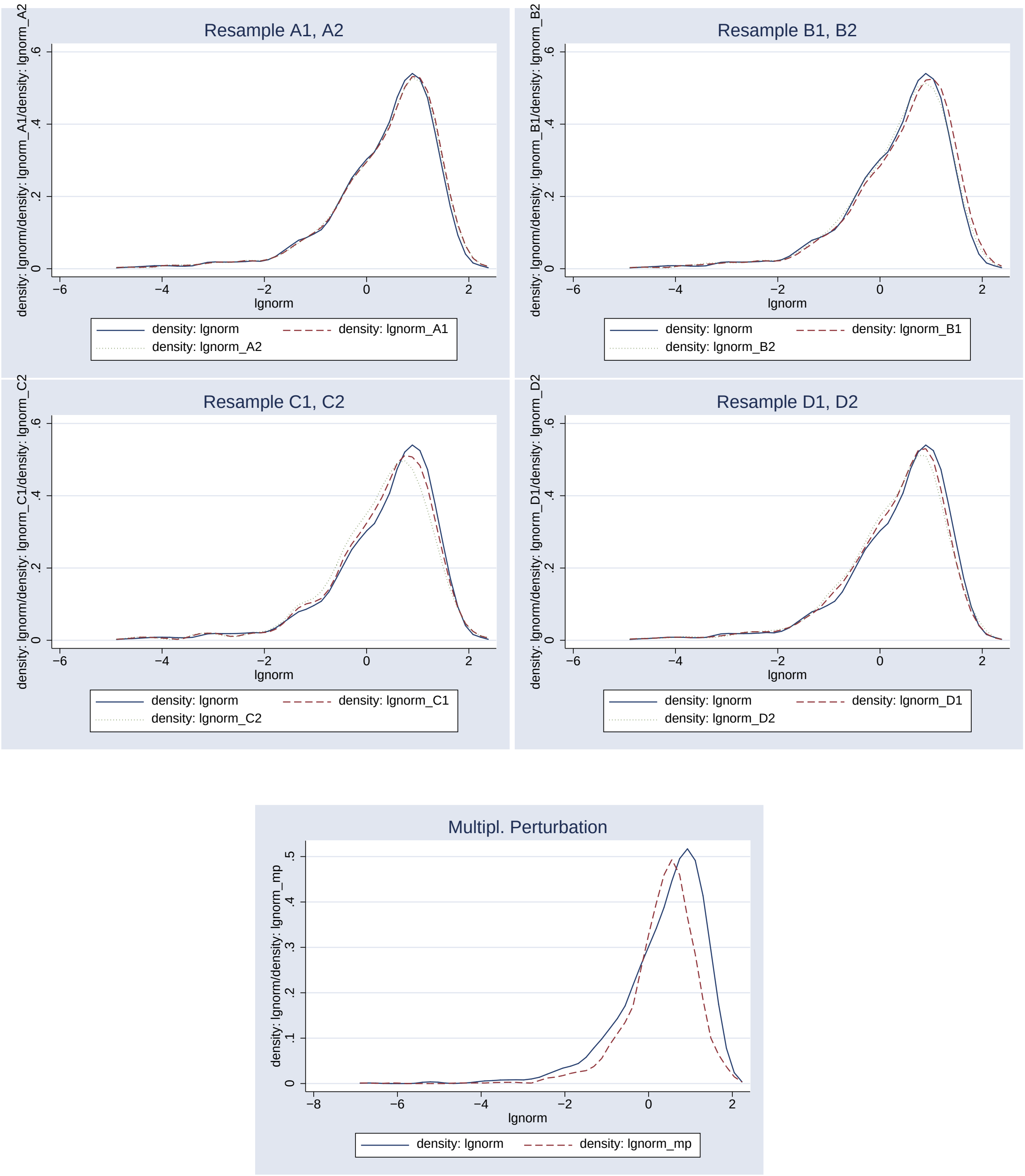


Figure 5: Mixed Distribution 1

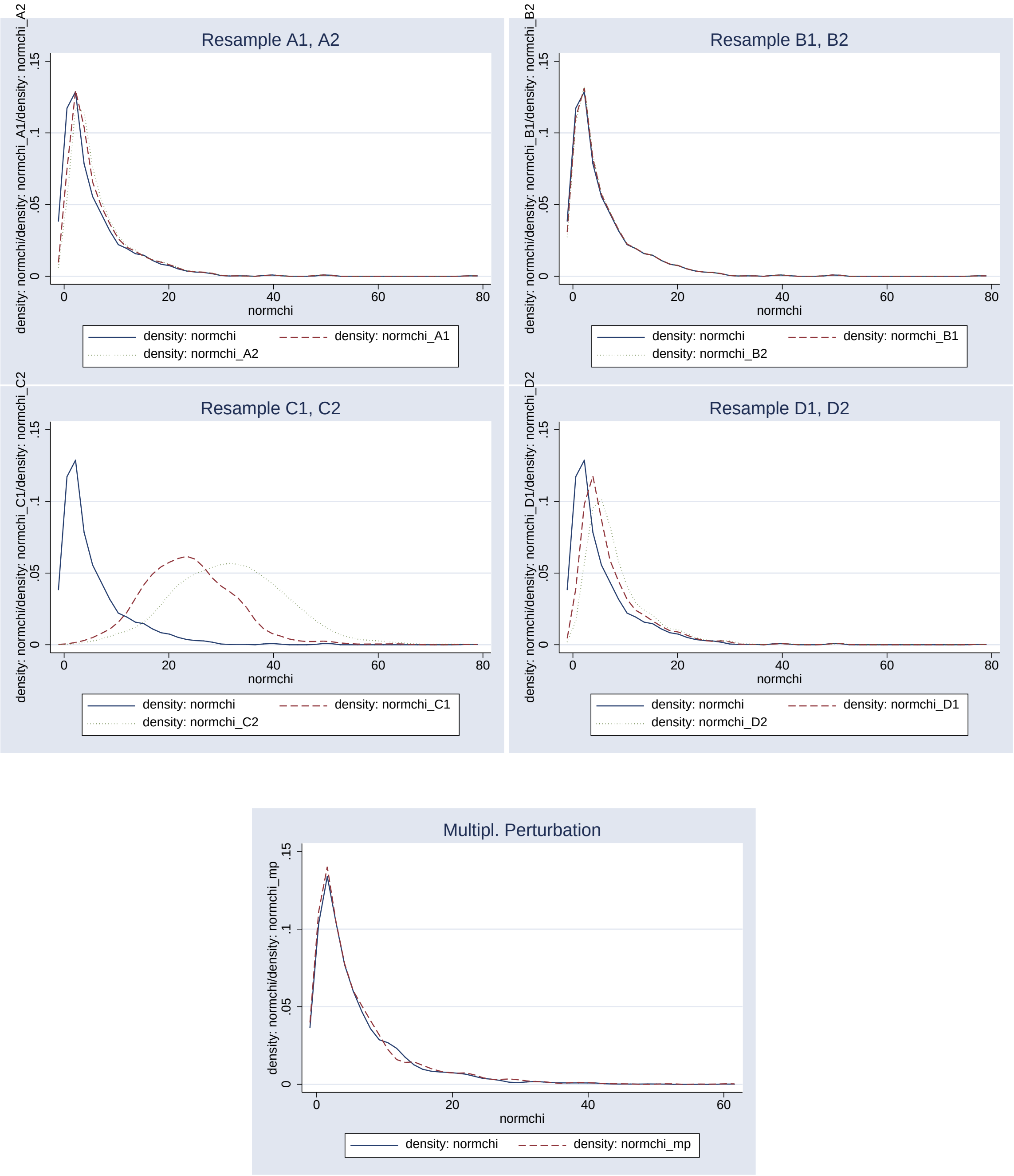
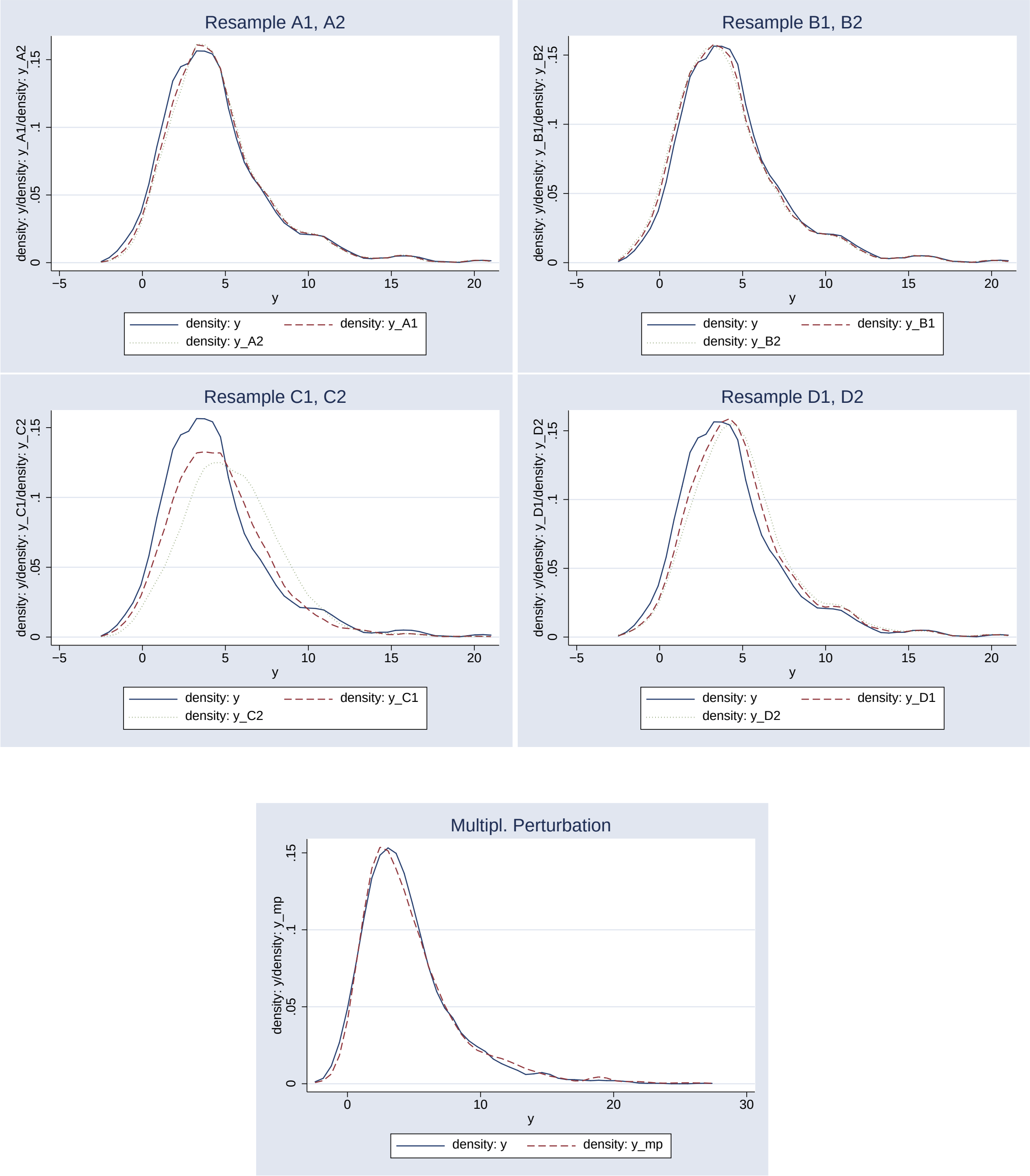


Figure 6: Mixed Distribution 2 - Endogenous Variable y


Resample A1, A2
density: y/density: y_A1/density: y_A2
y
density: y
density: y_A1
density: y_A2
Resample B1, B2
density: y/density: y_B1/density: y_B2
y
density: y
density: y_B1
density: y_B2
Resample C1, C2
density: y/density: y_C1/density: y_C2
y
density: y
density: y_C1
density: y_C2
Resample D1, D2
density: y/density: y_D1/density: y_D2
y
density: y
density: y_D1
density: y_D2
Multipl. Perturbation
density: y/density: y_mp
y
density: y
density: y_mp