Journal of Privacy and Confidentiality (2012) **4**, Number 2, 61–94

Estimation of Regression Parameters from Noise Multiplied Data

Yan-Xia Lin[*] and Phillip Wise[†]

Abstract. This paper considers the scenario that all data entries in a confidentialised unit record file are masked by multiplicative noises, regardless of whether unit records are sensitive or not and regardless of whether the masked variables are dependent or independent variables in the underlying regression analysis. A technique is introduced in this paper to show how to estimate parameters in a regression model, which is originally fitted by unmasked data, based on masked data. Several simulation studies and a real-life data application are presented.

AMS Subject Classification: 62D99, 62J99, 62P

Keywords: Confidential data, Masked data, Multiplicative noise, Regression analysis, Statistical disclosure limitation

1 Introduction

A data set is said to be a confidential data set when it contains information concerning an individual and/or entity which can be identified by someone analysing the data. Confidential data sets are ubiquitous, especially in national statistical institutions like the Australian Bureau of Statistics (ABS), who utilise large sample surveys of individuals and businesses. The information obtained by such institutions is necessarily confidential, as respondents expect their personal responses to carry an assurance of privacy.

Obvious identifying details such as name and address are automatically removed by statistical agencies and are replaced with random identification codes for analytical purposes. Further confidentiality protection can be provided by applying methods of statistical disclosure limitation (SDL), also known as statistical disclosure control (SDC).

Many masking methods have been proposed over the years as interest in data confidentiality has intensified. They include microaggregation of sensitive data, local suppression of unique data cells, top and bottom coding of continuous variables, rank swapping, rounding, adding noise, and imputation. More information can be found in Willenborg and de Waal (2001), Duncan and Lambert (1986 and 1989), Oganian (2010), Shlomo (2010b), and the references therein.

Masking data by multiplicative noise is not a new idea and carrying out regression analysis on confidential data is a well-covered topic. Nayak et al. (2011) gives

[*] School of Mathematics and Applied Statistics, University of Wollongong, NSW 2500, Australia, `mailto:yanxia@uow.edu.au`.

[†] Australian Bureau of Statistics, ACT 2617, Australia, `mailto:philip.wise@abs.gov.au`.

an interesting discussion of the statistical properties of multiplicative noise masking for confidentiality protection. Previous multiplicative noise techniques such as Evans (1996), Evans et al. (1998) and Krisinich and Piesse (2002) focus on applying noise to sensitive cells only, introducing the need to consider cut-offs to help determine those vulnerable cells in need of protection. Kim and Winkler (2003) discuss the use of multiplicative noise that follows a truncated normal distribution with mean 1 and derived formulas for evaluating the mean and variance of unmasked data in terms of the mean and variance of the masked data.

Current research on how to implement regression analysis with confidential data can be found in Karr et al. (2007, 2006, 2005) and Sanil et al. (2004). Their research does not consider the situation where an agency such as the ABS has a very large data set, but rather focuses on utilising integrated databases and information sharing amongst institutions to obtain information for regression analysis.

Hwang (1986) introduced a method of estimating regression parameters from noise multiplied data. His approach is only valid for the scenario where multiplicative noises are only applied to covariates, and he assumes that observations of each independent variable in regression analysis are considered an i.i.d. sample from a population with unknown distribution. Hwang claims that this assumption is especially appropriate for the energy problem addressed in his paper. Given the assumption of i.i.d. on observations of covariates, Hwang's result is derived based on the strong law of large numbers, i.e., $(\sum_{i=1}^{n} x_i)/n \to EX$ with probability 1, where $\{x_i\}$ is a sample from X. There are two limitations in the application of Hwang's results. The first limitation concerns the real life data applications. In real life, once a data set is made available to the public, data providers have no control on the use of the variables in the data set. Data users may assign any variable in the released data set as a dependent or independent variable based on their research purpose. In Hwang's formula, any dependent variables in a regression analysis are not masked. Not knowing which variables will be used as dependent variables in data users' study, data providers have difficulties making any decisions on which variables in data sets should not be masked before they are issued to the public. The second limitation is that, from a theoretical point of view, the assumption of i.i.d. is not necessary.

In this paper, we develop a technique for using masked microdata to estimate the parameters in linear regression models. To simplify the discussion in this preliminary study, we do not consider the scenario where the masked microdata arise from a survey with survey weights and design variables. All variables in this study are masked by independent multiplicative random noises, including the case where some variables are masked by noise equal to 1, i.e., unmasked. The technique is called the global-multiplicative noise regression method, or G-multiplicative noise regression method. The "G" indicates that regardless of whether data are sensitive or not and regardless of whether an original variable is an independent/dependent variable in the underlying regression analysis. The formula for the estimator of β, regression parameter(s), given by Hwang can be considered as a special case of this paper if multiplicative noises are uncorrelated. In contrast to Hwang's work, the formulae for the estimator of β provided in this paper are derived without the assumption of i.i.d.. The formulae given in this paper are derived based on Theorem 1 in Section 3, which provides a more general result than that given by the strong law of large numbers. The results presented in this paper are more practical.

An alternative to multiplicative noise is to use additive noise, which is popular because it can perturb records in an unbiased manner by adding noise with mean 0. More information on additive noise microdata protection can be found in Brand (2002), Fuller (1993), Shlomo (2010a) and references therein. This method generally does not preserve variances and correlation coefficients as well as the regression coefficients. However, when the covariance matrix of noise is proportional to the covariance matrix of the original data, the additive noise method can preserve the correlation coefficients (Domingo-Ferrer et al., 2004) and preserve regression coefficients. Due to the restriction on the covariance matrix of noise, the additive noise method will lead to asymptotically **biased** regression estimates for subpopulations; see Tendick and Norman (1987); Domingo-Ferrer et al. (2004).

Ting et al. (2008) introduced the Random Orthogonal Matrix Masking (ROMM) procedure for perturbing data in regression analysis. This method preserves sample means and covariances (and hence preserves linear regression estimates), and the magnitude of the perturbation is controlled. However, if basic uncorrelated additive noise is used, then biased, inconsistent regression estimates are obtained with higher variance. When using bias corrected and correlated additive noise, the estimates from the true data are maintained but variability is added. Also, ROMM preserves inferences under normality using correlated random noise, but is indistinguishably effective when compared to additive noise regression estimation. It is difficult to determine the effectiveness of ROMM on orthogonal matrices when small perturbations are involved, and it is too computationally intensive to assess the disclosure risk and data utility of ROMM, since a Bayesian framework involving posterior distributions is necessary.

Compared to the additive noise method, the G-multiplicative noise regression method has the following advantages: (i) all multiplicative noises used by the G-multiplicative noise regression method can be independently decided; (ii) the property of asymptotically unbiased estimators given by the G-multiplicative noise regression method is retained for the whole data set as well as its subsets, regardless of whether there is a connection between the covariance matrix of noise and the covariance matrix of the original data.

The Post RAndomisation Method (PRAM) was introduced by Kooiman et al. (1997) and Gouweleeuw et al. (1998) as a method for disclosure protection of categorical variables in microdata files. Microdata files might usually contain special structures between variables, e.g., a hierarchical structure when all members of a household are present in a data file. It is of importance to make sure that a masked data set conserves the structure, otherwise it might give an intruder a clue as to which data were altered as a result of masking. But, this is not an issue for the multiplicative noise method, as all categorical covariates are independently masked and might be converted to non-categorical covariates (see Example 2 in Section 5) and, then, the hierarchical structures among categorical variables becomes unobservable in masked microdata files. The valuable idea we obtained from the PRAM discussion given by de Wolf et al. (1998) is about the concept of "uncertainty." The authors develop the concept of reasonable uncertainty in the mind of an analyst in relation to whether a rare combination of scores in the perturbed data file is the result of applying PRAM or if the rare combination existed in the population. The concept of "reasonable uncertainty" is of benefit to the determination of disclosure risk developed in Section 2.

Evans et al. (1998) introduced the idea of measuring cell value changes after multiplicative noise perturbation, which is useful in this paper as a way of measuring the re-identification risk after multiplicative noise has been applied. The measure is essentially a relative distance of the perturbed value from the underlying value, expressed as a percentage. In this paper it is interpreted as a means of assessing the accuracy of the true data estimated by a hacker: the greater the distance between the true data and the analyst's estimate, the greater the level of confidentiality protection on the data.

Duncan et al. (2001, 2004) introduce an R-U confidentiality map that traces the joint impact on risk and utility of changes in the parameters of a disclosure limitation procedure. The aim is to give a means to compare procedures and the trade-offs between disclosure risk and data utility. The concept of the R-U map is used in this paper to explore the trade-off between estimation accuracy and confidentiality protection.

This paper is organized as follows: Section 2 discusses the risk measure for the multiplicative noise method. The theory on how to estimate the parameters in linear regression models based on masked data is presented in Section 3, followed by derivation of the asymptotic distribution and the variance of the estimator. In Section 4, we give a formula using masked data to estimate the standard errors of the ordinary least squares (OLS) estimator based on the model fitted by unmasked data. This piece of information can be used to construct hypothesis tests on regression parameters without accessing confidential data. Section 5 gives the applications of the G-multiplicative noise regression method to simulation data and real-life data. All proofs, some large tables and figures, and extra examples are listed in Appendix A–F.

2 Protecting data by multiplied noise

A basic discussion on masking data by multiplicative noise can be found in Nayak et al. (2011). For reading convenience, some basic concepts and definitions on noise multiplied data are introduced in this section. This section also addresses the effectiveness of multiplicative noise as a method for protecting confidential data by a distance-based risk measure which evaluates the re-identification risk of masked data. This risk measure is used as a basis of selecting a distribution for the noise multipliers.

The multiplicative noise method is described as follows: *let C be a random noise with mean $E(C) > 0$ and $Var(C)$. Given an observation y, independently and randomly draw an observation c from C. Then multiply c by y and release $y^* = c * y$ to the public. Thus y is protected by y^*.*

The value of $E(C)$ in the definition of multiplicative noise is usually assigned as 1 in the literature. If $E(C)$ is public, there is no difference between having the assumption $E(C) > 0$ and $E(C) = 1$ as $C/E(C)$ will have mean 1 if $E(C) \neq 1$. Giving various developed security techniques, it becomes feasible to pass masked data to data users in a portable storage, e.g., CD, with encoded values of $E(C)$ and $Var(C)$, and it is also feasible for data providers to build software, for example an R package, allowing data users to freely read masked data from the publicly issued CD without accessing the values of $E(C)$ and $Var(C)$ and to carry out linear regression by themselves. Since the values of $E(C)$ and $Var(C)$ are read in the background and the values of $E(C)$ and $Var(C)$ are invisible, intruders cannot expect that the released masked y^* is an unbiased estimator of y and they will have less chance to guess the true value of y based on the value of y^* only. However, the multiplicative noise method still works well in protecting original data even if the values $E(C)$ and $Var(C)$ are publicly issued, if appropriate multiplicative noise is used. (see Example A2 in Appendix B).

Clear attention must be paid to any observations that are in fact equal to zero. Clearly, multiplying noise into these observations will have no effect. To protect these points, special consideration needs to be paid before the multiplicative noise method is applied to the underlying data set, including randomly adding noise to these points or randomly deleting some of these points from their original data set. When the G-multiplicative noise regression method is considered in conjunction with zero records partially masked by other means, the regression analysis outputs given by the G-multiplicative noise regression method is for the modified data set, instead of for the original data set. It is a challenge how to perturb the values of zero appropriately in the original data set before the multiplicative noise such that the regression analysis outcomes for the modified data set are as close as the analysis outcomes for the original data set. However, this issue is out of the scope of this paper and will not be discussed here.

The measure of re-identification risk of an individual record for different masking schemes and different types of data might have different focus. Risk measures for microdata arising from survey data are typically assessed through probabilistic models (Bethlehem, et al. 1990; Elamir and Skinner, 2006) or through record-linkage methods (Torra, et al. 2006; Yancey, et al. 2002). They mainly aim to estimate the probability that a sample unique is a population unique through a set of categorical identifying variables as the probability is an indicator of how confident we are to issue the survey data to public.

de Wolf et al. (1998) discussed the expectation ratio of a rare score in PRAM. Given the value of a masked variable, the expectation ratio of a rare score of the variable is the ratio of the probability of the original variable taking the value to the probability of the original variable not taking the value. de Wolf et al. (1998) pointed out that the smaller the value of the expectation ratio, the more likely it is that a record in the perturbed file with the masked value does not originally belong to the original value, and thus the safer the perturbed file is. In other words, to use the measurement of uncertainty to measure if a masked file is safer.

A number of formulae for measuring "information loss" for masked data are introduced in Yancey et al. (2002). Relative difference between the original value and its masked value is one of the alternative ways for measuring the "information loss." In terms of protecting original data, the more difference there is between the masked data and its original data, the safer the original data will be.

We regard "risk" in this context as being the likelihood of an analyst identifying original values in underlying microdata sets. Adopting the concepts of the measurement of uncertainty and the measurement of information loss, this section explains the measurement of the risk proposed and used in this paper.

Many interesting statistical properties of noise multiplied data can be found in Nayak et al. (2011). Those properties in Nayak et al. (2011) are given based on $E(C) = 1$. Because the mean value of C is not specified in this paper, we repeat some relevant properties in Nayak et al. (2011) here and present them based on $E(C) \neq 1$.

Although it is possible to conceal the values of $E(C)$ and $Var(C)$ from data users, the discussion of data protection below is still based on the assumption that the values are publicly known. It is of interest how confident we are on data protection if both of these pieces of information are available to public.

Given the values of y^*, $E(C)$, and $Var(C)$, a simple and reasonable way to estimate y is $\hat{y} = y^*/E(C)$, which is an unbiased estimator of y conditional on y. This unbiased estimator is adopted in this paper when the multiplicative noise method is used. The variance of $\hat{y}$ conditional on y is

$$Var(\hat{y}|y) = y^2 \frac{Var(C|y)}{E(C)^2} = y^2 \frac{Var(C)}{E^2(C)}, \tag{1}$$

where the three factors y, $Var(C)$, and $E(C)$ can leverage $Var(\hat{y}|y)$.

We suggest use of

$$P[|\frac{\hat{y} - y}{y}| < \delta] = P[|\frac{y^*/E(C) - y}{y}| < \delta] = P[|C/E(C) - 1| < \delta], \tag{2}$$

where $\delta > 0$, to evaluate the protection level for the data which were masked by multiplicative noise C.

2.1 Data protection evaluated by the relative error measurement

No one is able to make a 100% correct guess on the value of y based on knowledge of $\hat{y} = y^*/E(C)$ only. Any intruders will allow a certain level error in their guess on y, based on the value of $\hat{y} = y^*/E(C)$. Therefore, we should take into account this fact when we evaluate disclosure risk of the multiplicative noise method. In this paper, we assume that an intruder uses the acceptance rule below to decide whether the guess he/she made is correct.

Acceptance Rule: *Denote $\tilde{y}$ as a guess given by an intruder on the value of y based on $\hat{y} = y^*/E(C)$. The value of $\tilde{y}$ is accepted as a correct guess on y if the relative difference between the true value of y and $\tilde{y}$ is less than 0.05, that is $|(y - \tilde{y})/y| < 0.05$.*

Different intruders may use different upper bounds of relative difference. In this paper, we use 0.05 as an example for our discussion.

Any multiplicative noise C, with $P[|C/E(C)-1| < 0.05] < 1$, will be able to provide a certain level of protection on any non-zero numerical data protected by C. The protection level will be affected by the probability distribution of C. For an appropriately chosen C, the smaller the probability $P[|C/E(C)-1| < 0.05]$ is, the higher the data protection will be.

Knowing y^*, $E(C)$, and $Var(C)$, different intruders may follow their own ways to guess the true value of y. Therefore, the methods used to evaluate the success guess rate will be different. In Appendix A, we suggest a manner for guessing the true value of data. We show that the larger $\delta_0 = \min_\delta\{\delta \mid P(|\frac{C}{E(C)}-1| < \delta) = 0.9999\}$ is, the higher the data protection will be if intruders follow the manner we suggested. Appendix A gives an example of evaluating data protection. Example A1 (in Appendix A) shows that, among the four considered multiplicative noises, the multiplicative noise with a bi-modal distribution is more powerful than a normal distribution in protecting data. The protection might be lost if an inappropriate distribution is selected.

All multiplicative noises used in this paper enable us to provide a reasonable level of data protection, although they might not be the most efficient. Which type of multiplicative noise will provide more efficient protection of the data is of interest in practice. It needs further intensive discussions and those discussions are out of the scope of this paper. In practice, it may not be wise to limit the choice of multiplicative noises into certain types of families. A clever intruder may be able to find an efficient way to guess the true value of y, if he/she knows the curve of the probability density function of C. Not putting any restrictions on the type of multiplicative noise may be a strategy of protecting data.

2.2 $R-U$ plot

Duncan et al. (2004) discussed data utility U and disclosure risk R. An R-U map is suggested for evaluating disclosure risk. We employ an R-U map to display the relationship between the measure of statistical disclosure risk $R(\delta)$ and data utility U. For the multiplicative noise method, the data utility U and the measure of statistical disclosure risk related to $\delta > 0$, $R(\delta)$, are defined as follows:

$$U = [E(\frac{y^*/E(C)-y}{y})^2]^{-1} = [\frac{Var(C)}{E^2(C)}]^{-1}, \tag{3}$$

and

$$R(\delta) = P(|\frac{y^*/E(C)-y}{y}| < \delta) = P(|C/E(C)-1| < \delta). \tag{4}$$

The data utility U is the reciprocal of the mean of the square of the relative differences between y and its unbiased estimator. Therefore, the larger the value U, the less the relative difference between y and its unbiased estimator will be.

Example 1 demonstrates how to use an R-U map to show how efficient a multiplicative noise is in protecting data.

Example 1. Consider two types of multiplicative noises, C_4 and C_1. C_4 has probability density function

$$f(x) = [f_1(x) + f_2(x) + f_3(x) + f_4(x)]/4,$$

where $f_i(x)$ is the probability density function of $N(a_i, 1)$ with $a_i = a_1 + (i-1)d$, a_1 a real number, and $d = 450$, $i = 2, 3, 4$. Noise C_1 is normally distributed with the same mean and variance as C_4. The ratio of $Var(C)$ to $E^2(C)$ as well as the utility U will be the same for both C_1 and C_4.

Let a_1 vary in value from 250 to 8150 with increment 100 in this example. As shown in Fig.1, the values of $Var(C)/E^2(C)$ decrease as a_1 increases. Thus, utility $U = [Var(C)/\ E^2(C)]^{-1}$ increases as a_1 increases. The plots $R(\delta) = P(|C/E(C) - 1| < \delta)$ against U for $\delta = 0.05$ and 0.25 are shown in the second panel of Fig.1.

The R-U map is useful in showing which noise might provide better protection of data under the multiplicative noise method. As explained in Section 2.1 and Appendix A, a multiplicative noise with smaller value of $R(0.05)$ and larger value of δ_0 (defined in Section 2.1) will enable protection of the data. Using C_4 and C_1 as an example, Fig.1 shows that all C_4 and C_1 with different distribution parameters demonstrate their capability in protecting data. As expected, $R(\delta)$ will increase to 1 as Utility increases (see the second panel of Fig.1). Fig.1 shows that, comparing with the value for C_1, $R(0.05)$ for C_4 is generally much less than 1 when Utility is less than 300 and the δ_0 given by C_4s tends to be larger than those given by C_1s (not shown here). It indicates that a four-modal noise might provide more protection on data than a normal noise in this example when the values of Utility determined by their distribution are less than 300.

If the upper bound of the relative difference 0.25 is considered by an intruder, Fig.1 shows that $R(0.25)$ given by C_4 with $a_1 = 2150$ rises to 1 faster than those given by C_1. This noise C_4, determined by $a_1 = 2150$, gives $Var(C_4)/E^2(C_4)$ as small as 0.0317 or $U = 31.55$. It might suggest that this C_4 can not protect data properly, but this is not true (see Example A2 in Appendix B).

We use Example 1 to show a basic idea of how to select an appropriate multiplicative noise from candidate multiplicative noises based on their R-U plots. However, making a decision of choosing multiplicative noise for an underlying data set is a complex processes and is beyond the purpose of this paper. Therefore, we will not discuss this issue further in this paper.

2.3 Comparison between the methods of additive noise and multiplicative noise

An interesting discussion on the comparison between the methods of additive noise and multiplicative noise can be found in Nayak et al. (2011). In this section, we use

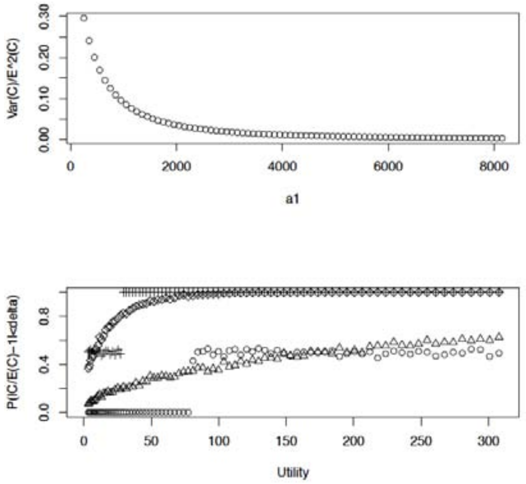

Figure 1: The top plot shows the relationship between $Var(C)/E^2(C)$ and a_1. The bottom plot shows the plot of $R(\delta) = P(|C/E(C) - 1| < \delta)$ versus U for $a_1 = 150 + 100m$, $m = 1, 2, \cdots, 80$, and $d = 450$. Triangles and diamonds represent $R(0.05)$ and $R(0.25)$ given by C_1 respectively; circles and crosses represent $R(0.05)$ and $R(0.25)$ given by C_4 respectively.

a simulation study to compare the multiplicative noise and additive noise approaches in terms of which method is able to provide more protection on data under the risk measurement suggested in this paper.

Denote y as a true observation, $\tilde{Y}$ as the observation masked by additive noise $\tilde{C}$, and Y^* as the observation masked by multiplicative noise C. Thus, $\tilde{Y} = y + \tilde{C}$ with $E(\tilde{C}) = 0$; $Y^* = Cy$ with $E(C) > 0$. Given $\tilde{Y}$ and Y^*, the unbiased predictions of y are $\hat{Y}_{add} = \tilde{Y}$ and $\hat{Y}_{multi} = Y^*/E(C)$, respectively.

The conditional variances of the difference between the true observation and its prediction given by the two methods are:

$$Var(\hat{Y}_{add} - y|y) = Var(\tilde{C}) = \sigma^2_{add}, \text{ and } Var(\hat{Y}_{multi} - y|y) = \frac{Var(C)}{E^2(C)} y^2.$$

The variance of $(\hat{Y}_{add} - y)$ depends on σ^2_{add} only, but the variance of $(\hat{Y}_{multi} - y)$ is related to the ratio of $Var(C)$ to $E^2(C)$ as well as the unaccessible true observation y.

Since σ^2_{add} is not available to the public, the additive noise method provides a level of protection. The value of σ^2_{add} should not be too small, otherwise the true value of y might be easily correctly estimated from $\hat{y}_{add}$. However, σ^2_{add} should not be too large either, otherwise too much noise is added to the underlying data and it might affect the quality of inference results based on masked data.

Comparing to the additive noise method, the multiplicative noise method has less restriction on the variance of C and so it can be flexibly applied to data in practice. Because the value of y is unknown, data will be protected through the uncertainty of $Var(\hat{Y}_{multi} - y)$ even if the values of $Var(C)$ and $E(C)$ are available for the public.

The additive noise method and the multiplicative noise method are two totally different methods. Each method has its own application areas and has its own advantages in different scenarios. It may be meaningless to discuss which method, the additive noise method or the multiplicative noise method, will offer better protection of data without considering the nature of the underlying data and the measurement of data protection. However, to show that sometimes the multiplicative noise method can be an alternative to the additive noise method in data protection, we use a simulation example in Appendix C to demonstrate that for a given additive noise $\tilde{C}$, there is a multiplicative noise C such that (i) C has the same type of probability distribution as $\tilde{C}$ and (ii) C is able to provide better protection of data than $\tilde{C}$ in terms of the relative error measurement defined in this paper.

3 Estimating the parameters in linear regression models by using masked data

In this section, we introduce a method for estimating the parameters in a regression model, which is originally fitted to unmasked data, when only masked data is available. The method, named the G-multiplicative noise regression method, is based on the following well known theorem (VII.8 and Theorem 1, Feller, 1966, p.236; Csorgo, 1968):

Theorem 1. *Let $X_1, X_2, \cdots$ be random variables on some probability space $(\Omega, \mathcal{B}, P)$ satisfying*

$$E(X_1) = 0, \qquad E(X_n | X_1, \cdots, X_{n-1}) = 0, \qquad n \geq 2,$$

with probability 1. *Define $S_n = X_1 + X_2 + \cdots + X_n$. If $0 < b_1 < b_2 < \cdots \to \infty$ and $\sum_1^\infty b_k^{-2} E(X_k^2) < \infty$, then*

$$\lim_{n\to\infty} b_n^{-1} S_n = 0 \qquad \textit{and} \qquad \sum_1^\infty b_k^{-1} X_k < \infty \qquad \textit{with probability 1.}$$

From Theorem 1, we have the following Corollary.

Corollary 1. *Let $0 < b_1 < b_2 < \cdots \to \infty$ be a sequence of positive real numbers and $\{C_i\}$ be a sequence of i.i.d. random variables with finite moments up to order r. If a sequence of real numbers $\{w_k\}$*

(i) is bounded and $\sum_{k=1}^\infty b_k^{-2} < \infty$; or

(ii) $\sum_{k=1}^\infty w_k^2 / b_k^2 < \infty$,

then

$$\frac{1}{b_n}\sum_{i=i}^{n}[C_i^{r/2} - E(C_i^{r/2})]w_i \to 0 \qquad \textit{with probability 1.}$$

The formulae for the estimation of parameters calculated from masked data are derived in this section in two steps. In the first step, we derive a formula where only the dependent variable is masked. Then, we extend the formula for general situations where all variables in the underlying regression model are masked.

3.1 Only dependent variable masked

Let dependent variable Y and covariates $X_1, \cdots, X_p$ satisfy the following linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{5}$$

where ε is a random error with mean 0 and variance σ^2.

Let C be a random variable independent of Y with mean $E(C) > 0$ and finite moments up to order 4, and c_i, $i = 1, 2, \cdots, n$, be a random sample drawn from C. For observations $(y_{i,}, x_{i,1}, \cdots, x_{i,p})$, $i = 1, 2, \cdots, n$, let $\mathbf{y}^* = (c_1 y_1, \cdots, c_n y_n)' = (y_1^*, \cdots, y_n^*)'$ and $\mathbf{y} = (y_1, \cdots, y_n)'$. Denote $\hat{\beta}_{OLS}^{(n)}$ the Ordinary Least Squares (OLS) estimator of $\beta = (\beta_0, \cdots, \beta_p)'$ in the model

$$y_i^* = c_i(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) + \eta_i \qquad i = 1, 2, \cdots, n. \tag{6}$$

Therefore,

$$\hat{\beta}_{OLS}^{(n)} = [X'(C^{(n)})'(C^{(n)})X]^{-1}[X'(C^{(n)})'\mathbf{y}^*]$$

where $C^{(n)} = diag(c_1, c_2, \cdots, c_n)$ and $X = (\mathbf{1}, \mathbf{x_1}, \cdots, \mathbf{x_p})$ is the design matrix.

Theorem 2. *Adopting the notation above, if there is a sequence of real numbers* $0 < b_1 < b_2 < \cdots \to \infty$ *such that* $\sum_{n=1}^{\infty} b_n^{-2} < \infty$, $\{y_i, x_{i,j}\}_{i \geq 1, k=1, \cdots, p}$ *are bounded and*

(i) $(1/b_n)X'X$ *has a non-singular limit as* $n \to \infty$,

(ii) $(1/b_n)X'\mathbf{y}$ *has a limit with probability* 1 *as* $n \to \infty$,

then $\hat{\beta}_{OLS}^{(n)} - (X'X)^{-1}X'\mathbf{y}^*/E(C) \to 0$ *with probability* 1, *as* $n \to \infty$.

The proof of Theorem 2 is presented in Appendix D.

Remarks: (i) $\hat{\beta}_{OLS}^{(n)}$ is the least squares estimator of β for (6). Technically, it is different from the least squares estimator $\hat{\beta}_{OLS}$ for (5). But both of them are unbiased and converge with probability 1 to the same true parameter β when the sample size n tends to infinity (explained at the end of Section 3.2). Therefore, $\hat{\beta}_{OLS}^{(n)}$ is a reasonable estimator for β and can be used to replace the least squares estimator for (5). (ii) The important message of Theorem 2 is that $\hat{\beta}_{OLS}^{(n)}$ can be estimated by

$$\hat{\beta}_C^{(n)} = (X'X)^{-1}(X'y^*)/E(C). \tag{7}$$

Therefore, $\hat{\beta}_C^{(n)}$ can be used to estimate β. In fact, $\hat{\beta}_C^{(n)}$ is a consistent estimator of β (See Section 4). Formula (7) provides a fundamental technique for linear regression analysis under the multiplicative noise method when the dependent variable is confidential and masked. (iii) Usually the sequence of $\{b_n\}$ can be chosen as $\{n\}$ in many practical situations. (iv) If condition "$\{y_i, x_{i,j}\}_{i\geq 1,k=1,\cdots,p}$ are bounded" is replaced by "$\{x_{i,j}\}_{i\geq 1,k=1,\cdots,p}$ are bounded and $\{y_i\}_{i\geq 1}$ are bounded in probability", the result of Theorem 2 will hold with the probability measure for noise variable C and Y. The proof for such weak result is straightforward and omitted from this paper.

3.2 All data masked

Theorem 3. *Assume that Y satisfies the following model*

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where ϵ is a random error with mean 0 and variance σ^2; $\beta_0, \beta_1, \cdots, \beta_p$ are unknown parameters and need to be estimated. For observations $(y_i, x_{i,1}, \cdots, x_{i,p})$, $i = 1, 2, \cdots, n$, denote $\mathbf{y} = (y_1, \cdots, y_n)'$ and $X = (\mathbf{1}, \mathbf{x_1}, \cdots, \mathbf{x_p})$. Assume that $\{y_i, x_{i,j}\}_{i\geq 1,k=1,\cdots,p}$ are confidential, cannot be issued to the public, and are bounded. Let C, $Z_1, \cdots, Z_p$ be mutually independent random variables. All of them have positive mean and finite moments up to order 4.

The information available for regression analysis is $\mathbf{y}^* = (y_1^*, \cdots, y_n^*)$ and $X^* = (\mathbf{1}, \mathbf{x_1^*}, \cdots, \mathbf{x_p^*})$, $E(C)$, $Var(C)$, $\{E(Z_j)\}$, and $\{Var(Z_j)\}$, $j = 1, \cdots, p$, where $y_i^* = c_i y_i$ and $x_{i,j}^* = z_{i,j} x_{i,j}$, $i = 1, \cdots, n$, $j = 1, \cdots, p$; $\{c_i\}$ and $\{z_{i,j}\}$ are independent samples from C and $\{Z_j\}$ respectively. If there is a sequence of real numbers $0 < b_1 < b_2 < \cdots \to \infty$ such that

(i) $(1/b_n)X'X$ has a non-singular limit as $n \to \infty$,

(ii) $(1/b_n)X'\mathbf{y}$ has a limit with probability 1 as $n \to \infty$,

then $\hat{\beta}_{OLS}^{(n)} - \hat{\beta}_{C,Z}^{(n)} \to 0$ with probability 1 as $n \to \infty$, where $\hat{\beta}_{OLS}^{(n)}$ is the least squares estimator of β defined in Theorem 2 and

$$\hat{\beta}_{C,Z}^{(n)} = \frac{1}{E(C)} A^{-1} B' \mathbf{y}^*, \tag{8}$$

where

$$A^{-1} = \begin{pmatrix} n & \frac{\sum_{i=1}^n x^*_{i,1}}{E(Z_1)} & \frac{\sum_{i=1}^n x^*_{i,2}}{E(Z_2)} & \cdots & \frac{\sum_{i=1}^n x^*_{i,p}}{E(Z_p)} \\ \frac{\sum_{i=1}^n x^*_{i,1}}{E(Z_1)} & \frac{\sum_{i=1}^n x^{*2}_{i,1}}{E(Z_1^2)} & \frac{\sum_{i=1}^n x^*_{i,1}x^*_{i,2}}{E(Z_1)E(Z_2)} & \cdots & \frac{\sum_{i=1}^n x^*_{i,1}x^*_{i,p}}{E(Z_1)E(Z_p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n x^*_{i,p}}{E(Z_p)} & \frac{\sum_{i=1}^n x^*_{i,p}x^*_{i,1}}{E(Z_p)E(Z_1)} & \frac{\sum_{i=1}^n x^*_{i,p}x^*_{i,2}}{E(Z_p)E(Z_2)} & \cdots & \frac{\sum_{i=1}^n x^{*2}_{i,p}}{E(Z_p^2)} \end{pmatrix}^{-1}$$

$$B' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \frac{1}{E(Z_1)}x^*_{1,1} & \frac{1}{E(Z_1)}x^*_{2,1} & \cdots & \frac{1}{E(Z_1)}x^*_{n,1} \\ \frac{1}{E(Z_2)}x^*_{1,2} & \frac{1}{E(Z_2)}x^*_{2,2} & \cdots & \frac{1}{E(Z_2)}x^*_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{E(Z_p)}x^*_{1,p} & \frac{1}{E(Z_p)}x^*_{2,p} & \cdots & \frac{1}{E(Z_p)}x^*_{n,p} \end{pmatrix}.$$

Following the technique used in the proof of Theorem 2, Theorem 3 is straightforward. Remark (iv) under Theorem 2 also applies to Theorem 3.

Recalling Theorem 3 and Remarks under Theorem 2, the ordinary least squares estimator of β for (5) can be reasonably approximated by $\hat{\beta}^{(n)}_{C,Z}$ when all variables in (5), dependent variable and covariates, are masked by independent multiplicative noise. $\hat{\beta}^{(n)}_C$ is a special case of $\hat{\beta}^{(n)}_{C,Z}$ when all $Z_{i,j} = 1$, i.e., $\{x_{i,j}\}$ were not masked. To simplify notation, we use $\hat{\beta}^{(n)}$ to denote $\hat{\beta}^{(n)}_C$ or $\hat{\beta}^{(n)}_{C,Z}$, depending on whether only the dependant variable is masked.

Based on Lebesgues's dominated convergence theorem (Loeve, 1963) and asymptotic results in the next section, in practice we always have the mean of $\hat{\beta}^{(n)}$ converging to β, the true value of parameter in (5) (see examples in Section 5). An application of this property in the G-multiplicative noise regression method is explained in the conclusion at the end of this paper.

4 Asymptotic distribution of $\hat{\beta}^{(n)}$, the variance of $\hat{\beta}^{(n)}$ and the estimation of the standard error of the OLS estimator

In this section we show under very weak conditions that $\sqrt{n}\hat{\beta}^{(n)}$ has asymptotically normal distribution. This result can be used to construct approximate tests and confidence sets for regression parameters. We also identify the factors which can be used to reduce the variances of $\hat{\beta}^{(n)}$ in the G-multiplicative noise regression method and derive a formula for estimating the variance of the OLS estimator for (5) based on masked data.

When Y and the covariates are masked, from (8), $\hat{\beta}^{(n)} = \frac{1}{E(C)}A^{-1}B'\mathbf{y}^*$. In Appendix E, we prove that

Theorem 4. *If*

(i) $\lim_{n\to\infty} \frac{1}{n} X'X = Q_1$ *is a non-singular matrix, and* $\|(\frac{1}{n}A)^{-1}\|$ *is bounded with probability* 1*;*

(ii)

$$\lim_{n\to\infty} 1/n \begin{pmatrix} n & \sum_{i=1}^n x_{i,1} & \cdots & \sum_{i,p}^n x_{i,p} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1} x_{i,1}^2 \frac{EZ_1^2}{(EZ_1)^2} & \cdots & \sum_{i=1}^n x_{i,1}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i,p}^n x_{i,p} & \sum_{i=1}^n x_{i,1}x_{i,p} & \cdots & \sum_{i=1} x_{i,p}^2 \frac{EZ_p^2}{(EZ_p)^2} \end{pmatrix} = Q_3;$$

(iii)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \left[(\sum_{j=0}^p x_{i,j}\beta_j)^2 \begin{pmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \\ x_{i,1} & x_{i,1}^2 \frac{EZ_1^2}{(EZ_1)^2} & \cdots & x_{i,1}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,p} & x_{i,1}x_{i,p} & \cdots & x_{i,p}^2 \frac{EZ_p^2}{(EZ_p)^2} \end{pmatrix} \right] = Q_4,$$

then

$$\sqrt{n}[\hat{\beta}^{(n)} - A^{-1}(X'X)\beta] = \sqrt{n}A^{-1}[\frac{1}{EC}B'y^* - (X'X)\beta]$$

$$\xrightarrow{D} Q_1^{-1}N(0, \frac{Var(C)}{(EC)^2}Q_4 + \frac{E(C^2)}{(EC)^2}\sigma^2 Q_3),$$

as $n \to \infty$.

The variance of $\hat{\beta}^{(n)}$ conditional on X can also be directly evaluated as follows:

$$Var_{(X)}(\hat{\beta}^{(n)}) = \frac{\sigma^2 E(C^2)}{E^2(C)} A^{-1}B'BA^{-1} + \frac{Var(C)}{E^2(C)} A^{-1}B' diag(a_{11}, \cdots, a_{nn})BA^{-1}.$$

$Var_{(X)}$ denotes the variance conditional on X, and $a_{jj} = (\sum_{i=0} x_{j,i}\beta_i)^2$ with $x_{j,0} = 1$ for all $j = 1, \cdots, n$.

If $E^2(C) >> Var(C)$, the value of $Var_{(X)}(\hat{\beta}^{(n)})$ will be dominated by

$$\frac{\sigma^2 E(C^2)}{E^2(C)}(A^{-1}B'BA^{-1}).$$

Using the results in Appendix E, if further $E^2(Z_i) >> Var(Z_i)$, $i = 1, 2, \cdots, p$, and $\|A^{-1}\ B'BA^{-1}\|$ is bounded, from the Dominated Convergence Theorem, $Var_{(X)}(\hat{\beta}^{(n)})$ will be approximately dominated by $\frac{\sigma^2 E(C^2)}{E^2(C)}(X'X)^{-1}$.

In summary, (i) $Var_{(X)}(\hat{\beta}^{(n)})$ could be closer to $\sigma^2(X'X)^{-1}$ if appropriate multiplicative noises are used; (ii) given the relationship between A^{-1} and $(X'X)^{-1}$ shown in Appendix E, the variance of the OLS estimator for (5) can be estimated by $\sigma^2 A^{-1}$ provided that the sample size of the underlying data is large and appropriate multiplicative noise are used. These results are demonstrated by simulation studies and a real-life data study in the next section.

5 Simulation study and real-life data application

5.1 Simulation study

A simulation example is presented in this section. It is used to demonstrate the application of the G-multiplicative noise regression method and empirically check the theoretical results given in Section 4.

Example 2. Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon = 2 + 1.5X_1 + 3X_2 + \varepsilon, \tag{9}$$

where $\varepsilon \sim N(0,1)$, $X_1 \sim U(0,20)$, particularly, X_2 is a categorical variable. If X_2 takes two values 0 and 1, the values of X_2, which are equal to zero, will not be protected by the multiplicative noise method. In terms of estimating the parameters in (9), we can consider the model

$$Y = \beta_0^* + \beta_1^* X_1 + \beta_2^* X_2^* + \varepsilon = (\beta_0 - x_0\beta_2) + \beta_1 X_1 + \beta_2 X_2^* + \varepsilon, \tag{10}$$

where $X_2^* = X_2 + x_0 \neq 0$ for any given real number $x_0 \neq -1$ and 0. Thus, X_2^* can be well protected by multiplicative noise and the estimation of $\beta = (\beta_0, \beta_1, \beta_2)$ can be obtained from the estimation of $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ in (10). Therefore, without loss of generality, we assume that X_2 in (9) takes two values, 1 and 2.

We simulated a sample with size 1000 from (9). The noises used to mask Y, X_1, and X_2 are C, Z_1, and Z_2, respectively, where C is four-modal distributed as defined in Example 1 with $a_1 = 150 + 18 \times 100$ and $d = 450$; both Z_1 and Z_2 have normal distribution $N((a+b)/2, \sqrt{1 + (a-b)^2/4})$ with $a = 170$ and $b = 80$. The estimator of $\hat{\beta}^{(n)}$ of $\beta = (\beta_0, \beta_1, \beta_2)$ was given by (8).

Given the sample $\{y_i, x_{i,1}, x_{i,2}\}_{i=1,\cdots 1000}$, we independently apply the G-multiplicative noise regression method to the sample 1000 times. The means of the estimates of β_0, β_1, and β_2 given by the G-multiplicative noise regression method were calculated and are reported in Table 1. The OLS estimators of parameters and the standard errors for the OLS estimators using the true data, as well as the means of the estimators of the standard errors calculated by the formula $\sigma^2 A^{-1}$ (see Section 4) are also reported in Table 1. Table 1 clearly shows that the means of estimates of parameters given by the G-multiplicative noise regression method are very close to their true values as well as the OLS estimators using the true data. The standard error for $\hat{\beta}_0^{(n)}$ is relatively large, the standard error for $\hat{\beta}_2^{(n)}$ is large compared to that for $\hat{\beta}_1^{(n)}$. Having a relatively larger standard error for some estimators of β is expected as the results are given by data with extra noise or the size of underlying sample is not sufficiently large. However, the analysis result can be improved in several ways. The standard errors for $\hat{\beta}^{(n)}$s can be reduced by increasing sample size. For example, if the sample size increases to 2000 in this example, we will have the mean of $\hat{\beta}^{(n)} = (1.999427, 1.501545, 2.981981)$ and standard error $(0.8975643, 0.04097169, 0.5483137)$. For more examples and discussion, see Example A4 in the Appendix F. Obviously, the standard errors for estimators can also be improved through sacrificing a certain level of data protection, for example choosing a multiplicative noise which has less capability in protecting data. Therefore, data providers might need to do a balance act between confidentiality protection and information loss. At the end of Section 5, we introduce another approach, an approach of multiple datasets, for improving the estimation of βs.

In Example 2, the categorical variable X_2 is masked by a continuous random variable Z_2 and converted to a non-categorical variable Z_2X_2. It is of interest to assess the safety of such a masking scheme. Given $Z_2 \sim N(125, 45.01)$ and X_2 takes two possible values, 1 and 2, we have $W_1 = Z_2X_2 \sim N(125, 45.01)$ given $X_2 = 1$ and $W_2 = Z_2X_2 \sim N(250, 90.02)$ given $X_2 = 2$. With probability 0.954, the values of W_1 and W_2 drop in $[34.978, 215.022]$ and $[69.956, 430.044]$, respectively. These two intervals overlap. The probabilities of W_1 and W_2 taking values in the overlapping interval $[69.956, 215.022]$ are 0.867 and 0.326. When an observation of Z_2X_2 drops outside the interval, the corresponding true value of X_2 can be easily identified. However, the true value of X_2 will not be easily identified if its masked value Z_2X_2 drops inside the overlapping interval. The probability of making a correct guess on the true value of X_2 will depend on $P(X_2 = 1)$, $P(W_1 \in [69.956, 215.022])$, and $P(W_2 \in [69.956, 215.022])$. Our investigation (omitted from this paper) shows a categorical variable (without taking value 0) can be well protected by a multiplicative noise if the differences between the values assigned to the categorical variable are relatively small compared to the variance of the multiplicative noise.

Example 2 demonstrates that the G-multiplicative noise regression method is practical for handling linear regression models with categorical covariates masked by continuous noise. Commonly data agencies tend to mask a categorical variable into a categorical variable. It might give data users an illusion that the data used is or is similar to the original data. Given the two facts that (i) by using the multiplicative noise method, most of the values that appear in masked data do not make sense to data users in general; and (ii) we only focus on linear regression with non-categorical dependent variables in this paper, it is not necessarily either to maintain the type of variable of a categorical variable after masking or to create an illusion for categorical variables in a masked data set.

More examples, showing how the sample size, the distribution of multiplicative noise, and the ratio $Var(C)$ to $E^2(C)$ impact on regression analysis are presented in Appendix F.

Table 1: The analysis outputs given by Example 2

	β_0	β_1	β_2
True value	2	1.5	3
OLS estimation	2.14172	1.50190	2.92109
mean of $\hat{\beta}^{(n)}$	2.10128 (1.318701)	1.507154 (0.05958237)	2.917266 (0.7494551)
	$se(\hat{\beta}_0)$	$se(\hat{\beta}_1)$	$se(\hat{\beta}_2)$
se. of OLS estimator	0.11665	0.00556	0.06391
mean of the estimations of $se(\hat{\beta})$	0.1180168 (0.004151345)	0.005626648 (0.0001337672)	0.06473129 (0.002564009)

5.2 Real-life data study

The G-multiplicative noise regression method is applied to a real-life data set taken from the United States Energy Information Authority. The data set can be found in the R package 'sdcMicro', and is also available from the United States Energy Information Authority website[1]. There are 4092 observations on 15 variables generally concerning income and sales data. The multiplicative noise method has no capability to protect any observations which take value "0". As mentioned in Section 1, to protect those observations, combining other methods with the multiplicative noise method is necessary. The purpose of this paper is to present the framework of the G-multiplicative noise regression method. To simplify our study at this initial stage, we do not consider how to efficiently protect observations with value "0" in this real-life data study and do not remove those observations from the data set.

Example 3. The response variable in this example was selected as 'othrevenue' (y), the revenue from sales to other consumers, whilst the explanatory variables were selected as: 'resrevenue' (x_1), the revenue from sales to residential consumers; 'ressales' (x_2), the sales to residential consumers; 'comrevenue' (x_3), the revenue from sales to commercial consumers; 'comsales' (x_4), the sales to commercial consumers; 'indsales' (x_5), the sales to industrial consumers; and 'othrsales' (x_6), the sales to other consumers. These variables were chosen because a preliminary regression analysis showed a model containing these variables to be the most appropriate extended model for the dataset.

The real data used by the OLS method gave the following fitted equation

$$\begin{aligned}\hat{y} &= 39.541929 + 0.028841x_1 - 0.002215x_2 + 0.010463x_3 \\ &\quad -0.001227x_4 + 0.009390x_5 + 0.064602x_6.\end{aligned} \tag{11}$$

The data were then masked using the multiplicative noise method to see whether the G-multiplicative noise regression method holds for real-life data. In this example, y and x_5 were independently masked by C and Z_5, respectively; both noise variables have a bimodal normal distribution with mean 145 and variance 626; x_1 was masked by

[1] http://www.eia.doe.gov/cneaf/electricity/page/eia826.html; see year 1996.

Z_1 normal distributed with mean 145 and variance 626; x_2 was masked by Z_2 Gamma distributed with mean 145 and variance 626; x_3 and x_4 were independently masked by Z_3 and Z_4, respectively, both noise variables have a uniform distribution with mean 145 and variance 626; finally, x_6 was masked by Z_6 Weibull distributed with scale 1 and shape 12. The plot of $R(\delta)$ versus δ for each multiplicative noise is presented in Fig.2. The plots are used to show the data protection capability provided by different multiplicative noises. All the ratios of the variance of noise to the square of the mean of noise are the same and equal to 0.02977408, except for $Var(Z_6)/E^2(Z_6) = 0.01024476$. But the plots of $R(\delta)$ show that some noise have better data protection capability than others. This indirectly shows that data protection capability is not unique determined by the ratio of the variance of noise to the square of the mean of noise.

There is no particular reason why we choose noises with the above different probability distributions. Basically, we want to use this example to demonstrate the application of the G-multiplicative noise regression method to real-life data and, in the meanwhile, show that any types of probability distributions, symmetric, skewness, one modal, multi-modal, or none modal, will work for the G-multiplicative noise regression method. Following the discussion in Section 2, if an intruder knows the probability distribution of multiple noise, he/she may have more chance to make a correct guess on the true values of data based on masked data. Therefore, we suggest that, when the multiplicative noise method is applied to a microdata file, combining varied types of noises with different levels of data protection may bring more uncertainty to intruders and provide more protection on the whole set of data in general.

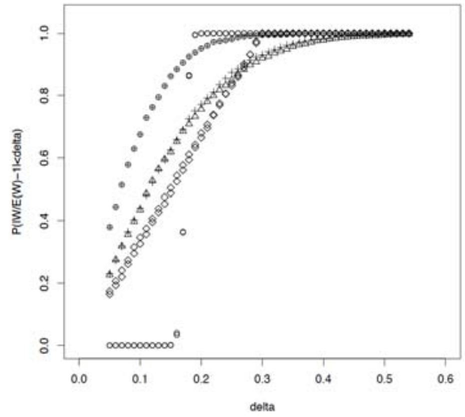

Figure 2: The plot of $P(|W/E(W) - 1| < \delta)$ vs δ. Circles are for y and x_5; triangles for x_1; crosses for x_2; diamonds for x_3 and x_4; sun crosses for x_6.

A summary on the absolute difference between the true observation and its unbiased estimation is reported in Table 2. Since the values of “0” were not removed from the original data and the multiplicative noise method does not provide any protection

Table 2: Summary statistics on the absolute difference between true observation and its unbiased estimation.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\lvert x_1 - x_1^*/E(Z_1)\rvert$	0	135.1	552.1	3022.0	2467.0	135000.0
$\lvert x_2 - x_2^*/E(Z_2)\rvert$	0	1719	7628	36710	36200	1168000
$\lvert x_3 - x_3^*/E(Z_3)\rvert$	0	70.91	372.00	2498.00	2069.00	96220.00
$\lvert x_4 - x_4^*/E(Z_4)\rvert$	0	998.7	5683.0	32330.0	32000.0	747800.0
$\lvert x_5 - x_5^*/E(Z_5)\rvert$	0	123.6	509.7	2002.0	2579.0	33390.0
$\lvert x_6 - x_6^*/E(Z_6)\rvert$	0	24.51	158.10	1784.00	989.80	172500.00
$\lvert y - y^*/EC\rvert$	0	9.323	44.210	284.700	235.700	12110.000

on value "0", the minimum of the absolute difference between true observation and its unbiased estimate provides no information on the level of protection. Since the underlying variables are skewed to the right, to assess the mean of the true observations of a variable and the mean of the unbiased estimators of the true observations of the variable, values of medians should be considered. As expected, the difference between the two medians is close to 0 (not shown in this paper). But, Table 2 shows that the medians of the **absolute of difference** between the true observation and its unbiased estimate for all variables are far away from 0. It indicates that it is not an optimal way to guess the true observation based on its unbiased estimate and the multiplicative noise method does provide reasonable protection on the data.

The G-multiplicative noise regression method was independently applied to the data 1000 times. The sample means and sample standard errors for all the estimated parameters as well as the estimates of the standard errors of the OLS estimators evaluated by $\hat{\sigma}^2 A^{-1}$ are reported in Table 3, where

$$\hat{\sigma}^2 = \frac{1}{n-6}\sum_{i=1}^{n}(y_i - \hat{\beta}_0^{(n)} - \hat{\beta}_1^{(n)} x_{i,1} - \cdots - \hat{\beta}_6^{(n)} x_{i,6})^2$$

$$\approx \frac{1}{n-6}(1, \hat{\beta}_0^{(n)}, \cdots, \hat{\beta}_6^{(n)}) D (1, \hat{\beta}_0^{(n)}, \cdots, \hat{\beta}_6^{(n)})',$$

$$D = \begin{pmatrix} \sum_{i=1}^n \frac{y_i^{*2}}{E(C^2)} & -\sum_{i=1}^n \frac{y_i^*}{E(C)} & -\sum_{i=1}^n \frac{y_i^* x_{i,1}^*}{E(C)E(Z_1)} & \cdots & -\sum_{i=1}^n \frac{y_i^* x_{i,6}^*}{E(C)E(Z_6)} \\ -\sum_{i=1}^n \frac{y_i^*}{E(C)} & n & -\sum_{i=1}^n \frac{x_{i,1}^*}{E(Z_1)} & \cdots & -\sum_{i=1}^n \frac{x_{i,6}^*}{E(Z_6)} \\ -\sum_{i=1}^n \frac{y_i^* x_{i,1}^*}{E(C)E(Z_1)} & \sum_{i=1}^n \frac{x_{i,1}^*}{E(Z_1)} & \sum_{i=1}^n \frac{x_{i,1}^{*2}}{E(Z_1^2)} & \cdots & -\sum_{i=1}^n \frac{x_{i,1}^* x_{i,6}^*}{E(Z_1)E(Z_6)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\sum_{i=1}^n \frac{y_i^* x_{i,6}^*}{E(C)E(Z_6)} & \sum_{i=1}^n \frac{x_{i,6}^*}{E(Z_6)} & \sum_{i=1}^n \frac{x_{i,1}^* x_{i,6}^*}{E(Z_6^2)E(Z_1^2)} & \cdots & -\sum_{i=1}^n \frac{x_{i,1}^* x_{i,6}^{*2}}{E(Z_6^2)} \end{pmatrix}$$

and A is defined in Theorem 3. Table 3 shows that the basic regression analysis information can be well obtained by using the G-multiplicative noise regression method without accessing the true data.

Table 3: Table of regression parameter estimation accuracy and reliability. The second column contains the OLS estimates of parameters based on the true data. The third column contains the sample means of the parameter estimates based on masked data. The fourth column contains the standard errors for OLS estimators based on the true data and the fifth column contains the means of the estimates of $se(\hat{\beta}_{OLS})$ based on masked data. The sample standard errors are listed in brackets.

Parameter	$\hat{\beta}_{OLS}$	mean of $\hat{\beta}^{(n)}$	$se(\hat{\beta}_{OLS})$	mean of estimations $se(\hat{\beta}_{OLS})$
β_0	39.541929	44.491090 (316.707400)	21.283367	20.903840 (2.852181)
β_1	0.028841	0.030303 (0.229575)	0.003180	0.003683 (0.002013)
β_2	-0.002215	-0.002289 (0.019781)	0.000255	0.000298 (0.000170)
β_3	0.010463	0.010560 (0.271125)	0.003008	0.003501 (0.001974)
β_4	-0.001227	-0.001308 (0.021121)	0.000268	0.000303 (0.000150)
β_5	0.009390	0.008889 (0.017454)	0.001448	0.001439 (0.000202)
β_6	0.064602	0.064666 (0.008127)	0.000276	0.000274 (4.413135e-05)

Conclusion

The applications of the G-multiplication noise regression method in Section 5 showed that the increase of the sample variance for the estimate of β_0 is larger than that for the estimates of other parameters after the original data are replaced by masked data.

The estimation of regression parameters under the G-multiplicative noise regression method can be further improved based on the asymptotically unbiased results in Section 4 if statistical agencies adopt the following practice.

Consider the following scenario which is used to explain the practice we are suggesting. Assume that a data user wants to carry out linear regression analysis on a set of confidential data and the data owner, a statistical agency, has obligations to provide information on the confidential data to the data user, but by law the agency should not allow the data user to access the original confidential data. Assume that the agency has decided a masking scheme for the data set, e.g., the multiplicative noises applied to each variable are decided. The practice involves two parts.

Part I. Multiple Masked datasets preparation. Independently apply the masking scheme to the underlying data set n times and obtain n sets of masked datasets. To avoid the observations of each variable that can be estimated from the n sets of masked datasets, the statistical agency has to independently randomly assign identification codes to each masked dataset before the n masked datasets are issued to the data user.

The process used to generate the multiple datasets in Part I is different from that used to generate multiple synthetic datasets discussed in the literature. In Part I, we do not simulate any data from the population distribution of the original underlying dataset and only simulate datasets from multiplicative noise, which are not related to the population distribution of the original underlying dataset.

Part II. Data analysis. After receiving the n masked datasets, Step 1, apply formulae (7) or (8) and $\hat{\sigma}^2 A^{-1}$ for each masked data set and obtain the estimates of regression parameters and the estimates of the standard errors of the OLS estimators, respectively; Step 2, for each parameter, calculate the sample mean of the estimates of the parameter and the sample mean of the estimates of standard error of the OLS estimator of the parameter. The final estimate of parameter and the estimate of the standard error of OLS estimator are given by these sample means.

In the multiplicative noise method, masked data as well as the means and variances of multiplicative noises are issued to the public. Based on the data protection discussed in Section 2, releasing the means and variances of multiplicative noises will not provide much help for data analysts in detecting the true values of data, especially when the type of distributions of multiplicative noises are not made public. However, publishing the information of the means and variances of multiplicative noises might still make statistical agencies uncomfortable. This issue could be solved by a software engineer as described in Section 2.

Acknowledgments

The authors would like to thank Natalie Shlomo, reviewers and the editor for their suggestions and comments on the original version of this work. The authors would also like to thank David Steel and Carole Birrell for their editorial contribution to the manuscript.

Appendix A: Data protection evaluated by the relative error measurement

In the following we discuss the protection of y evaluated by relative error measurement. Since the multiplicative noise method does not provide any protection on observations with value 0, we do not consider the scenario $y = 0$.

Choose random noise C satisfying conditions $E(C) > 0$ and

$$P(|\frac{C}{E(C)} - 1| < 0.05) \leq \alpha < 1, \text{ for a real number } \alpha > 0. \tag{12}$$

For this C, there is a $\delta > 0.05$ such that $P(|C/E(C) - 1| < \delta) = 0.9999$. Let

$$\delta_0 = \min_{\delta}\{\delta \mid P(|\frac{C}{E(C)} - 1| < \delta) = 0.9999\}, \tag{13}$$

i.e., $P(|(y^*/E(C) - y)/y| \leq \delta_0) = P(|C/E(C) - 1| \leq \delta_0) \approx 1$. Therefore, regardless of the probability 0.0001,

$$(1 - \delta_0)y \leq \frac{y^*}{E(C)} \leq (1 + \delta_0)y, \qquad \text{for } y > 0$$

and

$$(1 + \delta_0)y \leq \frac{y^*}{E(C)} \leq (1 - \delta_0)y, \qquad \text{for } y < 0.$$

To simplify our discussion, we further assume $y > 0$. The discussion for $y < 0$ can be similarly followed and is not given here.

(a) If $\delta_0 \geq 1$, the true value of y is allocated anywhere within the interval $[max\{\frac{y^*/E(C)}{1+\delta_0}, 0\}, \infty)$. The length of this interval is too large for an analyst to correctly identify the true value y from the interval.

(b) If $\delta_0 < 1$, the true value of y is allocated anywhere within $[\frac{y^*/E(C)}{1+\delta_0}, \frac{y^*/E(C)}{1-\delta_0}]$ except for very small probability 0.0001, given $y^*/E(C)$ is known.

What kind of strategy is used to guess the true value y from this interval will depend on an intruder's personality and their knowledge of y. An intruder might always choose the midpoint of the interval $[\frac{y^*/E(C)}{1+\delta_0}, \frac{y^*/E(C)}{1-\delta_0}]$, i.e., $\tilde{y} = y^*/[E(C)(1-\delta_0^2)]$, as a guess for y, or he/she might always choose the value at the qth percentile position in the interval, i.e.,

$$\tilde{y}_{(q)} = \frac{y^*/E(C)}{1+\delta_0} + q\frac{2\delta_0 y^*/E(C)}{1-\delta_0^2}, \;\; 0 < q < 1, \delta_0 < 1.$$

An intruder might use other rules to make his/her guess on the value of y. Basically, if the guessing strategy is unknown, it will be difficult to evaluate data protection.

If an intruder guesses the true value of y by $\tilde{y}_{(q)}$, the probability that $\tilde{y}_{(q)}$ is accepted as a correct guess is evaluated as

$$P(-0.05 < \frac{\tilde{y}_{(q)} - y}{y} < 0.05) = P(0.95\frac{(1-\delta_0^2)}{1-(1-2q)\delta_0} < \frac{C}{E(C)} < 1.05\frac{(1-\delta_0^2)}{1-(1-2q)\delta_0}). \tag{14}$$

If the δ_0 given by C is close to 0, then (14) will be close to $P(0.95 < C/E(C) < 1.05) = P(|C/E(C) - 1| < 0.05)$; if δ_0 is close to 1, (14) will be close to 0 if C is a continuous random variable. The value of (14) strongly depends on the properties of the probability distribution of C, i.e., whether the probability density function of C is symmetric or nonsymmetric; is single mode, or multiple modes. Basically, a noise C with lower $P(|C/E(C)-1| < 0.05)$ and higher δ_0 will provide better protection of data if intruders use the above strategy to guess the true values of data.

We present an example below to show which multiplicative noise enables us to provide more protection on data in terms of the value of $P(|C/E(C) - 1| < \delta)$.

Example A1. Two types of probability distributions for C are considered. The first one has probability distribution $(N(a, 1) + N(b, 1))/2$, which is the average of two normal distributions. The density function given by this distribution has two modes, named bi-modal normal density function. The noise is centered at $E(C) = (a+b)/2$ with $Var(C) = 1 + (a - b)^2/4$. The second one has normal density function with the same mean $(a + b)/2$ and variance $1 + (a - b)^2/4$. Therefore, both probability distributions give the same ratio of $Var(C)$ to $E^2(C)$, i.e., $[4 + (a - b)^2]/(a + b)^2$.

We compare the values of $P(|C/E(C) - 1| < \delta)$ given by the two probability distributions for different δ values and two sets of (a, b).

Fig.3 shows that the values of $P(|C/E(C)-1| < 0.05)$ are much less than 1 for all the underlying multiplicative noises. It indicates that all the multiplicative noises possess a certain level of capability in protecting data. Since the values of $P(|C/E(C)-1| < 0.05)$ given by bi-modal distributions are lower than those given by normal distributions, it indicates that the multiplicative noise with bi-modal distribution is more powerful than the noise with a normal distribution in protecting data, while both the distributions have the same ratio of $Var(C)$ to $E^2(C)$. As mentioned in Section 2.1, different intruders may use a different upper bound of relative difference to define his/her acceptance rule. The value of the upper bound of relative difference is unknown by the data provider. Therefore, to reduce the risk of data identification, the data provider should choose a multiplicative noise C such that $P(|C/E(C)-1| < \delta)$ takes smaller values and the speed of the value $P(|C/E(C)-1| < \delta)$ increasing to 1 is slower as δ increases. The level of data protection can be improved through choosing an appropriate type of distribution and parameter(s) for the multiplicative noise. Based on the above rule, in this example, using the pair parameters ($a = 12$, $b = 19$) in either normal or bi-modal normal distributions will provide better protection of data than using ($a = 170$, $b = 120$) in terms of a smaller value of $P(|C/E(C) - 1| < 0.05)$ and a larger value of δ_0.

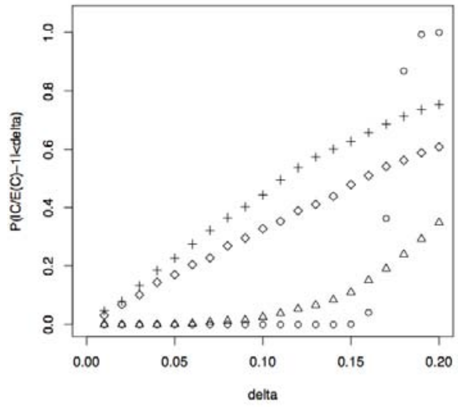

Figure 3: The probability $P[|C/E(C) - 1| < \delta]$ where C has mean $(a+b)/2$ and variance $1 + (a-b)^2/4$. The crosses and circles represent C having normal distribution and bi-modal normal distribution with $a = 170$, $b = 120$, and $Var(C)/E^2(C) = 0.02977408$, respectively. The diamonds and triangles represent C having normal distribution and bi-modal normal distribution with $a = 12$, $b = 19$, and $Var(C)/E^2(C) = 0.0551508$, respectively.

Appendix B: $R - U$ plot

Example A2 below shows that the level of protection provided by C_4 (defined in Section 2.2, Example 1) with $a_1 = 2150$ is still acceptable.

Example A2. A sample of size $n = 500$ was simulated for y from a uniform distribution $U(50, 250)$. The values of y were then masked by C_4 to obtain y^*. The plot of $y^*/E(C_4)$ versus y in Fig.4 clearly shows that it is very difficult to correctly guess the true value of y if only $y^*/E(C)$ is available. This occurs especially when $R(0.05)$ is much less than 0.5. For example, if $y^*/E(C) = 150$, many different values of y between 110 to 200 have the similar masked value.

Appendix C: Comparison between the methods of additive noise and multiplicative noise

In the following, we use a simulation example to demonstrate that, for a given additive noise $\tilde{C}$, there is a multiplicative noise C such that (i) C has the same type of probability distribution as $\tilde{C}$ and (ii) C is able to provide better protection of data than $\tilde{C}$ in terms of the relative error measurement.

Example A3. Consider a sample $\{y_i\}_{i \leq 500}$ simulated from the uniform distribution $U(50, 250)$. Let $\tilde{C} \sim (\sum_{i=1}^{4} N(a_{add,i}, 1))/4$, where $a_{add,1} = -3d/2$, $a_{add,i} = a_{add,1} + (i-1)d$ and d is a positive real number, $i = 2, 3, 4$. Obviously $E(\tilde{C}) = 0$ regardless of the

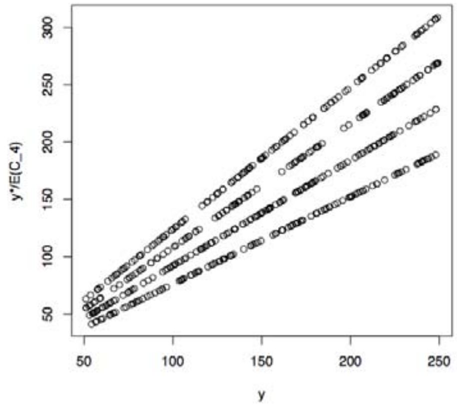

Figure 4: The plot of $y^*/E(C_4)$ vs y. Data variation cannot be shown due to the lower resolution of the image.

value of d. We choose $d = 1.003992$ in this example and give $Var(\hat{Y}_{add} - y) = \sigma^2_{add} = 4.6^2$ (see Section 2.3). This value for the variance of additive noise has been used in one study by Duncan and Mukherjee (2000).

Now we conduct a multiplicative noise C such that the variance of C is the same as $\sigma^2_{add} = 4.6^2$. By noting that $Var(\hat{Y}_{multi} - y|y) = [Var(C)/E^2(C)]y^2$ (see Section 3.2), the value of $Var(\hat{Y}_{multi} - y|y)$ will depend on $E(C)$, given $Var(C)$ and y are fixed. The ratio of $Var(C)$ to $E^2(C)$ is an issue in the multiplicative noise method, which is related to whether too much noise had been added into the underlying data by the multiplicative noise method. In one study of the multiplicative noise method carried out by the Energy Information Administration in U.S. Department of Energy (Kim and Winkler, 2003), the ratio is suggested as 0.0225. To meet these restrictions, we let the noise $C \sim (\sum_{i=1}^{4} N(a_{multi,i}, 1))/4$, where $a_{multi,1} = 24.64271$, $a_{multi,i} = a_{multi,1} + (i-1) \times 4.015968$, $i = 2, 3, 4$. Thus, $E(C) = 30.66668$, $Var(C) = 4.6^2$, and the ratio of $Var(C)$ to $E^2(C)$ is 0.0225.

Now both noises, $\tilde{C}$ and C, can be considered as acceptable noises in practice in terms of $\sigma^2_{add} = 4.6^2$ and $Var(C)/E^2(C) = 0.0225$, and both noises are four-modal distributed. The true data $\{y_i\}_{i\leq 500}$ are independently masked by $\tilde{C}$ and C which produce $\{\tilde{y}_i\}_{i\leq 500}$ and $\{y^*\}_{i\leq 500}$, respectively.

Since $\tilde{C}$ is an acceptable noise, it is expected that $\tilde{C}$ will provide reasonable protection on the data. However, the mean of square errors given by the additive noise method is $\sum_{i=1}^{500}(\hat{y}_{add,i} - y_i)^2/500 = 2.20003$. It is much smaller than the mean of square errors $\sum_{i=1}^{500}(\hat{y}_{multi,i} - y_i)^2/500 = 582.5439$ given by the multiplicative noise method. It clearly shows that, on average, the multiplicative noise method provides more protection on the data than the additive noise method, although both noise masking schemes are accepted in practice. The plot of disclosure risk $R(\delta)$ for both methods are presented in Fig.5. It shows that, (i) the $R(0.05)$ given by C is much smaller than that given by $\tilde{C}$; (ii) as δ increases, $R(\delta)$ given by $\tilde{C}$ increases to 1 much faster than that given by C. Based on the discussion in Section 2.1, it further supports that C provides more protection on data than $\tilde{C}$ in terms of relative error measurement.

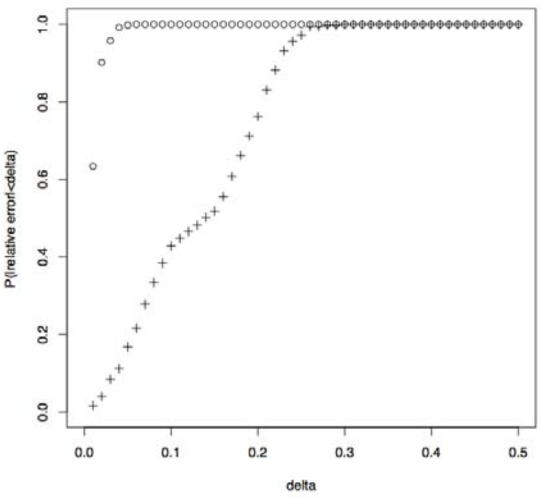

Figure 5: The plot of $P(|(\hat{y}_{add} - y)/y| < \delta)$ is in circles and the plot of $P(|(\hat{y}_{multi} - y)/y| < \delta)$ is in crosses.

Appendix D: The proof of Theorem 2

The proof of Theorem 2. Since $\mathbf{y}^* = C^{(n)}\mathbf{y} = C^{(n)}X\beta + C^{(n)}\epsilon$, where $C^{(n)} = diag(c_1, c_2, \cdots, c_n)$, the OLS estimator of β given by the above model is

$$\hat{\beta}_{OLS}^{(n)} = [X'(C^{(n)})'(C^{(n)})X]^{-1}(X'(C^{(n)})'\mathbf{y}^*) = A_n^{-1}W_n,$$

where

$$A_n = \frac{1}{b_n}\begin{pmatrix} \sum_{i=1}^n c_i^2 & \sum_{i=1}^n c_i^2 x_{i,1} & \sum_{i=1}^n c_i^2 x_{i,2} & \cdots & \sum_{i=1}^n c_i^2 x_{i,p} \\ \sum_{i=1}^n c_i^2 x_{i,1} & \sum_{i=1}^n c_i^2 x_{i,1}^2 & \sum_{i=1}^n c_i^2 x_{i,1}x_{i,2} & \cdots & \sum_{i=1}^n c_i^2 x_{i,1}x_{i,p} \\ \sum_{i=1}^n c_i^2 x_{i,2} & \sum_{i=1}^n c_i^2 x_{i,2}x_{i,1} & \sum_{i=1}^n c_i^2 x_{i,2}^2 & \cdots & \sum_{i=1}^n c_i^2 x_{i,2}x_{i,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n c_i^2 x_{i,p} & \sum_{i=1}^n c_i^2 x_{i,p}x_{i,1} & \sum_{i=1}^n c_i^2 x_{i,p}x_{i,2} & \cdots & \sum_{i=1}^n c_i^2 x_{i,p}^2 \end{pmatrix}$$

and

$$W_n = \frac{1}{b_n}\left(\begin{array}{c} \sum_{i=1}^n c_i y_i^* \\ \sum_{i=1}^n x_{i,1} c_i y_i^* \\ \sum_{i=1}^n x_{i,2} c_i y_i^* \\ \vdots \\ \sum_{i=1}^n x_{i,p} c_i y_i^* \end{array}\right).$$

From Corollary 1, we have

(a) $\frac{1}{b_n}\sum_{k=1}^n c_k^2 x_{k,j} - \frac{E(C^2)}{b_n}\sum_{k=1}^n x_{k,j} \to 0$ with probability 1, as $n \to \infty$;

(b) $\frac{1}{b_n}\sum_{k=1}^n c_k y_k^* - \frac{E(C^2)}{b_n E(C)}\sum_{k=1}^n y_k^* \to 0$, with probability 1, as $n \to \infty$;

(c) $\frac{1}{b_n}\sum_{k=1}^n c_k^2 x_{k,i} x_{k,j} - \frac{E(C^2)}{b_n}\sum_{k=1}^n x_{k,i} x_{k,j} \to 0$ with probability 1, as $n \to \infty$;

(d)$\frac{1}{b_n}\sum_{k=1}^n c_k x_{k,i} y_k^* - \frac{E(C^2)}{b_n E(C)}\sum_{k=1}^n x_{k,i} y_k^* \to 0$, with probability 1, as $n \to \infty$.

In the following, we only give the proof of (d). For (d),

$$\frac{1}{b_n}\sum_{k=1}^n c_k x_{k,i} y_k^* - \frac{E(C^2)}{b_n E(C)}\sum_{k=1}^n x_{k,i} y_k^*$$

$$= [\frac{1}{b_n}\sum_{k=1}^n c_k^2 x_{k,i} y_k - \frac{1}{b_n}E(C^2)\sum_{k=1}^n x_{k,i} y_k] + \frac{E(C^2)}{E(C)}[\frac{1}{b_n}E(C)\sum_{k=1}^n x_{k,i} y_k - \frac{1}{b_n}\sum_{k=1}^n c_k x_{k,i} y_k].$$

Apply Corollary 1 to

$$(1/b_n)[\sum_{k=1}^n c_k^2 x_{k,i} y_k - E(C^2)\sum_{k=1}^n x_{k,i} y_k] \quad \text{and} \quad (1/b_n)[E(C)\sum_{k=1}^n x_{k,i} y_k - \sum_{k=1}^n c_k x_{k,i} y_k].$$

Therefore,

$$\frac{1}{b_n}\sum_{k=1}^n c_k x_{k,i} y_k^* - \frac{E(C^2)}{b_n E(C)}\sum_{k=1}^n x_{k,i} y_k^* \to 0, \qquad \text{with probability 1,}$$

as $n \to \infty$, $i, j = 1, 2, \cdots, p$.

From conditions (i), (ii) and (a)–(d) above,

$$\begin{aligned} &\hat{\beta}_{OLS}^{(n)} - [(X'X)^{-1}X'\mathbf{y}^*]/E(C) \\ &= \left\{[\frac{1}{b_n}X'C^{(n)}C^{(n)}X]^{-1} - [\frac{E(C^2)}{b_n}X'X]^{-1}\right\}[\frac{1}{b_n}X'C^{(n)}\mathbf{y}^* - \frac{E(C^2)}{b_n}X'\mathbf{y}] \\ &\quad +[(\frac{1}{b_n}X'C^{(n)}C^{(n)}X)^{-1} - (\frac{E(C^2)}{b_n}X'X)^{-1}]\frac{E(C^2)}{b_n}X'\mathbf{y} \\ &\quad +(\frac{E(C^2)}{b_n}X'X)^{-1}[\frac{1}{b_n}X'C^{(n)}\mathbf{y}^* - \frac{E(C^2)}{b_n E(C)}X'\mathbf{y}^*] \to 0, \end{aligned}$$

with probability 1 as $n \to \infty$, as required.

Appendix E: The proof of Theorem 4

To prove Theorem 4, we need the following results.

Following the technique used in the proof of Theorem 2, we are able to show

$$\frac{1}{n}\left[\frac{\sum_{i=1}^{n} x_{i,j}^*}{E(Z_j)} - \sum_{i=1}^{n} x_{i,j}\right] \to 0, \quad \frac{1}{n}\left[\frac{\sum_{i=1}^{n} x_{i,j}^* x_{i,k}^*}{E(Z_j)E(Z_k)} - \sum_{i=1}^{n} x_{i,j}x_{i,k}\right] \to 0$$

with probability 1 for $j \neq k$ and $j, k = 1, 2, \cdots, p$. Thus,

$$\frac{1}{n}A - \frac{1}{n}X'X \to 0 \tag{15}$$

with probability 1 and

$$A^{-1} - (X'X)^{-1} = \frac{1}{n}(\frac{1}{n}A)^{-1}(\frac{1}{n}X'X - \frac{1}{n}A)(\frac{1}{n}X'X)^{-1} \to 0 \tag{16}$$

with probability 1 if $\|(\frac{1}{n}A)^{-1}\|$ and $\|(\frac{1}{n}X'X)^{-1}\|$ are bounded. These up bounded conditions can be easily satisfied in practice. Following the same technique, we also have

$$\frac{1}{n}B'B - \frac{1}{n}[X'X - D] \to 0$$

with probability 1, where

$$D = diag(0, \sum_{i=1}^{n} x_{i,1}^2[1 - E(Z_1^2)/E^2(Z_1)], \cdots, \sum_{i=1}^{n} x_{i,p}^2[1 - E(Z_p^2)/E^2(Z_p)]). \tag{17}$$

Using (15)-(17), if $E^2(Z_i) >> E(Z_i^2)$, i.e., $E(Z_i^2)/E^2(Z_i) \approx 1$, $i = 1, 2, \cdots, p$, we have

$$diag(0, \sum_{i=1}^{n} x_{i,1}^2(1 - E(Z_1^2)/E^2(Z_1)), \cdots, \sum_{i=1}^{n} x_{i,p}^2(1 - E(Z_p^2)/E^2(Z_p))) \approx \mathbf{0}$$

and $A^{-1}B'BA^{-1} \approx (X'X)^{-1}$, with probability 1, as $n \to \infty$, subject to $\|(\frac{1}{n}A)^{-1}\|$ and $\|(\frac{1}{n}X'X)^{-1}\|$ are bounded.

The proof of Theorem 4: Rewrite

$$\sqrt{n}(\hat{\beta}^{(n)} - A^{-1}(X'X)\beta) = (\frac{1}{n}A)^{-1}[\frac{1}{\sqrt{n}E(C)}B'y^* - \frac{1}{\sqrt{n}}(X'X)\beta].$$

We have

$$\frac{1}{\sqrt{n}E(C)}B'y^* - \frac{1}{\sqrt{n}}(X'X)\beta$$

$$= \frac{1}{\sqrt{n}E(C)}\{B'diag(c_1, \cdots, c_n)y - [X'diag(E(C), \cdots, E(C))X]\beta\}$$

$$= \frac{1}{\sqrt{n}E(C)}\{[B'diag(c_1, \cdots, c_n) - X'diag(E(C), \cdots, E(C))]X\beta + B'diag(c_1, \cdots, c_n)\epsilon\}.$$

Denote

$$B'diag(c_1, \cdots, c_n) - X'diag(E(C), \cdots, E(C)) = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n)$$

with $\mathbf{w}_i = (c_i - E(C), \frac{C_i x^*_{i,1}}{E(Z_1)} - E(C)x_{i,1}, \cdots, \frac{C_i x^*_{i,p}}{E(Z_p)} - E(C)x_{i,p})'$, and $B' = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n)$, with $\mathbf{b}_i = (1, x^*_{i,1}/E(Z_1), \cdots, x^*_{i,p}/E(Z_p))'$. Random vectors $\{\mathbf{w}_i\}$ and $\{\mathbf{b}_i\}$ are mutually independent and have 0 mean, respectively. Thus, the characteristic function of $\frac{1}{\sqrt{n}E(C)}B'y^* - \frac{1}{\sqrt{n}}(X'X)\beta$ is

$$\phi_n(\mathbf{t}) = E\{exp[\frac{i\mathbf{t}'}{\sqrt{n}E(C)}\sum_{i=1}^{n}(\mathbf{w}_i \sum_{j=0}^{p} x_{i,j}\beta_j + \mathbf{b}_i c_i \epsilon_i)]\}$$

$$= \Pi_{i=1}^{n}\{1 - \frac{\mathbf{t}'}{2nE^2(C)}[(\sum_{j=0}^{p} x_{i,j}\beta_j)^2 E(\mathbf{w}_i\mathbf{w}_i') + E(c_i^2\epsilon_i^2\mathbf{b}_i\mathbf{b}_i')]\mathbf{t} + o(n^{-1})\}$$

where

$$E(\mathbf{w}_i\mathbf{w}_i')$$

$$= Var(C)\begin{pmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \\ x_{i,1} & [\frac{E(C^2)E(Z_1^2)}{E^2(Z_1)} - E^2(C)]\frac{x_{i,1}^2}{Var(C)} & \cdots & x_{i,1}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,p} & x_{i,1}x_{i,p} & \cdots & [\frac{E(C^2)E(Z_p^2)}{E^2(Z_p)} - E^2(C)]\frac{x_{i,p}^2}{Var(C)} \end{pmatrix}$$

and

$$E[(c_i\epsilon_i)^2\mathbf{b}_i\mathbf{b}_i'] = E(C^2)\sigma^2 \begin{pmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \\ x_{i,1} & x_{i,1}\frac{E(Z_1^2)}{E^2(Z_1)} & \cdots & x_{i,1}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,p} & x_{i,1}x_{i,p} & \cdots & x_{i,p}\frac{E(Z_p^2)}{E^2(Z_p)} \end{pmatrix}.$$

Therefore,

$$\log\phi_n(\mathbf{t}) = \sum_{i=1}^{n}\log\{1 - \frac{\mathbf{t}'}{2nE^2(C)}[(\sum_{j=0}^{p} x_{i,j}\beta_j)^2 E(\mathbf{w}_i\mathbf{w}_i') + E(c_i^2\epsilon_i^2\mathbf{b}_i\mathbf{b}_i')]\mathbf{t} + o(n^{-1})\}$$

$$= -\frac{1}{2nE^2(C)}\mathbf{t}'[\sum_{i=1}^{n}(\sum_{j=0}^{p} x_{i,j}\beta_j)^2 E(\mathbf{w}_i\mathbf{w}_i') + E(C^2)\sigma^2\sum_{i=1}^{n}E(\mathbf{b}_i\mathbf{b}_i')]\mathbf{t} + o(1).$$

From conditions (i)–(iii), we have

$$\frac{1}{\sqrt{n}E(C)}B'y^* - (X'X)\beta \xrightarrow{D} N(0, \frac{Var(C)}{E^2(C)}Q_4 + \frac{E(C^2)}{E^2(C)}\sigma^2 Q_3)$$

and

$$\sqrt{n}(\hat{\beta} - A^{-1}X'X\beta) = (\frac{1}{n}A)^{-1}[\frac{1}{\sqrt{n}E(C)}B'y^* - (X'X)\beta]$$

$$\xrightarrow{D} Q_1^{-1}N(0, \frac{Var(C)}{E^2(C)}Q_4 + \frac{E(C^2)}{E^2(C)}\sigma^2 Q_3), \quad \text{as } n \to \infty.$$

Appendix F: Simulation examples

Example A4. We use this example to show the impact of the sample size and the distribution of noise on the final regression analysis.

Assume that true data with size 1000 and 2000 were simulated respectively from

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon = 2 + 1.5X_1 + 3X_2 + \varepsilon,$$

where X_1 and X_2 were simulated from uniform(0,20) and uniform(3,40), respectively, and $\varepsilon \sim N(0,1)$. Let Y, X_1, and X_2 be masked by C, Z_1, and Z_2, respectively. The estimator $\hat{\beta}^{(n)}$ of $\beta = (\beta_0, \beta_1, \beta_2)$ was given by (8) in Section 3.

Apply the multiplicative noise method to each data set 1000 times, respectively. Seven different types of masking schemes are considered in this example. In the first six masking schemes, we assigned all the noise, C, Z_1, and Z_2, are i.i.d. and have the same distribution. In the last masking scheme, two types of noises are involved. The estimated values of β based on different masking schemes and different sizes of the sample are reported in Table 4.

Table 4: Results for Example A4. Ratio $Var(C)/E^2(C)$ and approximate δ_0 (see Section2.1) are presented in the second column. The four-modal multiplicative noise is defined in Example 1 with $a_i = a_1 + (i-1)d$ and $d = 450$, $i = 2, 3, 4$.

		Size 1000		
	δ_0,ratio	$\hat{\beta}_0^{(n)}$ (se.)	$\hat{\beta}_1^{(n)}$ (se.)	$\hat{\beta}_2^{(n)}$ (se.)
Bimodal $a = 170$, $b = 120$	$\delta_0 \approx 0.18$ 0.029774	2.101065 (1.686983)	1.496296 (0.118900)	2.998384 (0.062488)
Normal $N(145, 626)$	$\delta_0 \approx 0.44$ 0.029774	1.865960 (1.757619)	1.506336 (0.117061)	3.006882 (0.068524)
four-modal $a_1 = 150 + 18 \times 100$	$\delta_0 > 0.6$ 0.036735	1.827599 (1.946144)	1.511865 (0.135546)	3.006121 (0.073774)
Bimodal $a = 12$, $b = 19$	$\delta_0 \approx 0.58$ 0.055150	1.799981 (2.445379)	1.512276 (0.171112)	3.006960 (0.088390)
Normal $N(15.5, 13.25)$	$\delta_0 > 0.6$ 0.055150	1.768212 (2.609527)	1.508968 (0.168562)	3.010818 (0.098706)
four-modal $a_1 = 150 + 8 \times 100$	$\delta_0 \approx 0.24$ 0.095858	1.621189 (3.648085)	1.526405 (0.242774)	3.012771 (0.130503)
X_1 and X_2, bimodal $a = 170$, $b = 120$ Y, four-modal $a_1 = 150 + 18 \times 100$		1.951158 (1.806469)	1.504587 (0.128307)	3.005366 (0.069712)
		Size 2000		
	δ_0,ratio	$\hat{\beta}_0^{(n)}$ (se.)	$\hat{\beta}_1^{(n)}$ (se.)	$\hat{\beta}_2^{(n)}$ (se.)
Bimodal $a = 170$, $b = 120$	$\delta_0 \approx 0.18$ 0.029774	1.98265 (1.058435)	1.501076 (0.082355)	3.000109 (0.045061)
Normal $N(145, 626)$	$\delta_0 \approx 0.44$ 0.029774	2.00303 (1.156801)	1.501282 (0.081986)	2.998305 (0.045389)
four-modal $a_1 = 150 + 18 \times 100$	$\delta_0 > 0.6$ 0.036735	2.045022 (1.283234)	1.498827 (0.095387)	2.99728 (0.052755)
Bimodal $a = 12$, $b = 19$	$\delta_0 \approx 0.58$ 0.055150	2.024373 (1.581899)	1.502247 (0.121158)	2.996331 (0.063052)
Normal $N(15.5, 13.25)$	$\delta_0 > 0.6$ 0.055150	1.986769 (1.708734)	1.562076 (0.117172)	2.998596 (0.065870)
four-modal $a_1 = 150 + 8 \times 100$	$\delta_0 \approx 0.24$ 0.095858	2.025701 (2.401262)	1.498422 (0.172018)	2.998039 (0.093221)
X_1 and X_2, bimodal $a = 170$, $b = 120$ Y, four-modal $a_1 = 150 + 18 \times 100$		2.079868 (1.170097)	1.493703 (0.088023)	2.998082 (0.048445)

The results show that the standard error of the parameter estimates will decrease as sample size increases. A noise with a larger ratio of the variance to the square of the mean tends to give larger standard error of the estimator of the parameter. In terms of protecting data, we tend to chose a noise with larger δ_0. In this example, a distribution with larger δ_0 always shows a larger ratio. Therefore, to decide a masking scheme for a variable, a balance between δ_0 and the ratio needs to be considered. The last masking scheme in Table 4 shows that variables masked by different noises might decrease the standard errors of estimators and maintain the protection level for some variables. If multiple masked datasets are used (see Part I in Conclusion), the means of estimates of parameters are always close to the true values of parameters no matter which masking scheme was used. Therefore, using a noise with higher protection level is an option.

References

[1] Bethlehem, J., Keller, W., and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association*, 85:38–45.

[2] Brand, R. (2002). Microdata protection through noise addition. In *Inference Control in Statistical Databases*, vol. 2316 of *LNCS*. Springer Berlin Heidelberg. 61–74.

[3] Csorgo, M. (1968). On the Strong Law of Large Numbers and the Central Limit Theorem for Martingales, *Transactions of the American Mathematical Society*, 13:259–275.

[21] Domingo-Ferrer, J., Sebé, F., and Castellà-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. In J. Domingo-Ferrer and V. Torra (eds.), *Privacy in Statistical Databases*, vol. 3050 of *LNCS*. Springer-Verlag Berlin Heidelberg. 149–161.

[6] Duncan, G.T. and Lambert, D. (1986). Disclosure limited data dissemination (with comment), *Journal of the American Statistical Association*, 81:10–28.

[6] Duncan, G.T. and Lambert, D. (1989). The risk of disclosure for microdata, *Journal of Business and Economic Statistics*, 7:207–217.

[7] Duncan, G.T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise, *Journal of the American Statistical Association*, 95:720–729.

[8] Duncan, G.T., Keller-McNulty, S.A., and Stokes, S.L. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map, Technical Report 121, National Institute of Statistical Sciences.

[9] Duncan, G.T., Keller-McNulty, S.A., and Stokes, S. L. (2004). Database Security and Confidentiality: Examining Disclosure risk vs. Data utility through the R-U Confidentiality Map, Technical Report Number 142, National Institute of Statistical Sciences.

[10] Elamir, E. and Skinner, C.J. (2006). Record-level measures of disclosure risk for survey micro-data, *Journal of Official Statistics*, 22:525–539.

[12] Evans, T. (1996). Effects on Trend Statistics of the Use of Multiplicative Noise for Disclosure Limitation, U.S. Bureau of the Census, `http://www.census.gov/srd/sdc/papers.html`, accessed 5/12/2008.

[12] Evans, T., Zayatz, L., and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data, *Journal of Official Statistics*, 14:537–551.

[13] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. II. New York: Wiley.

[14] Fuller, W. (1993). Masking procedures for microdata disclosure limitation, *Journal of Official Statistics*, 9:383–406.

[15] Gouweleeuw, J., Kooiman, P., Willenborg, L., and De Wolf, P.P. (1998). Post randomisation for statistical disclosure control: Theory and implementation, *Journal of Official Statistics*, 14:463–478.

[16] Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy, *Journal of the American Statistical Association*, 81:680–688.

[17] Karr, A., Lin, X., Sanil, A. and Reiter, J. (2005). Secure regression on distributed databases, *Journal of Computational and Graphical Statistics*, 14:263–279.

[18] Karr, A., Lin, X., Sanil, A., and Reiter, J. (2006). Secure statistical analysis of distributed databases. In A. Wilson, G. Wilson, and D. Olwell (eds.), *Statistical Methods in Counterterrorism: Game Theory, Modelling, Syndromic Surveillance, and Biometric Authentication*. New York: Springer. 237–262.

[19] Karr, A., Fulp, W., Vera, F., Young, S., Lin, X., and Reiter, J. (2007). Secure, privacy-preserving analysis of distributed databases, *Technometrics*, 49:335–345.

[22] Kim, J.J. and Winkler, W.E. (2003). Multiplicative Noise for Masking Continuous Data, Research Report Series (Statistics ♯2003-01), Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

[21] Kooiman, P., Willenborg, L., and Gouweleeuw, J. M. (1997). PRAM: A Method for Disclosure Limitation of Microdata, Research paper no. 9705, Statistics Netherlands.

[22] Krisinich, F. and Piesse, A. (2002). Multiplicative Microdata Noise for Confidentialising Tables of Business Data: Application to AES99.

[24] Loeve, M. (1963). *Probability Theory.* D.Van Nostrand Company Inc.

[24] Nayak, T.K., Sinha, B. and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection, *Journal of Official Statistics*, 27:527–544.

[25] Oganian, A. (2010). Multiplicative noise protocals. In J. Domingo-Ferrer et al. (eds.), *Privacy in Statistical Databases (PSD 2010)*, vol. 6344 of *LNCS*. Berlin: Springer. 107–117. UNESCO Chair in Data Privacy, International Conference, Corfu, Greece, September 22-24, 2010.

[26] Sanil, A., Karr, A., Lin, X. and Reiter, J. (2004). Privacy preserving regression modelling via distributed computation. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, USA. 677–682.

[27] Shlomo, N. (2010a). Measurement error and statistical disclosure control. In *Privacy in Statistical Databases*, vol. 6344 of *LNCS*. 118–126.

[28] Shlomo, N. (2010b). Releasing microdata: Disclosure risk estimation, data masking and assessing utility, *Journal of Privacy and Confidentiality*, 2:73–91.

[29] Tendick, P. and Norman, N.S. (1987). Recent result on the noise addition method for database security. In *Proceedings of the 1987 Joint Meetings, American Statistical Association/Institute of Mathematical Statistics (ASA/IMA)*, Washington, D.C.

[30] Ting, D., Fienberg, S., and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release, *International Journal of Information and Computer Security*, 2:86–105.

[31] Torra, V., Abowd, J.M., and Domingo-Ferrer, J. (2006). Using Mahalanobis distance-based record linkage for disclosure risk assessment. In J. Domingo-Ferrer L. Franconi (eds.) *Privacy in Statistical Databases*, vol. 4302 of *LNCS* Berlin: Springer. 233–242.

[32] Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, vol. 155 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

[33] de Wolf, P., Gouweleeuw, J., Kooiman, P., and Willenborg, L. (1998). Reflections on PRAM. In *Proceedings of the Conference on Statistical Data Protection '98*, 25–27 March 1998, Lisbon.

[34] Yancey, W.E., Winkler, W.E., and Creecy, R.H. (2002). Disclosure risk assessment in perturbation micro-data protection. In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*. New York: Springer. 135–151.