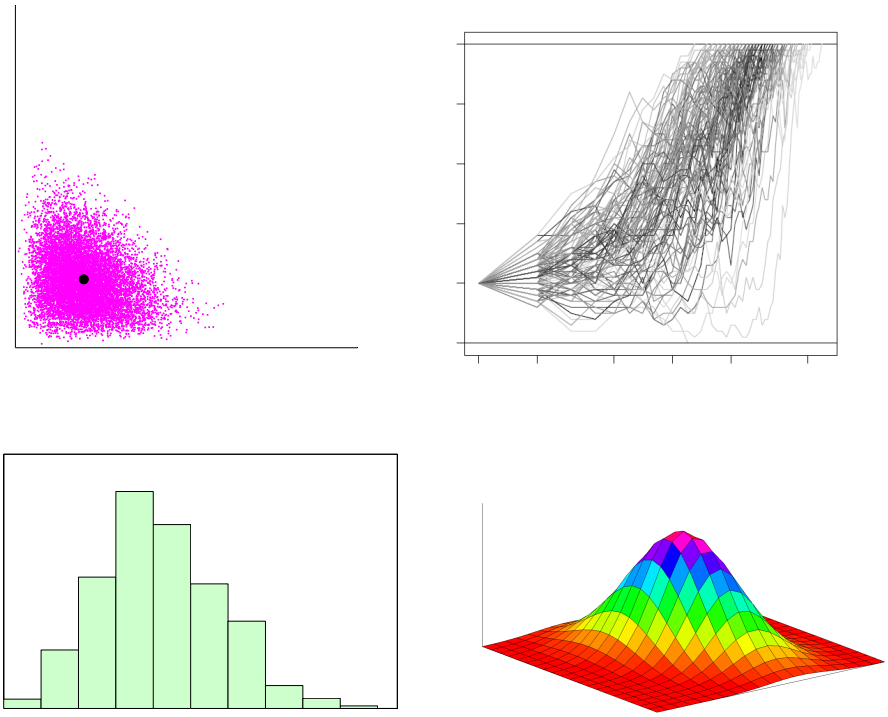
Philipps Universität Marburg

Discussion Papers on Statistics and Quantitative Methods

A Covariate Nonrandomized Response Model for Multicategorical Sensitive Variables

Heiko Groenitz

2 / 2013



Download from:
http://www.uni-marburg.de/fb02/statistik/forschung/discpap

Coordination: Prof. Dr. Karlheinz Fleischer • Philipps-University Marburg
Faculty of Business Administration • Department of Statistics
Universitätsstraße 25 • D-35037 Marburg
E-Mail: k.fleischer@wiwi.uni-marburg.de

A Covariate Nonrandomized Response Model for Multicategorical Sensitive Variables

Heiko Groenitz[1]

05.03.2013

Abstract

The diagonal model (DM) is a recently published nonrandomized response (NRR) survey method to collect data on categorical sensitive characteristics Y^*. Based on DM data, the distribution of Y^* can be estimated. In contrast to randomized response (RR) techniques, NRR schemes avoid the use of a randomization device. Due to this fact, survey complexity and study costs decrease. In this article, we assume that not only Y^*, but also nonsensitive characteristics $X_1^*, ..., X_p^*$ are involved in the survey. Then, the aim of this paper is to provide methods to investigate the dependence of Y^* on $X^* = (X_1^*, ..., X_p^*)$. For instance, the influence of gender and profession on income (recorded in income classes) may be under study. In particular, we describe two estimation procedures: Stratum-wise estimation and LR-DM estimation. Stratum-wise estimation is suitable if only few covariate levels appear in the sample. LR-DM estimation is based on a logistic regression model for the relation between Y^* and X^* and requires several techniques for generalized linear models (e.g., Fisher scoring). In simulations, we first investigate the convergence behavior of the Fisher scoring algorithm. Subsequently, we illustrate the connection between efficiency of the LR-DM estimation and the degree of privacy protection. Finally, the efficiency of the LR-DM estimation is compared with the efficiency of the stratum-wise estimation.

Zusammenfassung

Das Diagonal-Modell (DM) ist eine kürzlich publizierte Nonrandomized-Response-Methode für die Erhebung von Daten über ein sensitives, kategoriales Merkmal Y^*. Basierend auf Diagonal-Modell-Daten ist die Schätzung der Verteilung von Y^* möglich. Anders als bei Randomized-Response-Techniken ist bei Nonrandomized-Response-Verfahren die Ausführung eines Zufallsexperimentes durch die Befragten nicht nötig. Dadurch werden die Komplexität der Umfrage und die Umfragekosten reduziert. In diesem Artikel gehen wir davon aus, dass neben Y^* auch nicht-sensitive Merkmale $X_1^*, ..., X_p^*$ in die Umfrage involviert sind. Ziel dieser Arbeit ist es, Methoden zu entwickeln, die die Untersuchung des Einflusses von $X^* = (X_1^*, ..., X_p^*)$ auf Y^* ermöglichen. Zum Beispiel könnte die Abhängigkeit des Einkommens (erfasst in Klassen) von Geschlecht und Beruf von Interesse sein. In diesem Aufsatz werden zwei Schätzprozeduren beschrieben: Die schichtweise Schätzung und die LR-DM-Schätzung. Schichtweise Schätzung ist geeignet, wenn nur wenige Kovariablenlevel in der Stichprobe auftreten. LR-DM-Schätzung basiert auf einem logistischen Regressionsmodell für die Beziehung zwischen Y^* und X^* und benötigt verschiedene Methoden für generalisierte lineare Modelle (z.B. den Fisher-Scoring-Algorithmus). In umfangreichen Simulationen untersuchen wir zunächst das Konvergenzverhalten des Fisher-Scoring-Algorithmus. Anschließend illustrieren wir die Beziehung zwischen der Effizienz der LR-DM-Schätzung und dem Grad an Schutz der Privatsphäre. Schlussendlich vergleichen wir die Effizienz von LR-DM-Schätzung mit der Effizienz der schichtweisen Schätzung.

KEYWORDS: Untruthful answers; Answer refusal; Logistic regression; Generalized linear model; Fisher scoring

[1]Philipps-University Marburg, Department for Statistics (Faculty 02), Universitätsstraße 25, 35032 Marburg, Germany (e-mail: groenitz@staff.uni-marburg.de).

1 Introduction

To gather data about sensitive characteristics like income and tax evasion, it is not recommendable to ask directly, because direct questioning provokes answer refusal (i.e., missing values) or untruthful answers. Instead, survey designs that protect the respondents' privacy should be applied, because they can improve the respondents' cooperation. The first privacy-protecting survey method was the randomized response (RR) model by Warner (1965). Today there are several RR procedures which enable the estimation of the distribution of a sensitive characteristic. However, in practice, the investigator is sometimes not only interested in the distribution of the sensitive characteristic, but also in the dependence of the sensitive characteristic on certain covariates. For instance, the influence of age and profession on income might be under study.

The first covariate extension of a RR technique can be found in the book of Maddala (1983), p. 54-56, who proposes a model that enables the analysis of the relation between nonsensitive exogenous variables and a binary sensitive variable.

The paper by Scheers and Dayton (1988) extends the randomized response model by Warner (1965) and the unrelated question (UQM) model (see Greenberg et al. (1969)) with covariates. A survey according to the covariate Warner model proceeds as follows: Consider a sensitive characteristic Y^* with two outcomes, say $Y^* = 1$ and $Y^* = 2$, and an arbitrary respondent. Initially, he or she is asked directly for his or her values of p nonsensitive covariates. Subsequently, he or she draws randomly one of the questions:

$$Q^* = 1: \text{ "Is your value of } Y^* \text{ equal to 1"?} \qquad Q^* = 2: \text{ "Is your value of } Y^* \text{ equal to 2"?} \tag{1}$$

The question might be selected by spinning a spinner for example. The selection occurs hidden and the selected question is not revealed to the interviewer. The respondent replies either "yes" or "no", but the interviewer can not identify the respondent's value of the sensitive characteristic. The authors model the dependence of Y^* on the covariables, for example, by a logistic regression model, and describe methods to maximize the likelihood function. In the case of the UQM, question $Q^* = 2$ would contain a nonsensitive attribute, such as "Are you born in the first quarter of the year?". Within a real data study, the influence of the GPA (grade point average) on academic cheating behavior is investigated. Additional details of this study, especially a comparison between the estimations based on the covariate UQM and an anonymous questionnaire, are available in Scheers and Dayton (1987).

The work by van der Heijden and van Gils (1996) presents a covariate version of the RR method by Kuk (1990). Van den Hout et al. (2007) deal with the analysis of the relation between multiple sensitive characteristics and covariates where the sensitive data are gathered by randomized response methods. They present a real data example regarding social benefit fraud, more precisely the illegal receipt of unemployment benefit in the Netherlands. In particular, the relation between the binary sensitive questions "Is the number of your job applications less then required?" and "Do you conduct any work without reporting this?" and certain covariates (sex, age and an indicator whether the respondent is the main earner in the household) is studied.

In the publications of the previously mentioned authors, RR models are involved in the survey. That means that the respondents have to conduct a random experiment with the help of a randomization device (e.g., spinner or deck of cards). In contrast, nonrandomized response (NRR) techniques, which have been proposed increasingly in the last years, do not need a randomization device. The absence of a randomization device causes a reduction in survey complexity and study costs. Moreover, the respondent would always give the same answer if the survey was conducted again. One such NRR method is the diagonal model (DM) by Groenitz (2012) that is suitable for categorical sensitive characteristics.

After reviewing the DM in Section 2, we consider in Section 3 a survey which includes a sensitive

$Y^* \in \{1, ..., k\}$ and nonsensitive characteristics $X_1^*, ..., X_p^*$ where the DM is applied to elicit data about Y^*. Here, the aim of Section 3 is to investigate the influence of $X^* = (X_1^*, ..., X_p^*)$ on Y^*. For this, we present a stratum-wise estimation as well as an estimation that is based on a logistic regression model (LRM). For the latter, extensive material regarding generalized linear models (e.g., Fisher scoring) is required. In Section 4, ample simulations are presented: After a discussion about the convergence behavior of the Fisher scoring algorithm, we analyze the relation between efficiency of the estimation based on a LRM and the degree of privacy protection. Subsequently, we compare the efficiency of the estimation based on a LRM with the efficiency of the stratum-wise estimation.

2 The diagonal model

Groenitz (2012) proposes a nonrandomized response model for multichotomous sensitive variables, namely the diagonal model. This model enables the estimation of the distribution of a sensitive characteristic Y^* with codomain $\{1, ..., k\}$ by the frequencies of certain nonrandomized answers A^*, which depend on an auxiliary variable $W^* \in \{1, ..., k\}$. The auxiliary variable is assumed to be nonsensitive and independent from Y^* with a known distribution $\mathbb{P}_{W^*}$. Moreover, we assume that the interviewer does not know the respondents' values for W^*. Every respondent should give an answer according to

$$A^* := [(W^* - Y^*) \bmod k] + 1. \tag{2}$$

Instead of presenting this formula to the respondents, who may be not familiar with the modular arithmetic, every respondent is given a table where he or she can find the answer to give. For example for $k = 5$, such a table is given by

Y^*/W^*	$W^* = 1$	$W^* = 2$	$W^* = 3$	$W^* = 4$	$W^* = 5$
$Y^* = 1$	1	2	3	4	5
$Y^* = 2$	5	1	2	3	4
$Y^* = 3$	4	5	1	2	3
$Y^* = 4$	3	4	5	1	2
$Y^* = 5$	2	3	4	5	1

Additionally, an example of an answer like "If your value of Y^* equals 3 and your value of W^* equals 1, please give the answer $A^* = 4$" should be included in the questionnaire. The interviewee searches his or her values of Y^* and W^* and gives an answer A^*. Since it is not possible to identify the correct Y^*-value with the help of the answer, we assume that the interviewees cooperate. For instance, W^* could describe the period of birthday of the respondent's mother.

We denote the proportion of units in the population having $Y^* = i$, $W^* = i$ and $A^* = i$ with π_i^*, c_i^* and μ_i^*, respectively. Moreover, let C be the $k \times k$ matrix where every row is a left-cyclic shift of the row above and the first row is equal to $c^* = (c_1^*, ..., c_k^*)$. The proportions $c_1^*, ..., c_k^*$ are the model parameters and C is referred to as "design matrix of c^*". We have

$$(\mu_1^*, ..., \mu_k^*)^t = C \cdot (\pi_1^*, ..., \pi_k^*)^t. \tag{3}$$

The paper by Groenitz (2012) describes the maximum likelihood (ML) estimation in the case of simple random sampling with replacement, where it turns out that finding an explicit form of the ML estimator is difficult for some samples. However, he shows that the estimation of π^* can be viewed as missing data problem and operated with the expectation maximization (EM) algorithm.

3 Influence of nonsensitive covariates on the sensitive variable

Let us consider a survey involving a categorical, sensitive characteristic $Y^* \in \{1, ..., k\}$ where $k = q+1$ and a vector of nonsensitive characteristics $X^* = (X_1^*, ..., X_p^*)$. Here, the respondents do not provide their values of Y^*, but give an answer A^* according to the diagonal model. This answer A^* depends on both Y^* and an auxiliary characteristic W^*. We define c^* and the matrix C as in Section 2 and assume throughout the remainder of this article:

- All components of c^* are nonzero (when a c_i^* equaled zero, every answer A^* would restrict the possible Y^*-values).
- The matrix C is invertible.

The aim of this section is to study the dependence of Y^* on X^*. The quantity Y^* is called endogenous characteristic and $X_1^*, ..., X_p^*$ are called exogenous characteristics or covariates or regressors. We consider both deterministic and stochastic covariates.

3.1 The case of deterministic covariates

In this subsection, we assume that the investigator chooses the values of the covariates X^* (i.e., they are fixed and known) and searches persons having the predefined covariate levels. Each person is then requested to give a response A^* according to (2).
For instance, for X_1^*, X_2^*, and Y^* representing sex, profession, and income, respectively, this procedure means that the investigator determines for any combination of sex and profession how many persons possessing this combination are involved into to survey. Then appropriate persons are selected and each person in the sample gives DM answer A^* depending on his or her income and his or her value of the nonsensitive characteristic W^*.

Say n persons are interviewed. Consider for $i = 1, ..., n$ and $j = 1, ..., k$

$$Y_{ij} = \begin{cases} 1, & \text{if person } i \text{ has attribute } Y^* = j \\ 0, & \text{else} \end{cases}, \qquad A_{ij} = \begin{cases} 1, & \text{if person } i \text{ answers } A^* = j \\ 0, & \text{else} \end{cases},$$

let W_i denote the value of W^* corresponding to the i-th person and let x_{ij} represent the value of X_j^* corresponding to the i-th person. Set

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_{11} & \cdots & Y_{1q} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nq} \end{pmatrix}, A = \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix} = \begin{pmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nq} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Notice, the realizations of the auxiliary variables W_i and the sensitive variables Y_i are not observed while data on the answers A_i and the regressors x_i are available. We introduce $\pi_{ij} = \mathbb{E}(Y_{ij})$ and $\pi_i = (\pi_{i1}, ..., \pi_{iq})$ as well as $\mu_{ij} = \mathbb{E}(A_{ij})$ and $\mu_i = (\mu_{i1}, ..., \mu_{iq})$. Eventually, we define

$$\pi_j^*(x^*): \text{ proportion of units with } Y^* = j \text{ among the units in the population having } X^* = x^*. \quad (4)$$

In this subsection, we assume throughout

(D1) $Y_1, ..., Y_n$ independent

(D2) $W_1, ..., W_n$ are independent and identically distributed.

(D3) The two quantities $(Y_1^t, ..., Y_n^t)^t$ and $(W_1, ..., W_n)^t$ are independent.

These conditions are fulfilled if (Y^*, X^*) and W^* are independent and if for each covariate level chosen by the investigator, the sample units are drawn by simple random sampling with replacement from the population units having this covariate level where the selection for one covariate level is independent from the selection for the other covariate levels.

Let x^* be one of the covariate levels specified by the investigator, i.e., there is a row of x equal to x^*. The quantity $\pi_j^*(x^*)$ can be estimated from the answers A^* of the persons in the sample having this covariate level x^* according to the estimation procedure in Groenitz (2012) for the diagonal model. Possibly, the EM algorithm must be applied for the estimation.
Let us now assume $g \leq n$ different covariate levels are available. This means that x has g different rows. Then, the set of sample units having the i-th covariate level can be interpreted as stratum i. For this reason, we call the just described estimation method "stratum-wise estimation". One can expect the stratum-wise estimation to be suitable if each stratum contains sufficiently large sample units.

In the sequel, we present an estimation method based on a logistic regression model (LRM). Occasionally, we will call this estimation technique briefly the "LR-DM estimation". LRMs are often applied to analyze the influence of certain covariates on a categorical endogenous characteristic. Some material on LRMs that we need in this article is collected in Appendix A. For the LR-DM estimation, we make the additional assumption:

(D4) There is a $\beta = (\beta^{(1)^t}, ..., \beta^{(q)^t})^t$ with $\beta^{(i)} \in \mathbb{R}^{p\times 1}$ so that (Y, x, β) is a logistic regression model.

Of course, the vector β has length $s := pq$ and (D4) includes the independence of $Y_1, ..., Y_n$. Define for $z = (z_1, ..., z_q)$

$$h : z \mapsto (h_1(z), ..., h_q(z)) = \left(\tfrac{e^{z_1}}{1+e^{z_1}+...+e^{z_q}}, \quad \ldots, \tfrac{e^{z_q}}{1+e^{z_1}+...+e^{z_q}}\right), \text{ and } \mathbf{x}_i := \begin{pmatrix} x_i & & \\ & \ddots & \\ & & x_i \end{pmatrix} \in \mathbb{R}^{q\times pq}, \tag{5}$$

and $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Then, we have $\pi_i = h((\mathbf{x}_i\beta)^t)$. To estimate β from the LRM (Y, x, β), we have to make a detour via the answers collected in A, because Y is not observed. Let $C(1:q, j) \in \mathbb{R}^q$, $j = 1, ..., q+1$, denote the j-th column of C without the last entry, set $\tilde{c}_j = C(1:q, j) - C(1:q, q+1)$ for $j = 1, ..., q$, and define the $q \times q$ matrix $\tilde{C} := [\tilde{c}_1|\tilde{c}_2|\ldots|\tilde{c}_q]$. We introduce the map

$$m(z) = m(z_1, ..., z_q) = \left[\tilde{C} \cdot \begin{pmatrix} h_1(z) \\ \vdots \\ h_q(z) \end{pmatrix} + C(1:q, k)\right]^t. \tag{6}$$

The following theorem contains an important observation:

Theorem 1 *$(A, x, \beta, \mathbf{x}, m)$ is a generalized linear model (GLM).*

Proof: We must verify that the definition for a GLM (see Appendix B.1) is fulfilled. Since A_i is a function of Y_i and W_i, the independence of $A_1, ..., A_n$ follows. The (discrete) density of A_i is given by

$$f_{A_i}(a_1, ..., a_q) = \mu_{i1}^{a_1} \cdots \mu_{iq}^{a_q} \cdot \mu_{ik}^{1-a_1-...-a_q} \cdot 1_{\mathcal{A}}(a_1, ..., a_q), \quad a_i \in \mathbb{R},$$

where $A = \{(a_1, ..., a_q) : a_i \in \{0, 1\}, a_1 + ... + a_q \leq 1\}$. Set $\Theta = \mathbb{R}^{1\times q}$, $\Psi = \{1\}$ and for $\theta \in \Theta$, $\psi \in \Psi$, $y \in R^{1\times q}$

$$f_{\theta,\psi}(y) = c(y, \psi) \cdot e^{\frac{\theta y^t - b(\theta)}{\psi}} \text{ where } c(y, \psi) = 1_{\mathcal{A}}(y) \text{ and } b(\theta) = log(1 + e^{\theta_1} + ... + e^{\theta_q}).$$

The distribution corresponding to $f_{\theta,\psi}(y)$ is denoted with $\mathbb{P}_{\theta,\psi}$. Consequently, $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ is a simple, q-parametric exponential family with scale parameter and we have for $\psi = 1$: For all $i = 1, ..., n$,

the distribution of A_i belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$. Thus, the distribution assumption in Appendix B.1 is satisfied.

The function h is invertible with

$$h^{-1}(w_1,...,w_q) = (\log\frac{w_1}{w^*},\ldots,\log\frac{w_q}{w^*}) \text{ where } w^* := 1-(w_1+...+w_q). \tag{7}$$

Applying a chain rule, it suffices to show that the matrix $\tilde{C}$ is regular to ensure the reversibility of m. Assume $\tilde{C}$ is not regular. Then, this matrix has eigenvalue zero, i.e., there is a vector $v=(v_1,...,v_q)^t \neq 0$ with $\tilde{C}v=0$. Denoting the $q\times q$ identity matrix by I_q we can write $\tilde{C}=[I_q|(0,...,0)^t]\cdot C\cdot[I_q|(-1,...,-1)^t]^t$. It follows that $0=[I_q|(0,...,0)^t]\cdot C\cdot(v_1,...,v_q,-\sum_{j=1}^q v_j)^t =: [I_q|(0,...,0)^t]\cdot U$. Thus, the first q entries of U are zero. By taking the sum of these q numbers, we can conclude that the k-th entry of U is also zero. Altogether, C has eigenvalue zero. Because we assumed C to be invertible, this is a contradiction. Hence, $\tilde{C}$ is regular.

Finally, we have

$$\mu_i=(\mu_{i1},...,\mu_{iq}) \;=\; m\left((\mathbf{x}_i\beta)^t\right), \tag{8}$$

which completes the proof.

□

Let a_i be an observed realization of A_i. The likelihood function $\beta\mapsto f_{A_1}(a_1)\cdots f_{A_n}(a_n)$ can be maximized via the Fisher scoring algorithm. Some details of this iterative method are provided in Appendix C.1. For our GLM $(A,x,\beta,\mathbf{x},m)$, we must specify quantities from C.1 as follows. The expectation vector $\mu_i=\mu_i(\beta)$ is given through (8). The Jacobi matrix of m from (6) equals $m'(z)=\tilde{C}\cdot h'(z)$. Here, the Jacobi matrix of h is $h'(z)=\left[diag(\exp(z)\cdot Q(z))-\exp(z^t)\exp(z)\right]/\left(Q(z)\right)^2$ with componentwise application of exp and $Q(z)=1+e^{z_1}+...+e^{z_q}$. Furthermore, we have

$$D_i(\beta)=\left[m'((\mathbf{x}_i\beta)^t)\right]^t \text{ and } \Sigma_i(\beta)=Var_\beta(Y_i)=diag(\mu_i(\beta))-\mu_i(\beta)^t\mu_i(\beta).$$

In GLMs, the asymptotic normality $(F(\hat{\beta}))^{\frac{1}{2}}(\hat{\beta}-\beta)\xrightarrow{\mathcal{L}}N(0,I)$ holds for $n\to\infty$ and $\hat{\beta}$ is approximately $N(\,\beta,\,F^{-1}(\hat{\beta})\,)$-distributed if the total sample size n is sufficiently large (Fahrmeir and Tutz (2010), p. 106). Here, $F(\hat{\beta})$ is the Fisher matrix calculated under $\hat{\beta}$ and $F^{-1}(\hat{\beta})$ can be taken from the last iteration of the Fisher scoring algorithm (cf. Appendix C.1). An estimate for the asymptotical standard error of the i-th component of $\hat{\beta}$ is given by

$$\hat{SE}_{AS}(\hat{\beta}_i)=\sqrt{[F^{-1}(\hat{\beta})]_{ii}}. \tag{9}$$

We now study the estimation of the population parameters $\pi_j^*(x^*)$ from (4). Let us choose a fixed value x^*. Once obtained a maximum likelihood estimate $\hat{\beta}$, we can calculate estimates

$$[\hat{\pi}_1^*(x^*),...,\hat{\pi}_q^*(x^*)]=h((\mathbf{x}^*\hat{\beta})^t),\quad \hat{\pi}_k^*(x^*)=1-\hat{\pi}_1^*(x^*)-...-\hat{\pi}_q^*(x^*), \tag{10}$$

where $\mathbf{x}^*$ is the $q\times s$ design matrix corresponding to x^*. The identity (10) implies that $\hat{\pi}_j^*(x^*)$ is a function of $\hat{\beta}$. In particular, with $H(\beta)=(H_1(\beta),...,H_q(\beta))=h((\mathbf{x}^*\beta)^t)$ and $H_k(\beta)=h_k((\mathbf{x}^*\beta)^t)$ where $h_k(z)=1-h_1(z)-...-h_q(z)$, we have the equations

$$(\hat{\pi}_1^*(x^*),...,\hat{\pi}_q^*(x^*))=H(\hat{\beta}) \text{ and } \hat{\pi}_k^*(x^*)=H_k(\hat{\beta}). \tag{11}$$

Using a first-order Taylor approximation of H at β, we obtain

$$\begin{aligned} Var(H(\hat{\beta})) &\approx Var[H(\beta)+J_H(\beta)\cdot(\hat{\beta}-\beta)] = J_H(\beta)\cdot Var(\hat{\beta})\cdot J_H^t(\beta) \\ &\approx J_H(\hat{\beta})\cdot\hat{Var}(\hat{\beta})\cdot J_H^t(\hat{\beta}) = J_h((\mathbf{x}^*\hat{\beta})^t)\cdot\mathbf{x}^*\cdot\hat{Var}(\hat{\beta})\cdot\mathbf{x}^{*^t}\cdot J_h^t((\mathbf{x}^*\hat{\beta})^t) =: \hat{Var}(H(\hat{\beta})) \end{aligned}$$

where J denotes the Jacobi matrix and $\hat{Var}(\hat{\beta})$ is given by $F^{-1}(\hat{\beta})$. Thus, to estimate the variance of $\hat{\pi}_j^*(x^*)$ $(j = 1, ..., q)$, we can use the j-th diagonal element of $\hat{Var}(H(\hat{\beta}))$. Analog, we can derive

$$\hat{Var}(H_k(\hat{\beta})) = J_{h_k}((\mathbf{x}^*\hat{\beta})^t) \cdot \mathbf{x}^* \cdot \hat{Var}(\hat{\beta}) \cdot \mathbf{x}^{*^t} \cdot J_{h_k}^t((\mathbf{x}^*\hat{\beta})^t)$$

with the Jacobi matrix $J_{h_k}(z_1, ..., z_k) = (-e^{z_1}, ..., -e^{z_q})/(Q(z))^2$. The estimated standard errors for the $\hat{\pi}_j^*(x^*)$ are given by taking the square root of the estimated variances for $\hat{\pi}_j^*(x^*)$.

Linear hypotheses concerning β

$$H_0: \quad \mathcal{C}\beta = \rho \quad \text{against} \quad H_1: \quad \mathcal{C}\beta \neq \rho \tag{12}$$

where $\mathcal{C}$ is a $r \times s$ matrix $(r \leq s)$ with full row rank can be tested with the well known Wald statistic (cf. Fahrmeir and Tutz (2010), p. 107)

$$w = (\mathcal{C}\hat{\beta} - \rho)^t \cdot [\mathcal{C} \cdot F^{-1}(\hat{\beta}) \cdot \mathcal{C}^t]^{-1} \cdot (\mathcal{C}\hat{\beta} - \rho),$$

which is asymptotically $\chi^2_{Rank(\mathcal{C})}$-distributed under H_0.

The LR-DM estimation is built on the model structure, especially on (8). To check whether the data fit the relation (8), the Pearson statistic can be applied, provided that we have grouped data such that there is a sufficiently large number of observations in each group. As in Appendix C.1, let $g \leq n$ be the number of different rows of x, i.e., the number of covariate levels, set for $r = 1, ..., g$

$$I_r = \{i \in \{1, ..., n\} : \text{ sample unit } i \text{ possesses covariate level } r\},$$

define n_r to be the number of elements in I_r and assume $i_1 \in I_1$,...,$i_g \in I_g$. The null hypothesis H_0 is given by

$$\mathbb{E}(A_{i_1}) = m((\mathbf{x}_{i_1}\beta)^t), ..., \mathbb{E}(A_{i_g}) = m((\mathbf{x}_{i_g}\beta)^t) \text{ for one } \beta \in \mathbb{R}^s. \tag{13}$$

Set $(\tilde{A}_{r1}, ..., \tilde{A}_{rk}) = n_r^{-1} \sum_{l \in I_r} (A_{l1}, ..., A_{lk})$ and $(\tilde{\mu}_{r1}, ..., \tilde{\mu}_{rq}) = m((\mathbf{x}_{i_r}\hat{\beta})^t)$ and $\tilde{\mu}_{rk} = 1 - \tilde{\mu}_{r1} - ... - \tilde{\mu}_{rq}$. The Pearson statistic P compares $\tilde{A}_{rj}$ and $\tilde{\mu}_{rj}$, in particular, P equals

$$P = \sum_{r=1}^{g} n_r \sum_{j=1}^{k} \frac{(\tilde{A}_{rj} - \tilde{\mu}_{rj})^2}{\tilde{\mu}_{rj}}.$$

If the n_r are sufficiently large, we have approximately $P \sim \chi^2_{(g-p)q}$ under H_0. For more details, see Fahrmeir and Tutz (2010), p. 107. We remark that $\mu_i = m((\mathbf{x}_i\beta)^t) \Leftrightarrow \pi_i = h((\mathbf{x}_i\beta)^t)$. Consequently, the rejection of H_0 from (13) implies that the LRM (Y, x, β) does not fit the observed data.

We provide the self-programmed MATLAB program `fisherscore1.m`, which computes ML estimates for β and $\pi_j^*(x^*)$ (with corresponding standard errors) and assesses the goodness-of-fit, as supplemental material.

3.2 The case of stochastic covariates

In practice, it may occur that the values of the exogenous characteristics are not deterministic (i.e., not determined by the interviewer), but realizations of random variables. For such stochastic regressors, a survey proceeds as follows. Each interviewee is asked directly for his or her values of the nonsensitive covariates $X_1^*, ..., X_p^*$. Afterwards, he or she is requested to give an answer A^* according to the DM answer formula (2).

Let n, Y, A, W_i be defined as in Subsection 3.1, let the random variable X_{ij} represent the value of X_j^*

corresponding to the i-th person in the sample and set $X_i = (X_{i1}, ..., X_{ip})$ as well as $X = (X_1^t, ..., X_n^t)^t$.

In this subsection, we have to incorporate the stochastic character of X into our assumptions. In particular, we assume throughout this subsection

(S1) $(Y_1, X_1), ..., (Y_n, X_n)$ are n iid vectors.

(S2) $W_1, ..., W_n$ are iid.

(S3) The two quantities $\begin{pmatrix} Y_1, X_1 \\ \vdots \\ Y_n, X_n \end{pmatrix}$ and $\begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}$ are independent.

These requirements are satisfied when (Y^*, X^*) and W^* are independent and the interviewees are selected by simple random sampling with replacement from the population.

Stratum-wise estimation can be conducted analog to Subsection 3.1. To convey the LR-DM estimation as presented in the previous subsection to the case of stochastic regressors, we further assume

(S4) There is a $\beta = (\beta^{(1)^t}, ..., \beta^{(q)^t})^t$ with $\beta^{(i)} \in \mathbb{R}^{p\times 1}$ so that (Y, X, β) is a LRM with stochastic covariates (see Appendix A.2).

We have that $(A_1, X_1), ..., (A_n, X_n)$ are n iid vectors and that $A_1, ..., A_n$ are independent given $X_1 = x_1, ..., X_n = x_n$ (for all values $x_1, ..., x_n$). Moreover, with $\mathbf{X}_i := \begin{pmatrix} X_i & & \\ & \ddots & \\ & & X_i \end{pmatrix} \in \mathbb{R}^{q\times pq}$ and $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)$ as well as m from (6), we have $\mathbb{E}(A_i|X) = m((\mathbf{X}_i\beta)^t)$ and $(A, X, \beta, \mathbf{X}, m)$ is a GLM with stochastic covariates (cf. Appendix B.2).

The maximum likelihood estimation for $\beta \in \mathbb{R}^{s\times 1}$ with $s = pq$ in this GLM with stochastic covariates can be traced back to the ML estimation in a GLM with deterministic covariates (see Appendix C.2). Thus, our program `fisherscore1.m` can also be applied to calculate MLEs in the case of stochastic covariates. The asymptotic normality $(F(\hat{\beta}))^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, I)$ of the MLE $\hat{\beta}$ also holds for GLMs with stochastic covariates (Fahrmeir and Tutz (2010), p. 106). Thus, $\hat{\beta}$ has the approximative distribution $N(\,\beta,\, F^{-1}(\hat{\beta})\,)$ when n is sufficiently large. Consequently, an estimate for the asymptotical standard error of $\hat{\beta}_i$ is $\sqrt{[F^{-1}(\hat{\beta})]_{ii}}$. Linear hypotheses (12) can be tested with the Wald statistic (Fahrmeir and Tutz, p.107)

$$W = (\mathcal{C}\hat{\beta} - \rho)^t \cdot [\mathcal{C} \cdot F^{-1}(\hat{\beta}) \cdot \mathcal{C}^t]^{-1} \cdot (\mathcal{C}\hat{\beta} - \rho),$$

which is also in the case of stochastic covariates asymptotically $\chi^2_{Rank(\mathcal{C})}$-distributed under the null hypothesis.

For a fixed covariate level x^*, the population parameters $\pi_j^*(x^*)$ from (4) can be estimated totally analog to Subsection 3.1 by (11). The estimated standard errors for this estimation can be obtained again as in Subsection 3.1.

For grouped data with a sufficiently large number of observations in each group, the goodness-of-fit can be assessed by the Pearson statistics P as in Subsection 3.1, where we have the approximative conditional distribution $P|X = x \quad \sim \quad \chi^2_{(g-p)q}$ under H_0.

4 Simulations

4.1 Convergence behavior of the scoring algorithm

The maximum likelihood estimation for a GLM according to Section 3 requires the maximization of

$$\beta \mapsto \sum_{i=1}^{n} (a_{i1}, ..., a_{ik}) \cdot \log \left(C \cdot (\pi_{i1}, ..., \pi_{ik})^t \right) \tag{14}$$

where a_{ij} is a realization of A_{ij}, π_{ij} depends on β, and log is applied componentwise. It may occur that the function (14) does not possess a maximum. A discussion about the existence of an MLE in general GLMs including further references can be found in Fahrmeir and Tutz (2010), p. 43. Nevertheless, the mathematical conditions for the existence are usually difficult to check in practice. We will illustrate the non-existence with some examples:

1. We first give an example for which we can show by simple analytic methods that a MLE does not exist. Let $Y^* \in \{1, ..., k\}$ (with $\pi_i^* > 0$) be a sensitive variable and assume we have conducted a survey due to the non-covariate diagonal model with n interviewees drawn by a simple random sample with replacement. Define Y, A, W_i as in Subsection 3.1. For observed values a_{ij} of A_{ij}, the log-likelihood is given by

$$\tilde{l}((\pi_1, ..., \pi_k)^t) = \left(\sum_{i=1}^{n} (a_{i1}, ..., a_{ik}) \right) \cdot \log \left(C \cdot (\pi_1, ..., \pi_k)^t \right).$$

Set $x = (1, ..., 1)^t \in \mathbb{R}^n$, $\mathbf{X}_i = I_q$, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $\beta = h^{-1}(\pi_1^*, ..., \pi_q^*)$ with the link function h^{-1} from (7). With the map m from (6), it follows that $(A, x, \beta, \mathbf{x}, m)$ is a GLM with log-likelihood function

$$l(\beta) = \left(\sum_{i=1}^{n} (a_{i1}, ..., a_{ik}) \right) \cdot \log \left(C \cdot H(\beta) \right),$$

where H is a function $\mathbb{R}^{q \times 1} \to \{(x_1, ..., x_k)^t : x_i \in (0, 1), \sum_{i=1}^{k} x_i = 1\}$ with $H(\beta) = (h_1(\beta^t), ..., h_q(\beta^t), 1 - h_1(\beta^t) - ... - h_q(\beta^t))^t$.

Let us now specify $k = 2$, $c^* = (0.6, 0.4)$ and let the number of respondents who give answer 1 and 2 equal 15 and 5, respectively. Suppose that l possesses on $\mathbb{R}$ a maximum $\hat{\beta}$. Then, $H(\hat{\beta})$ would be the maximum of $\tilde{l}$ on the set $\{(x_1, x_2)^t : x_i \in (0, 1), x_1 + x_2 = 1\}$. However, we can easily show that $\tilde{l}$ does not possess a maximum on $\{(x_1, x_2)^t : x_i \in (0, 1), x_1 + x_2 = 1\}$ for above specifications. Due to this contradiction, l has no maximum on $\mathbb{R}$.

2. Let us consider a sensitive Y^* with range $\{1, 2\}$ and exogenous characteristics $X^* = (X_1^*, X_2^*)$ where X_1^* is constant equal to one and $X_2^* \in \{1, 2, 3\}$. We assume stochastic covariates and make the following specifications taken from an example in Scheers and Dayton (1988), Section 3:

$$n = 200, \quad w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0.1587 \\ 0.6826 \\ 0.1587 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} -3.118 \\ 1.218 \end{pmatrix}$$

where w_i is defined to be the proportion of individuals in the universe having attribute $X_2^* = i$. Furthermore, we set $c^* = (0.7, 0.3)$. As before c^* describes the distribution of an auxiliary variable. We have simulated 1000 samples where realizations of A and X are available for each sample. To obtain one sample it suffices to generate absolute frequencies of the covariate levels $(n_1, n_2, n_3) \sim Multinomial(n, w)$ and to subsequently draw the frequencies of the answers $A^* = j$ for each covariate level from the multinomial distribution with number of trials equal to n_i and cell probabilities $(m(\beta_1 + i\beta_2), 1 - m(\beta_1 + i\beta_2))$.

For each sample, we tried to compute a MLE $\hat{\beta}$ with the self-programed MATLAB program `fisherscore1` and also with the function `glmfit` which is already available in MATLAB. A valid estimate is obtained for most samples, but for some samples the estimation fails. For instance, no problems occur for

covariate level (X_1^*, X_2^*)	$(1,\ 1)$	$(1,2)$	$(1,3)$
observations	32	137	31
frequency of $A^* = 1$	16	55	16

(15)

where $\hat{\beta} = \begin{pmatrix} -0.8750 & 0.0999 \end{pmatrix}^t$. Otherwise, the sample with

covariate level (X_1^*, X_2^*)	$(1,\ 1)$	$(1,2)$	$(1,3)$
observations	30	144	26
frequency of $A^* = 1$	8	68	18

(16)

leads to $\hat{\beta} = (NaN,\ NaN)^t$ in `fisherscore1` respectively to a complex-valued $\hat{\beta} = (-8.2030 + 6.2832\mathtt{i},\ 3.9660 - 3.1416\mathtt{i})^t$ using `glmfit`. The contour plots in Figures 1 and 2 give an illustration of the log-likelihood function for (15) and (16). According to our simulation, non-convergence occurred in 5.4% (`fisherscore1`) respectively 7.3% (`glmfit`) of the samples. The difference may be explained by the fact that `fisherscore1` has used several starting values whereas user-defined starting values cannot be inputted in `glmfit`.

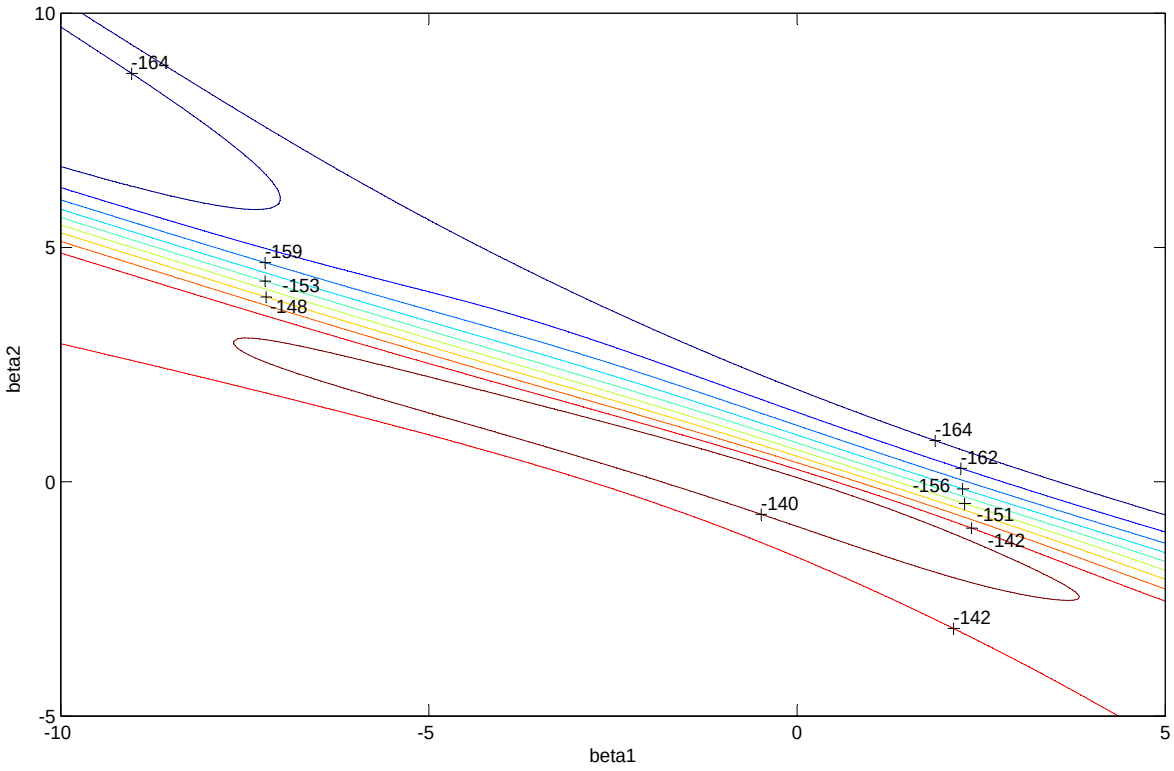


Figure 1: Contour plot for the log-likelihood l corresponding to (15): Consider the isoline $l = -140$. Outside this isoline we have $l < -140$ and the maximum of l is located in the domain $\{\beta : l(\beta) \geq -140\}$. In particular, the maximum is $(-0.875,\ 0.0999)$ with log-likelihood value -136.93.

4.2 Efficiency of LR-DM estimation and degree of privacy protection (DPP)

For the non-covariate diagonal model, Groenitz (2012), Sections 3.5 and 4.2, has shown how the distribution $c^* = (c_1^*, ..., c_k^*)$ of the auxiliary characteristic W^* influences the DPP and efficiency. The

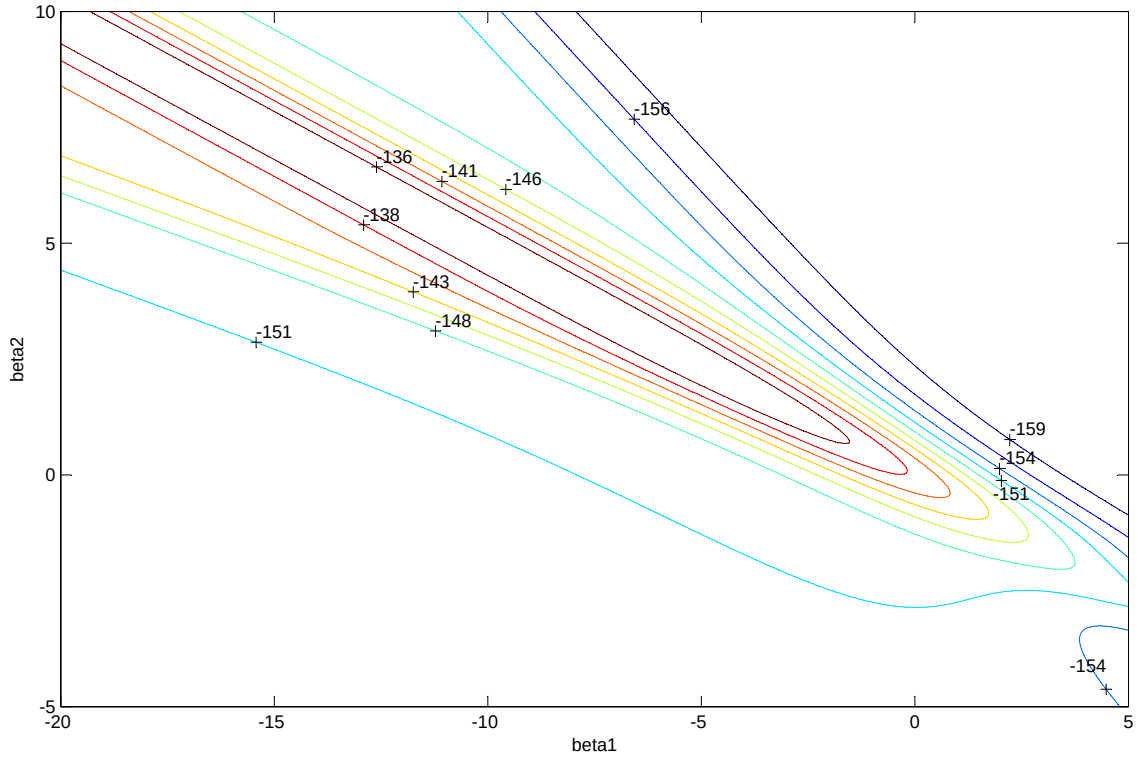


Figure 2: Contour plot for the log-likelihood l corresponding to (16): If we are located in $(0,0)^t$ and move upwards and to the left in the picture, we will successively find vectors β with increasing likelihood.

goal of this section is to illustrate the influence of c^* for the LR-DM estimation within a simulation study. Here, we consider $k = 4$, $n = 300$, $X^* = (X_1^*, X_2^*)$, where X_1^* is a constant equal to one and X_2^* has codomain $\{1, ..., 5\}$, as well as $\beta = (3.5, -1.25, 2.5, -0.5, 2, -0.25)^t$ and $w = (1, 2, 3, 2, 1)/9$. The i-th component of w denotes the proportion of people in the population with level $X_2^* = i$. The entry (i, j) of the matrix

$$\begin{pmatrix} 0.4015 & 0.3127 & 0.2435 & 0.0423 \\ 0.2143 & 0.3534 & 0.3534 & 0.0789 \\ 0.0975 & 0.3403 & 0.4370 & 0.1252 \\ 0.0399 & 0.2949 & 0.4863 & 0.1789 \\ 0.0153 & 0.2392 & 0.5063 & 0.2392 \end{pmatrix} \tag{17}$$

denotes the proportion of units with $Y^* = j$ among the units in the universe having covariate value $X_2^* = i$. That is, the matrix entries equal the $\pi_j^*(x^*)$ according to (4). Imagine that Y^* describes income classes where category $Y^* = 1$ ($Y^* = k$) represents low (high) income, and covariable X_2^* describes age classes where $X_2^* = 1$ ($X_2^* = 5$) indicates a low (high) age. Then, (17) might be realistic relative frequencies, because income often grows with increasing age.

We can measure the efficiency of estimators $\hat{\pi}_j^*(x^*)$ for each covariate level x^* (for our setup, we have $x^* \in \{(1, i) : i = 1, ..., 5\}$) by

$$trace\left[MSE\left(\hat{\pi}_1^*(x^*), ..., \hat{\pi}_k^*(x^*)\right)\right] = MSE(\hat{\pi}_1^*(x^*)) + ... + MSE(\hat{\pi}_k^*(x^*)). \tag{18}$$

In our simulations, we consider several vectors c^*. As in Groenitz (2012), we use the standard deviation of the vector c^*, denoted by $\sigma = std(c^*) \in [0, \sqrt{1/k}]$, to quantify the DPP. In other words, we measure the closeness of the distribution of W^* to a degenerate and a uniform distribution.

The simulations start with the draw of 500 vectors $c^* = (c_1^*, ..., c_4^*)$ which are uniformly scattered on $\{(x_1, ..., x_4) \in [0,1]^4 \ : \ x_1 + ... + x_4 = 1\}$. One such c^* can be generated as follows: Simulate (c_1^*, c_2^*, c_3^*) from a Dirichlet distribution with parameter $(1,1,1,1)$, see Gentle (1998), p. 111, and define $c_4 = 1 - (c_1 + c_2 + c_3)$.

For each drawn c^*, we compute the standard deviation of c^* as measure for the DPP and generate 100 samples. To obtain one sample, we draw $(n_1, ..., n_5) \sim Multinomial(n, w)$. This implies that we have stochastic covariates. Afterwards, we draw the frequencies of the responses $A^* = j$ for each covariate level x^* from the multinomial distribution with parameters n_i and

$$\left[m_1((\mathbf{x}^*\beta)^t), \quad ... \quad , m_q((\mathbf{x}^*\beta)^t), \quad 1 - \sum_{j=1}^{q} m_j((\mathbf{x}^*\beta)^t) \right]. \tag{19}$$

As before, $\mathbf{x}^*$ denotes the $q \times s$ design matrix corresponding to x^*. As already mentioned in Section 4.1, the ML estimation for β may fail. We delete all samples in which `fisherscore1` does not converge. For each of the remaining samples, we calculate $\hat{\pi}_j^*(x^*)$ from $\hat{\beta}$, see (10). Based on the realizations of $\hat{\pi}_j^*(x^*)$, we calculate the empirical MSE. That is, we compute an estimate $\hat{\mathbb{E}}((\hat{\pi}_j^*(x^*) - \pi_j^*(x^*))^2 \,|\, \mathcal{B})$ with the event $\mathcal{B} = \{$ MLE exists $\}$. The quantity (18) is then estimated by the simulated MSE sum $\sum_{j=1}^{4} \hat{\mathbb{E}}((\hat{\pi}_j^*(x^*) - \pi_j^*(x^*))^2 \,|\, \mathcal{B})$.

As soon as the simulations for the randomly drawn c^* have been completed, we repeat the procedure with the vectors $c^{*(1)}, ..., c^{*(6)} \in \mathbb{R}^4$ according to Theorem 2b in Groenitz (2012) for the corresponding degrees of privacy protection $\sigma_i = i/12$. Clearly, the σ_i $(i = 1, ..., 6)$ are equidistant points in the range of the standard deviation.

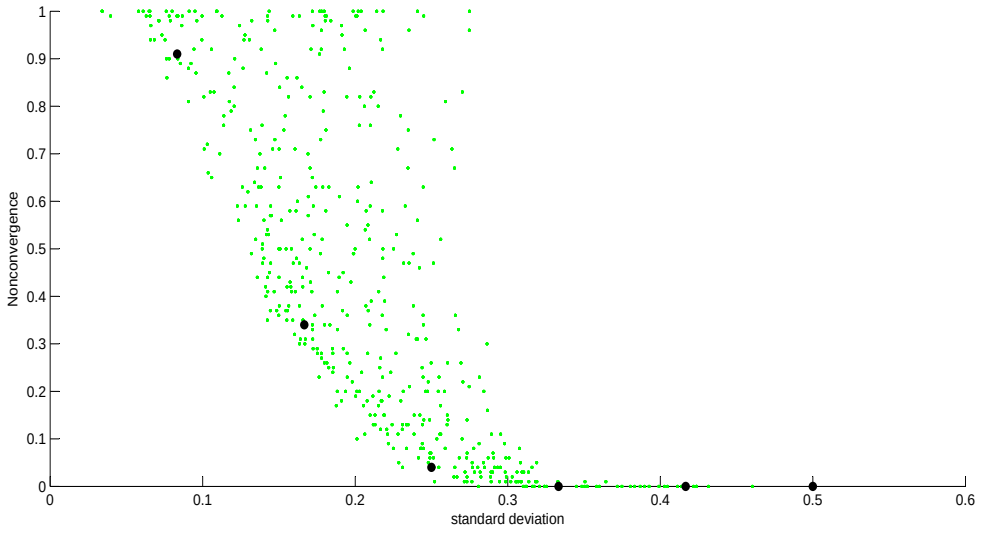


Figure 3: Nonconvergence rates in dependence of the standard deviation σ. A small point corresponds to a vector c^* that is drawn randomly. The boldfaced black dots belong to the $c^{*(i)}$.

Due to Figure 3, the nonconvergence probability seems to have a lower bound that depends on σ. The nonconvergence rates of $c^{*(i)}$ decrease from $c^{*(1)}$ to $c^{*(6)}$ and are close to this lower bound. However, $c^{*(1)}$ and $c^{*(2)}$ are impractical, because the ML estimation often fails. Let us now consider Figure 4. For any covariate level, the point cloud for the randomly drawn vectors has a lower bound. The crosses $(\times)$ for $c^{*(2)}, ..., c^{*(6)}$ ($c^{*(1)}$ was omitted due to the high nonconvergence rate) are located quite accurate

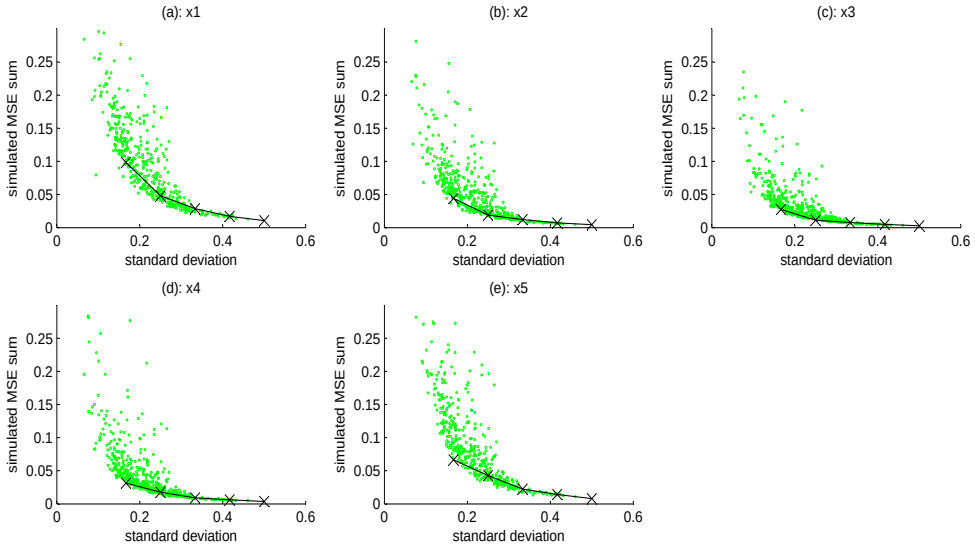


Figure 4: Plots of the simulated MSE sum against the standard deviation for each covariate level. E.g., `x1` in the heading of the first plot, means the covariate level $x^* = (1, 1)$. A small point corresponds to a vector c^* that is drawn randomly. The boldfaced black dots belong to $c^{*(2)}, ..., c^{*(6)}$.

on this bound. Thus, we conclude that the $c^{*(i)}$ are efficient choices for $\mathbb{P}_{W^*}$ for the corresponding degrees of privacy protection. If we connect the 5 crosses, we obtain a strictly monotonically decreasing polygonal curve. That means a larger degree of privacy protection is associated with smaller efficiency. Altogether, the observed influence of $\mathbb{P}_{W^*}$ on efficiency of the LR-DM estimation coincides with the results for the non-covariate diagonal model.

Hence, the interviewer should fix a medium value of σ and determine the vector c^* via Theorem 2b from Groenitz (2012). Finally, an auxiliary attribute W^* should adapted on the chosen c^*.

4.3 Efficiency comparison

Let us consider a sensitive characteristic $Y^* \in \{1, ..., k\}$ and covariates $X^* = (X_1^*, X_2^*)$ where X_1^* is constant equal to one and X_2^* is nonsensitive and can attain the outcomes $1, ..., g^*$. We specify $k = 3$, $c^* = (2/3,\ 1/6,\ 1/6)$, and $g^* \in \{3, 5\}$, i.e., either three or five covariate levels appear in the population. Moreover, we assume that the relation between Y^* and X^* follows a logistic regression model with $\beta = (3.50,\ -1.25,\ 2.50,\ -0.50)^t$. We have the following proportions of units having $Y^* = j$ among the units in the population with covariate level x^*:

$g^* = 3$ covariate levels

x^* / j	1	2	3
(1,1)	0.5307	0.4133	0.0559
(1,2)	0.3315	0.5465	0.1220
(1,3)	0.1732	0.6045	0.2224

$g^* = 5$ covariate levels

x^* / j	1	2	3
(1,1)	0.5307	0.4133	0.0559
(1,2)	0.3315	0.5465	0.1220
(1,3)	0.1732	0.6045	0.2224
(1,4)	0.0777	0.5741	0.3482
(1,5)	0.0310	0.4845	0.4845

(20)

Similar to Section 4.2 the proportions in (20) might be realistic proportions for Y^* and X_2^* describing income and age classes, respectively. Notice, the elements of the tables in (20) equal the $\pi_j^*(x^*)$ according to (4). We consider sample sizes $n \in \{100,\ 200,\ 300,\ 400\}$ and several specifications for w

where the i-th component of w denotes the relative frequency of units in the universe having $x^* = (1, i)$:

$$\begin{array}{ll} g^* = 3: & w^{(1)} = (1,\ 1,\ 1)/3 \text{ and } w^{(2)} = (1,\ 2,\ 3)/6 \\ g^* = 5: & w^{(1)} = (1,\ 1,\ 1,\ 1,\ 1)/5 \text{ and } w^{(2)} = (1,\ 2,\ 3,\ 2,\ 1)/9 \end{array} \tag{21}$$

The aim of this subsection is to compare the efficiency of two estimation procedures: On the one hand, we estimate $\pi_j^*(x^*)$ from (20) according to the LR-DM estimation. On the other hand, a stratum-wise estimation is conducted.

For each specification of (g^*, w, n), we simulate 1000 samples. Each sample consists of $n_i = round(w_i \cdot n)$ interviewees with covariate level $x^* = (1, i)$. Here, the operator *round* means rounding to the nearest integer and w_i is the i-th component of w. This situation corresponds to deterministic covariates. For covariate level $x^* = (1, i)$, we draw the frequencies of the replies A^* from a multinomial distribution analog to the description around (19). Since the ML estimation for β may fail, we delete all samples in which `fisherscore1` does not converge. For each of the remaining samples, we calculate estimates for $\pi_j^*(x^*)$ - once by LR-DM estimation and once by stratum-wise estimation.

For each considered estimator, we compute the average and the empirical mean squared error (MSE) from the available realizations. This means that we obtain estimates for expectation and MSE of the estimators. An excerpt of the simulation output can be found in the Tables 1 and 2.

LR-DM estimation average of the estimates				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n = 300$, $w^{(1)}$	$(1,1)$	0.5411	0.3972	0.0617
	$(1,2)$	0.3283	0.5478	0.1239
	$(1,3)$	0.1718	0.6127	0.2156
	$(1,4)$	0.0835	0.5774	0.3390
	$(1,5)$	0.0385	0.4757	0.4858
	non-conv	3		
$n = 300$, $w^{(2)}$	$(1,1)$	0.5289	0.4081	0.0629
	$(1,2)$	0.3247	0.5509	0.1244
	$(1,3)$	0.1712	0.6111	0.2176
	$(1,4)$	0.0851	0.5715	0.3434
	$(1,5)$	0.0414	0.4731	0.4855
	non-conv	4		

Stratum-wise estimation average of the estimates				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n = 300$, $w^{(1)}$	$(1,1)$	0.5295	0.3992	0.0713
	$(1,2)$	0.3322	0.5419	0.1259
	$(1,3)$	0.1738	0.5971	0.2292
	$(1,4)$	0.0918	0.5738	0.3344
	$(1,5)$	0.0569	0.4710	0.4721
$n = 300$, $w^{(2)}$	$(1,1)$	0.5052	0.4076	0.0872
	$(1,2)$	0.3288	0.5443	0.1269
	$(1,3)$	0.1739	0.6067	0.2194
	$(1,4)$	0.0931	0.5642	0.3427
	$(1,5)$	0.0694	0.4665	0.4641

Table 1: The left (right) part of the table contains the averages of the estimates for $\pi_j^*(x^*)$ according to the LR-DM estimation (stratum-wise estimation). The entry "non-conv" counts how often `fisherscore1` did not converge.

We first regard five covariate levels. It turns out, that the nonconvergence rates decrease strongly with increasing sample size (for $w^{(1)}$: reduction from 19,6% ($n = 100$) to 0,3% ($n = 400$); for $w^{(2)}$: reduction from 13% ($n = 100$) to 0,2% ($n = 400$)). This coincides with the theoretic result that the existence of a MLE for β in GLMs is asymptotically guaranteed (cf. Fahrmeir and Tutz (2010), p.44).

Let us now focus on the estimation of the conditional proportions $\pi_j^*(x^*)$. On average, the estimates calculated according to both LR-DM and stratum-wise estimation are close to the true values of $\pi_j^*(x^*)$. Regarding efficiency, the empirical MSEs of the estimates decreases if the sample size grows. Moreover, the empirical MSEs corresponding to LR-DM estimation are always smaller than the MSEs corresponding to stratum-wise estimation. The quotient of empirical MSE for LR-DM estimation divided by empirical MSE for stratum-wise estimation attains values between 17% and 93% where it is mostly less than 60%. That is, the estimation precision can be improved significantly by using the functional form (22) from Appendix A.1.

LR-DM estimation (MSEs)				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n=300$, $w^{(1)}$	$(1,1)$	0.0129	0.0102	0.0022
	$(1,2)$	0.0063	0.0057	0.0039
	$(1,3)$	0.0048	0.0059	0.0051
	$(1,4)$	0.0029	0.0053	0.0054
	$(1,5)$	0.0014	0.0103	0.0112
	non-conv.	3		
$n=300$, $w^{(2)}$	$(1,1)$	0.0178	0.0142	0.0024
	$(1,2)$	0.0066	0.0062	0.0035
	$(1,3)$	0.0044	0.0052	0.0038
	$(1,4)$	0.0029	0.0056	0.0056
	$(1,5)$	0.0017	0.0142	0.0159
	non-conv.	4		

Stratum-wise estimation (MSEs)				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n=300$, $w^{(1)}$	$(1,1)$	0.0147	0.0142	0.0065
	$(1,2)$	0.0146	0.0149	0.0094
	$(1,3)$	0.0117	0.0169	0.0132
	$(1,4)$	0.0077	0.0152	0.0143
	$(1,5)$	0.0058	0.0146	0.0147
$n=300$, $w^{(2)}$	$(1,1)$	0.0260	0.0242	0.0112
	$(1,2)$	0.0129	0.0144	0.0084
	$(1,3)$	0.0079	0.0104	0.0081
	$(1,4)$	0.0070	0.0133	0.0129
	$(1,5)$	0.0102	0.0264	0.0252

Table 2: Empirical mean squared errors (MSEs) of the estimates for $\pi_j^*(x^*)$ using the LR-DM procedure and the stratum-wise estimation.

The aforementioned observations for five covariate levels can be also found in the case of three covariate levels. The only noticeable difference is that higher nonconvergence rates of `fisherscore1` occur in the three level case. Altogether, we conclude the major result of this section: If the logistic regression model fits the data, the use of the functional structure (22) leads to a considerably reduction of the MSE.

5 Summary

In this article, we have considered a survey with a sensitive attribute $Y^* \in \{1,...,k\}$ and nonsensitive characteristics $X^* = (X_1^*,...,X_p^*)$ where the collection of data on Y^* is conducted with the nonrandomized diagonal model. To examine the dependence of Y^* on X^*, we have introduced the stratum-wise estimation and the LR-DM estimation, which is built on a logistic regression model for the relation between Y^* and X^*. For the LR-DM estimation, maximum likelihood estimates must be computed iteratively where the Fisher scoring algorithm is helpful. In simulations, we investigated the convergence probabilities of Fisher scoring and discussed how the efficiency of the LR-DM estimation depends on the degree of privacy protection. In a further part of the simulation study, we considered a situation where the data fit a logistic regression model. We found out that the application of the functional relation between the proportion of units in the population having outcome $Y^*=j$ and the covariates leads to considerably smaller mean squared errors than a stratum-wise estimation.

Acknowledgments:

The author would like to thank Prof. Dr. Karlheinz Fleischer for helpful comments.

Appendix

For the LR-DM estimation we need some material regarding logistic regression models (LRMs) and generalized linear models (GLMs). Although LRMs and GLMs are well-known (e.g., Fahrmeir and Tutz (2010)), we briefly mention some facts in this appendix to increase the readability of the paper.

A Logistic regression models (LRMs)

A.1 LRMs with deterministic covariates

Consider random variables Y_{ij} $(i = 1, ..., n;\ j = 1, ..., q)$, define the random vectors $Y_i = (Y_{i1}, ..., Y_{iq})$ and the random matrix $Y = (Y_1^t, ..., Y_n^t)^t$. Let x_{ij} $(i = 1, ..., n; j = 1, ..., p)$ be real numbers, define $x_i = (x_{i1}, ..., x_{ip})$ and the deterministic matrix $x = (x_1^t, ..., x_n^t)^t$. Moreover, assume $\beta^{(1)}, ..., \beta^{(q)} \in \mathbb{R}^{p\times 1}$ and set $\beta = ({\beta^{(1)}}^t, ..., {\beta^{(q)}}^t)^t$. The triple (Y, x, β) is called logistic regression model, if

1. $Y_1, ..., Y_n$ are independent and the random vector $(Y_{i1}, ..., Y_{iq}, 1 - \sum_{j=1}^q Y_{ij})$ is multinomially distributed with number of trials equal to one.

2. The equations

$$\mathbb{P}(Y_{ij} = 1) = \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + ... + e^{x_i\beta^{(q)}}} \quad (i = 1, ..., n; j = 1, ..., q) \tag{22}$$

   hold for the cell probabilities.

When (Y, x, β) is a LRM, we set $k = q + 1$, $Y_{ik} = 1 - \sum_{j=1}^q Y_{ij}$ and can conclude that

$$\mathbb{P}(Y_{ij} = 1) \,/\, \mathbb{P}(Y_{ik} = 1) = e^{x_i\beta^{(j)}} \quad (j = 1, ..., q). \tag{23}$$

In applications, LRMs are useful to study the dependence of a categorical characteristic $Y^* \in \{1, ..., k\}$ with $k = q+1$ on a vector of covariates $X^* = (X_1^*, ..., X_p^*)$. Here, one considers a sample of size n and the Y_{ij} are given by

$$Y_{ij} = 1 \ \ (Y_{ij} = 0) \text{ if sample unit } i \text{ possesses outcome } Y^* = j \ \ (Y^* \neq j),$$

whereas the value of X^* corresponding to the i-th sample unit is denoted with x_i. According to (23), the components of the parameter β can be interpreted in the following way: E.g., an increase by 1 in the second covariate causes a change in the odds ratio $\mathbb{P}(Y_{ij} = 1) \,/\, \mathbb{P}(Y_{ik} = 1)$ by the factor $e^{\beta_2^{(j)}}$.

A.2 LRMs with stochastic covariates

In practice, the values of the covariates are often not deterministic, but realizations of random quantities. This motivates to consider also LRMs with stochastic regressors. Define Y and β as in A.1, let X_{ij} $(i = 1, ..., n; j = 1, ..., p)$ be random variables, set $X_i = (X_{i1}, ..., X_{ip})$ and $X = (X_1^t, ..., X_n^t)^t$. The triple (Y, X, β) is called a LRM with stochastic covariates, if the following properties are satisfied for every value x of X:

1. The $Y_1, ..., Y_n$ are independent given $X = x$ and the conditional distribution of the vector $(Y_{i1}, ..., Y_{iq}, 1 - \sum_{j=1}^q Y_{ij})$ given $X = x$ is a multinomial distribution with number of trials equal to one.

2. The identities

$$\mathbb{P}(Y_{ij} = 1 \,|\, X = x) = \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + ... + e^{x_i\beta^{(q)}}} \quad (i = 1, ..., n; j = 1, ..., q) \tag{24}$$

   hold (x_i is the i-th row of x).

B Generalized linear models (GLMs)

As preparatory work, we need the following definition: A family $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ of distributions on the Borel σ-algebra over $\mathbb{R}^q$ is called "simple, q-parametric exponential family with scale parameter" if functions $c : \mathbb{R}^q \times \Psi \to [0,\infty)$ and $b : \Theta \to \mathbb{R}$ exist with the property: Any $\mathbb{P}_{\theta,\psi}$ has a density of the form

$$f_{\theta,\psi}(y) = f_{\theta,\psi}(y_1,...,y_q) = c(y,\psi)\cdot e^{\frac{\theta y^t - b(\theta)}{\psi}} \quad (y\in\mathbb{R}^q).$$

B.1 GLMs with deterministic covariates

Consider random variables Y_{ij} $(i = 1,...,n;\ j = 1,...,q)$, the random vectors $Y_i = (Y_{i1},...,Y_{iq})$ and the random matrix $Y = (Y_1^t,...,Y_n^t)^t$. Let x_{ij} $(i = 1,...,n; j = 1,...,p)$ be real numbers, $x_i = (x_{i1},...,x_{ip})$ and $x = (x_1^t,...,x_n^t)^t$. Moreover, let β be a vector in $\mathbb{R}^{s\times 1}$, $\mathbf{x}_i$ a $q\times s$ matrix created from x_i, $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_n)$, and $h : z = (z_1,...,z_q) \mapsto (h_1(z),...,h_q(z))$ an invertible function. Then, $(Y, x, \beta, \mathbf{x}, h)$ is called a generalized linear model, if (G1) and (G2) hold:

(G1) *Distribution assumption:*

(a) There is a simple, q-parametric exponential family with scale parameter $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ and one element $\psi\in\Psi$ with the property: For all $i = 1,...,n$, the distribution of Y_i belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$.

(b) $Y_1,...,Y_n$ are independent.

(G2) *Structure assumption:*

The expectation vector $\mu_i = \mathbb{E}(Y_i)$ and the linear predictor $\eta_i = (\mathbf{x}_i\beta)^t$ are connected by h, that is, $\mu_i = h(\eta_i)$.

In applications, n is the sample size while x_i and Y_i represent the values of the covariates and the endogenous characteristics corresponding to the i-th sample unit.

B.2 GLMs with stochastic covariates

Consider Y, β and h as in B.1. Let X_{ij} $(i = 1,...,n; j = 1,...,p)$ be random variables, $X_i = (X_{i1},...,X_{ip})$ and $X = (X_1^t,...,X_n^t)^t$. Moreover, let $\mathbf{X}_i$ be a $q\times s$ matrix created from X_i, $\mathbf{X} = (\mathbf{X}_1,...,\mathbf{X}_n)$. We call $(Y, X, \beta, \mathbf{X}, h)$ a GLM with stochastic covariates, if:

(G1) *Distribution assumption:*

(a) There is a simple, q-parametric exponential family with scale parameter $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ and one element $\psi\in\Psi$ with the property: For all $i = 1,...,n$ and all possible realizations x of X, the conditional distribution of Y_i given $X = x$ belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$.

(b) $Y_1,...,Y_n$ are independent given $X = x$ (for any value x of X).

(G2) *Structure assumption:*

The conditional expectation $\mu_i = \mathbb{E}(Y_i|X)$ and $\eta_i = (\mathbf{X}_i\beta)^t$ are connected by $\mu_i = h(\eta_i)$.

C Fisher scoring in GLM

Fisher scoring is an iterative method to compute maximum likelihood estimates. Notice, in (G1) from B.1 respectively B.2 the set of scale parameters Ψ appears. We describe Fisher scoring only for the case $\Psi = \{1\}$, because this case is relevant in this article.

C.1 Fisher scoring in GLMs with deterministic covariates

Let $(Y, x, \beta, \mathbf{x}, h)$ be a GLM and $y = (y_1^t, ..., y_n^t)^t \in \mathbb{R}^{n\times q}$ an observed value of Y. According to (G1), we need to maximize $l(\beta) = l(\beta, y) = \sum_{i=1}^n l_i(\beta)$ in β where $l_i(\beta) = l_i(\beta, y) = \theta_i y_i^t - b(\theta_i)$. To maximize l, the Fisher scoring algorithm generates a sequence of estimates $(\beta_\nu)_{\nu\in\mathbb{N}_0}$ as follows: When an estimate β_ν is available from the preceding iteration, the next estimate is computed by

$$\beta_{\nu+1} = \beta_\nu + F^{-1}(\beta_\nu) \cdot s(\beta_\nu). \tag{25}$$

Here, $s(\beta) = s(y, \beta) = (l'(\beta))^t$ is called score function, where $l'(\beta) \in \mathbb{R}^{1\times s}$ denotes the Jacobi matrix of l at β, and $F(\beta) = \mathbb{E}[-\frac{d^2}{d\beta^2}\, l(Y, \beta)] = Var(s(Y, \beta))$ is the Fisher matrix. Define the partial score functions $s_i(\beta) = s_i(y, \beta) = (l_i'(\beta))^t$ and the partial Fisher matrices $F_i(\beta) = Var(s_i(Y, \beta))$. We have $s(\beta) = \sum_{i=1}^n s_i(\beta)$ and can show by standard calculations that $s_i(\beta) = \mathbf{x}_i^t \cdot D_i(\beta) \cdot [\Sigma_i(\beta)]^{-1} \cdot (y_i - \mu_i(\beta))^t$ with

$$D_i(\beta) = \left[h'((\mathbf{x}_i\beta)^t)\right]^t, \quad \Sigma_i(\beta) = Var_\beta(Y_i), \quad \mu_i(\beta) = h((\mathbf{x}_i\beta)^t),$$

where $h'(z)$ represents the Jacobi matrix $(D_j h_i(z))_{i,j=1,...,q}$. Moreover, $F(\beta) = \sum_{i=1}^n F_i(\beta)$ and $F_i(\beta) = \mathbf{x}_i^t \cdot W_i(\beta) \cdot \mathbf{x}_i$ hold, where $W_i(\beta) = D_i(\beta) \left[\Sigma_i(\beta)\right]^{-1} D_i(\beta)^t$.

We notice that the number of computations for Fisher scoring can be reduced when the number of different covariate levels is smaller than the number of rows of x: Let $g \leq n$ be the number of different rows of x, i.e., we have g covariate levels. We introduce the sets $(r = 1, ..., g)$

$$I_r = \{i \in \{1, ..., n\} : \text{ sample unit } i \text{ possesses covariate level } r\},$$

define n_r to be the number of elements in I_r and assume $i_1 \in I_1$,...,$i_g \in I_g$. We remark that all units with the same covariate level have identical values for $\mu_i(\beta)$, i.e, $\mu_i(\beta) = \mu_j(\beta)$ for $i, j \in I_r$ $(r = 1, ..., g)$. An analog statement holds for $D_i(\beta)$, $\Sigma_i(\beta)$, $W_i(\beta)$ and $F_i(\beta)$. For this reason, we can conclude

$$F(\beta) = \sum_{r=1}^g n_r \cdot F_{i_r}(\beta) \text{ and } s(\beta) = \sum_{r=1}^g \mathbf{x}_{i_r}^t D_{i_r}(\beta) \left[\Sigma_{i_r}(\beta)\right]^{-1} n_r \left[\left(\frac{1}{n_r}\sum_{i\in I_r} y_i^t\right) - \mu_{i_r}(\beta)^t\right].$$

Hence, to obtain $F(\beta)$ and $s(\beta)$, we have to sum up each g terms. When g is considerably smaller than n, the effort to calculate $F(\beta)$ and $s(\beta)$ decreases significantly.

C.2 Fisher scoring in GLMs with stochastic covariates

Consider a GLM with stochastic covariates $(Y, X, \beta, \mathbf{X}, h)$ and assume y and x are observed realizations of Y and X respectively. As usual, let $f_{Y_i|X}(\cdot \mid x)$ denotes the density of Y_i given $X = x$. We have to maximize the function $\beta \mapsto \prod_{i=1}^n f_{Y_i|X}(y_i|x)$. However, this function is the likelihood function corresponding to a GLM with deterministic covariates. Thus, we can apply C.1.

References

[1] Fahrmeir L. / Tutz G. (2010): Multivariate statistical modelling based on generalized linear models. Springer.

[2] Gentle J.E. (1998): Random Number Generation and Monte Carlo Methods. Springer.

[3] Greenberg B.G. / Abul-Ela A.A. / Simmons W.R. / Horvitz D.G. (1969): The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association 64, 520-539.

[4] Groenitz, H. (2012): A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8.

[5] Kuk A.Y.C. (1990): Asking Sensitive Questions Indirectly. Biometrika 77, 436-438.

[6] Maddala G.S. (1983): Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press.

[7] Scheers N.J. / Dayton C.M. (1987): Improved Estimation of Academic Cheating Behavior Using the Randomized Response Technique. Research in Higher Education 26, 61-69.

[8] Scheers N.J. / Dayton C.M. (1988): Covariate Randomized Response Models. Journal of the American Statistical Association 83, 969-974.

[9] Van der Heijden P.G.M. / Van Gils G. (1996): Some logistic regression models for randomized response data. Proceedings of the 11th International Workshop on Statistical Modelling (Orvieto, Italy 15-19 July, 1996), 341-348.

[10] Van den Hout A. / Van der Heijden P.G.M. / Gilchrist R. (2007): The logistic regression model with response variables subject to randomized response. Computational Statistics & Data Analysis 51, 6060-6069.

[11] Warner S.L. (1965): Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60, 63-69.

Supplemental material

```
function [beta, Iter, SE,V_beta, p_beta, fit]=...
    fisherscore1(X,Y,model,C0,BETA0,epsilon)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Supplemental material for the manuscript
% Groenitz, H.: A Covariate Nonrandomized Response Model for
% Multicategorical Sensitive Variables.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%



%This program can be applied to estimate parameters (a) in logistic
%regression models and (b) according to LR-DM estimation

% I N P U T

% X: design matrix. The number of rows in X is the number of different
% covariate levels, the number of columns is the number of covariates.
% Y: response matrix with q+1 columns. The entry Y_ij represents the
% absolute frequency of category j among the units having the i-th
% covariate level.
% model: When intending to analyze an ordinary logistic regression model,
% type 'logreg'. When considering the diagonal model with covariates and
% intending to conduct a LR-DM estimation type 'diagcov'.
% C0: (q+1) x (q+1) design matrix in the diagonal model, every row is a
% left-cyclic shift of the row above (for model 'logreg' an arbitrary
% (q+1) x (q+1) matrix can be typed for C0).
% BETA0: starting values for Fisher scoring algorithm
% epsilon:  accuracy of calculation

% O U T P U T

% beta: vector of estimated parameters (maximum likelihood estimate, MLE)
% Iter: number of iterations of Fisher scoring algorithm
% SE: estimated standard errors for the estimation
% V_beta: estimated variance matrix of the estimator
% p_beta: p-values for the tests with H_0: beta_i=0
% fit=[chi2, pchi2, dev, pdev, df] where
% chi2: value of the test statistic for the chi^2-goodness-of-fit test
% pchi2: p-value for chi^2-goodness-of-fit test
% dev: value of the test statistic for the deviance test (this is another
% well-known goodness-of-fit test, cf. ``Multivariate Statistical Modelling
% Based on Generalized Linear Models'' by Fahrmeir and Tutz (2010),
% Springer, page 50)
% pdev: p-value for deviance test
% df: degrees of freedom for chi^2 / deviance test

% E X A M P L E 1 (estimation in logistic regression models)
% The data of this example are taken from an example in the book "Multivariate
% statistische Verfahren" by Fahrmeir et al. (1996), de Gruyter, page
% 263 / 267 where the sales of gasoline stations are investigated.

% The rows of the following matrix X represent the observed covariate
% levels
% x1=ones(1,12)';x2=[ones(1,6) -1 -1 -1 -1 -1 -1]';
% x3=[1 1 1  -1 -1 -1  1 1 1  -1 -1 -1]';x4=[1 0 -1 1 0 -1 1 0 -1 1 0 -1]';
% x5=[0 1 -1 0 1 -1 0 1 -1 0 1 -1]'; X=[x1 x2 x3 x4 x5];

% Each row of the following matrix Y contains the absolute frequencies of
% the categories 1 (low sales), 2 (medium sales), 3 (large sales) for the
% corresponding covariate level
```

```
% y1=[2 2 3 65 63 48 4 2 5  38 16 179]';y2=[3 0 4 32 24 12 4 0 12 19 7  55]';
% y3=[0 0 1 20  4  6 7 1  4 27 2  29]'; Y=[y1 y2 y3];

% Set
% beta0=zeros(10,1); para=eye(10);
% then the command
% [beta, Iter, SE,V_beta, p_beta, fit]=fisherscore1(X,Y,'logreg',para,beta0,10^-8)
% delivers among others the MLE beta:
% 1.2209 0.3735 -0.5320 -0.9716 0.6174 0.8744 0.3542 0.0978 -0.7246 0.5615

% EXAMPLE 2 (Diagonal model with covariates, LR-DM estimation)

% We introduce the following quantities
% X=[1 1 1 1 1; 1 2 3 4 5]';
% Y=[35 16 30; 27 18 35; 20 22 38; 16 27 36; 15 33 33];
% C0=[2/3 1/6 1/6; 1/6 1/6 2/3; 1/6 2/3 1/6];BETA0=[0 0 0 0; 1 -1 1 -1];
% That is, we have two covariates, and the available covariate levels are
% (1,1),...,(1,5). E.g., for covariate level (1,1), we have 35 respondents
% giving diagonal model answer 1, 16 respondents giving answer 2 and 30
% respondents giving answer 3. The command
% [beta, Iter, SE,V_beta, p_beta, fit]=fisherscore1(X,Y,'diagcov',C0,BETA0,10^-8)
% returns among others the estimate beta equal to
% 3.5691  -1.2722  2.5304  -0.5052

%-------------------------------------------------------------------
q=length(Y(1,:))-1; R=q+1; n=length(X(:,1)); p=length(X(1,:)); nn=sum(Y,2);
if min(nn)==0
    error('n_i equals 0 for some i; Remove corresponding rows in X and Y.')
end

%----- Def. of functions-------------------------------------------
Q =@(z)sum(exp([0 z])); %z row vector
Jh=@(z)( diag(exp(z)*Q(z)) - exp(z')*exp(z)) /(Q(z))^2;
%           D h1
% Jh = [     .      ]   Jacobi matrix of h
%           D hq
h =@(z)exp(z)/Q(z);

CC=C0(1:q,1:q);
for j=1:q
    CC(:,j)=CC(:,j)-C0(1:q,R);
end
m =@(z)h(z) * CC' + C0(1:q,R)';
function M=Jm(z,CC,q,Jh) %Jacobi matrix of m % "nested function"
M=zeros(q); JJ=feval(Jh,z);
for l=1:q
    M=M + CC(:,l)*JJ(l,:);
end
end
%-------------------------------------------------------------------

if strcmp(model,'logreg')   %compares strings
% Here, the case of a logistic regression model is studied.

beta0=BETA0;

for j=1:n
    Y(j,:)=Y(j,:)/nn(j);
end
Y=Y(:,1:q);
b0=beta0;

F=0;score=0;
```

```
for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:)); %block diagonal matrix
    end
    P_i     =(X_i*b0)' ; %predictor
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    F=F + X_i' * W_i * X_i;
    score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
end
b1=b0 + F\score; %A^-1 * b : A\b

Iter=1;

while norm(b1-b0)/norm(b0)>epsilon
    Iter=Iter+1;
    b0=b1;

    F=0;score=0;
    for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:));
    end
    P_i     =(X_i*b0)';  %predictor
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    F=F + X_i' * W_i * X_i;
    score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
    end

    b1=b0 + F\score; % A^-1 * b: A\b
end

beta=b1;

% Standard errors, testing H_0: beta_i=0, goodness-of-fit tests (chi^2 /
% deviance)

chi2=zeros(n,1); dev=zeros(n,1); F=0;
for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:));
    end
    P_i     =(X_i*beta)';  %beta: MLE
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    %for Fisher matrix at the MLE beta
    F=F + X_i' * W_i * X_i;

    %for chi2-goodness-of-fit test
    chi2(i)=(Y(i,:)-mu_i)* inv(Sigma_i) *(Y(i,:)-mu_i)';
```

```
190
191     % for deviance; mnpdf(X,PROB) X and PROB  1-by-k vectors, where k is the
192     % number of multinomial categories
193     Z_i=round( [Y(i,:) 1-sum(Y(i,:))]*nn(i) ); %abs. frequencies
194
195     L1=mnpdf(Z_i, [mu_i 1-sum(mu_i)]);    l1=log(L1);
196     L2=mnpdf(Z_i, Z_i/nn(i));  l2=log(L2);
197     dev(i)=l1-l2;
198 end
199
200 %Estimated standard errors for the components of the MLE
201 SE=sqrt(diag(inv(F)));
202
203 %Estimated variance matrix for the MLE
204 V_beta=inv(F);
205
206 %Testing H_0: beta_i=0 (t-statistics; p-values)
207 T=beta./SE; p_beta=2*(1-normcdf( abs(T) ) );
208
209 % goodness-of-fit
210 CHI2=sum(chi2); DEV=-2*sum(dev);
211 df=n*q-p*q; % degrees of freedom
212 %p-values
213 pCHI2=1-chi2cdf(CHI2,df); pDEV=1-chi2cdf(DEV,df);
214 fit=[CHI2,pCHI2,DEV,pDEV,df];
215
216 end
217
218 % ---------------------------------------------------------
219
220 if strcmp(model,'diagcov')
221 % Case of diagonal model with covariates, LR-DM estimation is conducted.
222
223 YY=Y; %for later calculation of the log-Likelihood
224 for j=1:n
225     Y(j,:)=Y(j,:)/nn(j);
226 end
227 Y=Y(:,1:q);
228
229 E=zeros(length(BETA0(:,1)),p*q+1);
230
231 for jj=1:length(BETA0(:,1))
232     beta0=BETA0(jj,:)';
233
234     b1=beta0;
235     cond=inf;Iter=1;
236     while cond>epsilon
237         Iter=Iter+1;
238         b0=b1;
239
240         F=0;score=0;
241         for i=1:n
242         X_i     =X(i,:);
243         for j=2:q
244          X_i=blkdiag(X_i,X(i,:));
245         end
246         P_i     =(X_i*b0)';
247         D_i     =Jm(P_i,CC,q,Jh)';
248         mu_i    =m(P_i);
249         Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
250         W_i     =D_i * inv(Sigma_i) * D_i';
251
252         F=F + X_i' * W_i * X_i;
```

```
253         score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
254         end
255
256         b1=b0 + F\score; %A^-1 * b = A\b
257         cond=norm(b1-b0)/norm(b0);
258
259         %To avoid endless loops
260         if Iter > 1000
261             b1=ones(p*q,1)*NaN;
262             cond=0;
263         end
264
265     end %endwhile
266
267     beta=b1;
268
269     %Plausibility check
270
271     if sum(isnan(beta))==0 && sum(isinf(beta))==0 && rcond(F)<10^-15
272         beta=ones(p*q,1)*NaN;
273     end
274     %now a beta for this starting value is available
275 E(jj,1:p*q)=beta';
276 mu=zeros(n,q+1);
277 for i=1:n
278     eta_i=zeros(1,q);
279       for j=1:q
280               eta_i(j)=X(i,:)*beta( (j-1)*p+1: j*p);
281       end
282     mu(i,1:q)=m(eta_i);
283 end
284 mu(:,q+1)=1-sum(mu(:,1:q),2);
285 E(jj,p*q+1)=sum(sum(YY.*log(mu)));
286 % value of the log-likelihood (Y: frequencies of the answers)
287 end  %end jj-loop
288
289 % Which starting value leads to the largest likelihood?
290 % The max function ignores NaNs. max([0 1 Nan])=1; max([NaN NaN])=NaN;
291
292 M=max(E(:,p*q+1));
293
294 if isnan(M)==1
295     beta=ones(p*q,1)*NaN;
296 else
297     ind=find(E(:,p*q+1)==M);
298     ind=ind(1);
299     beta=E(ind,1:p*q)';
300 end
301
302 % Standard errors, testing H_0: beta_i=0, goodness-of-fit tests (chi^2 /
303 % deviance)
304
305 chi2=zeros(n,1); dev=zeros(n,1); F=0;
306 for i=1:n
307     X_i     =X(i,:);
308     for j=2:q
309         X_i=blkdiag(X_i,X(i,:));
310     end
311     P_i     =(X_i*beta)';
312     D_i     =Jm(P_i,CC,q,Jh)';
313     mu_i    =m(P_i);
314     Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
315     W_i     =D_i * inv(Sigma_i) * D_i';
```

```
316
317     %for Fisher matrix at the MLE beta
318     F=F + X_i' * W_i * X_i;
319
320     %for chi2-goodness-of-fit test
321     chi2(i)=(Y(i,:)-mu_i)* inv(Sigma_i) *(Y(i,:)-mu_i)';
322
323     % for deviance test; mnpdf(X,PROB) X and PROB  1-by-k vectors, where k is the
324     % number of multinomial categories
325     Z_i=round( [Y(i,:) 1-sum(Y(i,:))]*nn(i) ); %abs. frequencies
326
327     L1=mnpdf(Z_i, [mu_i 1-sum(mu_i)]);    l1=log(L1);
328     L2=mnpdf(Z_i, Z_i/nn(i));  l2=log(L2);
329     dev(i)=l1-l2;
330 end
331
332 %Estimated standard errors for the components of the MLE
333 SE=sqrt(diag(inv(F)));
334
335 %Estimated variance matrix for the MLE
336 V_beta=inv(F);
337
338 %Testing H_0: beta_i=0 (t-statistics, p-values)
339 T=beta./SE; p_beta=2*(1-normcdf( abs(T) ) );
340
341 %for goodness-of-fit tests
342 CHI2=sum(chi2); DEV=-2*sum(dev);
343 df=n*q-p*q; % degrees of freedom
344 %p-values
345 pCHI2=1-chi2cdf(CHI2,df); pDEV=1-chi2cdf(DEV,df);
346 fit=[CHI2,pCHI2,DEV,pDEV,df];
347 end
348
349 end
350
351
352
353
354
355
```