Doi.
Artículo de investigación

Estimación Bayesiana de los parámetros del modelo de regresión probit ordinal aleatorizado

Bayesian estimation of the parameters of the randomized ordinal probit regression model

Estimação Bayesiana dos parâmetros do modelo de regressão probit ordinal aleatorizado

Deisy Lozano-Salado[1] ID. 0000-0003-1056-3774
Flaviano Godínez-Jaimes[1*] ID. 0000-0001-5531-8989
Ramón Reyes-Carreto[1] ID. 0000-0003-4120-5718
María Guzmán-Martínez[1] ID. 0000-0001-9035-2699
Sergio Pérez-Elizalde[2] ID. 0000-0002-1605-0817
Agustín Santiago-Moreno[1] ID. 0009-0004-9101-608X

[1]Maestría en Matemáticas Aplicadas, Universidad Autónoma de Guerrero: Av. Lázaro Cárdenas s/n, Ciudad Universitaria, 39087, Chilpancingo, Guerrero, México.
[2]Departamento de Estadística, Colegio de Postgraduados. Km. 36.5 Carretera México-Texcoco, Montecillo, 56264, Texcoco, Estado de México, México.

*Autor de correspondencia fgodinezj@uagro.mx

Recibido: 12/07/2024

Revisado: 18/08/2024

Aprobado: 21/10/2024

Publicado: 20/12/2024

Resumen

En un gran número de artículos se estima la prevalencia de una variable binaria aleatorizada, pero pocos modelan el efecto de covariables no sensibles en una variable respuesta binaria aleatorizada. En un número pequeño de artículos se estudia una variable ordinal sensible, y no hay artículos en que se modele el efecto de las covariables no sensibles en una variable respuesta ordinal aleatorizada. El objetivo de este trabajo es usar el enfoque Bayesiano para medir el efecto de covariables no sensibles en una variable respuesta ordinal aleatorizada obtenida bajo el diseño de respuesta forzada y evaluar el desempeño de los estimadores propuestos. Se utilizaron cuatro distribuciones *a prioris*: doble Exponencial, Normal, t y Cauchy. Se realizó una simulación para comparar los estimadores Bayesianos estudiados considerando dos números de categorías de la variable respuesta ordinal aleatorizada, dos tamaños de muestra y dos números de covariables no sensibles. Los criterios de comparación fueron el error cuadrático medio, la longitud y la cobertura de los intervalos de credibilidad. Los estimadores Bayesianos propuestos aproximaron adecuadamente los parámetros verdaderos del modelo de regresión probit ordinal aleatorizado, aun cuando se usó el diseño de respuesta forzada inducido por el dispositivo de aleatorización de Hopkins para producir una variable respuesta ordinal aleatorizada. El estimador Bayesiano con la distribución a priori doble exponencial fue el mejor en cuanto a los criterios utilizados.

Palabras clave: variable respuesta ordinal aleatorizada, regresión probit ordinal, inferencia Bayesiana, distribución, doble exponencial.

Abstract

A large number of papers estimate the prevalence of a randomized binary variable, but few model the effect of non-sensitive covariates on a randomized binary response variable. A small number of papers study a sensitive ordinal variable, and there are no papers that model the effect of non-sensitive covariates on a randomized ordinal response variable. The objective of this work is to use the Bayesian approach to measure the effect of non-sensitive covariates on a randomized ordinal response variable obtained under the forced response design and evaluate the performance of the proposed estimators. Four prior distributions were used: Double exponential, Normal, t and Cauchy. A simulation was performed to compare the Bayesian estimators studied considering two numbers of categories of the randomized ordinal response variable, two sample sizes, and two numbers of non-sensitive covariates. The comparison criteria were the mean squared error, the length and the coverage of the credible intervals. The proposed Bayesian estimators adequately estimated the true parameters of the randomized ordinal probit regression model, even when the forced response design induced by the Hopkins randomization device was used to produce a randomized ordinal response variable. The Bayesian estimator using the double exponential prior distribution was the best in terms of the criteria used.

Keywords: randomized ordinal response variable, ordinal probit regression, bayesian inference, double exponential, distribution.

Resumo

Um grande número de artigos estima a prevalência de uma variável binária aleatorizada, mas poucos modelam o efeito de covariáveis não sensíveis em uma variável resposta binária aleatorizada. Um número reduzido de artigos estuda uma variável ordinal sensível, e nenhum modela o efeito de covariáveis não sensíveis em uma variável resposta ordinal aleatorizada. O objetivo deste trabalho é utilizar a abordagem Bayesiana para mensurar o efeito de covariáveis não sensíveis em uma variável resposta ordinal aleatorizada obtida sob um delineamento de resposta forçada e avaliar o desempenho dos estimadores propostos. Quatro distribuições a priori foram utilizadas: exponencial dupla, normal, t e Cauchy. Uma simulação foi realizada para comparar os estimadores Bayesianos estudados, considerando dois números de categorias para a variável resposta ordinal aleatorizada, dois tamanhos de amostra e dois números de covariáveis não sensíveis. Os critérios de comparação foram o erro quadrático médio, o comprimento e a cobertura dos intervalos de credibilidade. Os estimadores Bayesianos propostos aproximaram adequadamente os parâmetros verdadeiros do modelo de regressão probit ordinal aleatorizado, mesmo quando o delineamento de resposta forçada induzido pelo dispositivo de aleatorização de Hopkins foi usado para produzir uma variável de resposta ordinal aleatorizada. O estimador Bayesiano com distribuição a priori exponencial dupla apresentou o melhor desempenho de acordo com os critérios utilizados.

Palavras-chave: Variável de resposta ordinal aleatorizada, Regressão probit ordinal, Inferência Bayesiana, Distribuição exponencial dupla.

Introducción

Los modelos estadísticos comunes permiten medir el efecto de un conjunto de covariables $X_1,\ldots,X_p$ en una variable respuesta Y. Un supuesto fundamental en estos modelos es que la variable respuesta y las covariables se miden sin error. Sin embargo, esto no siempre es posible, especialmente cuando la variable respuesta mide atributos sensibles.

Una variable sensible es la medición de un atributo sensible que se refiere a información personal del entrevistado que no es bien visto por la sociedad, o sobre la práctica de actividades ilícitas como la evasión de impuestos, o la opinión acerca de la práctica del aborto ilegal.

La técnica de respuesta aleatorizada (TRA) es una herramienta para realizar entrevistas diseñada para proteger la privacidad del encuestado y evitar el estigma social o el miedo a las represalias y al mismo tiempo reduce las respuestas deshonestas y el sesgo de no respuesta. En la TRA la respuesta a una pregunta sensible depende de 1) el verdadero estado del entrevistado respecto de la pregunta sensible y 2) el resultado de un dispositivo de aleatorización (Cruyff *et al.*, 2008). El primero en proponer una TRA fue Stanley Warner en 1965. La TRA de Warner tiene a) un mecanismo para obtener información sobre una variable sensible binaria, y b) un estimador de la prevalencia del atributo sensible binario (Warner, 1965). La TRA de Warner consta de dos afirmaciones complementarias: *Tengo el atributo sensible* y *No tengo el atributo sensible*. El entrevistado responde solo a una de las dos afirmaciones y la elección de cuál debe responder se realiza con un dispositivo aleatorio, como un

dado o una ruleta. La elección de la pregunta a contestar se oculta al entrevistador, que solo recibe una respuesta de *Sí* o *No* sin saber qué pregunta fue respondida.

Las variables sensibles pueden ser dicotómicas, politómicas o cuantitativas según el atributo sensible estudiado. En el primer y segundo caso, el parámetro de interés es la proporción de individuos (prevalencia) en cada categoría del atributo sensible. En el último caso el parámetro de interés es la media poblacional.

Varias modificaciones se han propuesto a la TRA de Warner. Una familia de modificaciones cambia la pregunta complementaria por una pregunta incorrelacionada (Greenberg *et al.*, 1969) o por una respuesta forzada (Boruch, 1971). Otra familia de modificaciones cambia la respuesta del entrevistado por una respuesta inocua como *Rojo* (*Verde*) en lugar de *No* (*Sí*) (Kuk, 1990) o por un valor codificado o mezclado (Eichhorn y Hayre, 1983; Bar-Lev *et al.* 2004).

Al usar una TRA se introduce un ruido aleatorio adicional a la respuesta del entrevistado lo que agrega más incertidumbre a los estimadores. Aun así, es posible obtener estimadores insesgados de parámetros unidimensionales, prevalencias o medias, pero estos estimadores tienen mayores varianzas estimadas que en el caso no sensible (Ardah y Oral, 2017).

La mayoría de artículos se enfocan en la estimación de la prevalencia de una variable aleatorizada binaria sensible. Un reducido número de artículos estiman la asociación entre una variable aleatorizada binaria y otra del mismo tipo, o con una variable binaria no sensible (Barabesi *et al.*, 2012; Drane, 1976; Ewemooje y Amahia, 2015; Lee *et al.*, 2013; Tamhane, 1981). Muy pocos artículos estiman el efecto de covariables no sensibles y factores en una variable respuesta binaria sensible. Este escenario es más interesante y desafiante porque el ruido aleatorio añadido mediante el uso de una TRA en la variable respuesta sensible provoca varianzas estimadas muy grandes, además porque es necesario estimar parámetros adicionales.

La regresión logística es el modelo más frecuentemente usado para modelar el efecto de covariables no sensibles en una variable respuesta aleatorizada binaria. Este modelo se ha utilizado cuando la variable aleatorizada binaria se obtuvo bajo los diseños de Warner, de preguntas incorrelacionadas de Greenberg, de respuesta forzada de Boruch y de respuesta enmascarada de Kuk (Scheers y Dayton, 1988; Blair *et al.*, 2015, van der Heijden y van Gils, 1996; Lensvelt-Mulders *et al.*, 2006; van den Hout *et al.*, 2007).

Abul-Ela *et al.* (1967) y Eriksson (1973) fueron los primeros en estimar las prevalencias de las categorías de variables aleatorizadas politómicas. Algunas preguntas que se han utilizado para obtener una variable aleatorizada ordinal son: *¿Cuántos días usó drogas ilegales la semana pasada?*, donde las posibles respuestas son 0, 1, 2 o ≥ 3 (Kim y Warde, 2005), *¿Cuántos abortos tuvo una mujer?*, *¿En cuántas ocasiones se tomaron drogas?*, y *¿Cuánto dinero no reportaste en tu declaración de impuestos?* (Eichhorn y Hayre, 1983). Otras preguntas producen naturalmente variables ordinales aleatorizadas, como *¿Qué tan de acuerdo estás con la interrupción legal del embarazo antes de las doce semanas?*, donde las respuestas pueden ser muy en desacuerdo, en desacuerdo, de acuerdo y muy de acuerdo.

El número de artículos en que se estima la prevalencia de las categorías de variables ordinales aleatorizadas es mucho menor que para el caso binario, incluso hay menos artículos que estiman asociaciones entre una variable ordinal aleatorizada y otra del mismo tipo, o con covariables no sensibles (Kim y Warde, 2005).

El modelo de regresión ordinal se utiliza cuando la variable respuesta es ordinal y no sensible. La estimación de los parámetros de este modelo cuando la variable respuesta es no sensible se ha realizado con ambos paradigmas estadísticos, frecuentista y Bayesiano. Sin embargo, no se encontraron artículos donde la variable respuesta ordinal sea sensible.

Este trabajo tiene como objetivo proponer un modelo para estudiar el efecto de covariables no sensibles en una variable respuesta ordinal aleatorizada bajo el diseño de respuesta forzada, usar la inferencia Bayesiana para estimar los parámetros y evaluar el desempeño de los estimadores propuestos.

El objetivo de este trabajo es usar el enfoque Bayesiano para medir el efecto de covariables no

sensibles en una variable respuesta ordinal aleatorizada obtenida bajo el diseño de respuesta forzada y evaluar el desempeño de los estimadores propuestos.

Material y métodos

Considere que hay un atributo ordinal sensible que se mide por una variable ordinal *Z* con categorías 1, 2,…,J, es decir, tiene un total de J categorías. Si al entrevistado se le hace una pregunta directa sobre su pertenencia a una categoría de *Z*, él o ella puede no responder honestamente o puede no contestar. La información sobre la pertenencia a una de las J categorías de *Z* se puede obtener mediante una TRA que generará una variable ordinal aleatorizada *Y* estrechamente relacionada con *Z*. El uso de la TRA permite motivar la participación del entrevistado al darle confianza para responder su verdadera pertenencia a la categoría de la variable ordinal aleatorizada *Y* y al mismo tiempo proteger su privacidad.

Para cada individuo i, sea $Z_i = j$; $j=1,\ldots,J$, la categoría verdadera pero desconocida para la variable ordinal sensible y $Y_i = j$ la categoría de la variable ordinal aleatorizada observada obtenida utilizando el dispositivo de aleatorización de Hopkins que se describe a continuación.

Dispositivo de aleatorización de Hopkins

Liu y Chow (1976) propusieron un modelo de respuesta aleatorizado cuantitativo discreto utilizando el dispositivo de aleatorización de Hopkins. El dispositivo consiste en una jarra que contiene bolas de dos colores diferentes, por ejemplo, verde y azul, en proporciones g y $1-g$, respectivamente. Las bolas azules están marcadas con los números 1,…,J en proporciones q_1 ,…, q_J (Figura 1). Se le pide al entrevistado que dé la vuelta a la jarra, la agite bien y permita que

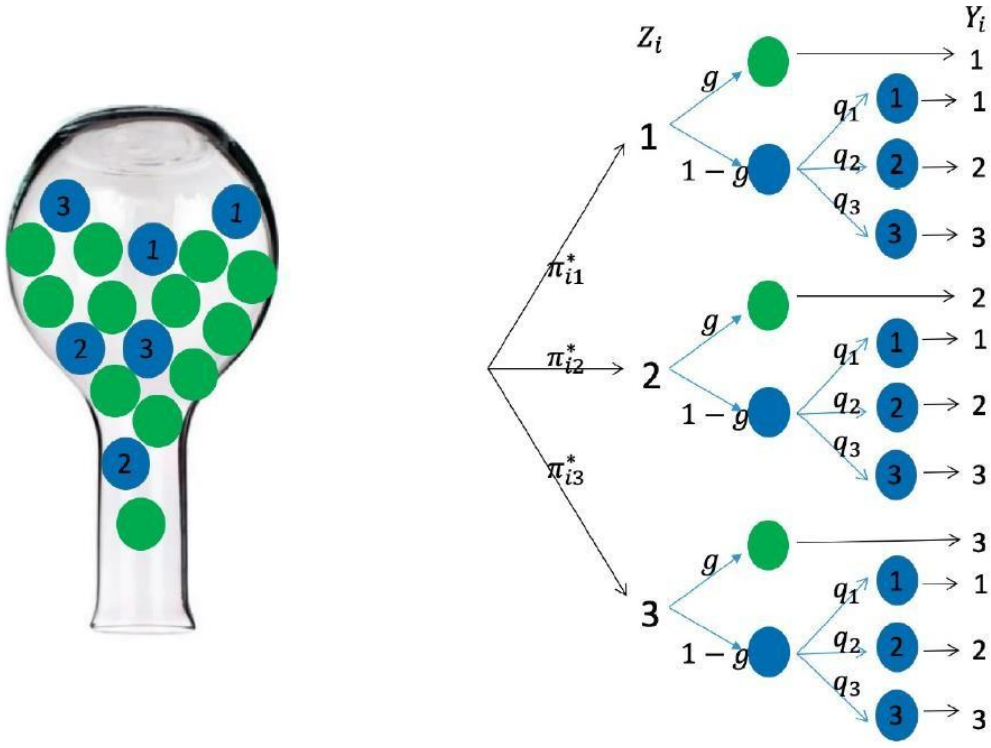

Figura 1: Dispositivo de aleatorización de Hopkins para el diseño de respuesta forzada para variables ordinales.

aparezca una bola en el cuello. Si la bola es verde, el entrevistado debe reportar su verdadera categoría del atributo ordinal sensible, pero si la bola es azul debe informar el número marcado en la bola. q_j tiene el efecto de aumentar las respuestas para la *j*-ésima categoría. Este proceso se lleva a cabo sin la observación del entrevistador lo que protege la privacidad del entrevistado y motiva su participación. El método descrito implementa un diseño de respuesta forzada para obtener respuestas sobre un atributo ordinal sensible.

El dispositivo de aleatorización de Hopkins relaciona la probabilidad de la j-ésima categoría para la variable ordinal sensible para el i-ésimo individuo, $\pi_{ij} = P(Z_i = j)$, a la probabilidad de la j-ésima categoría de la variable ordinal aleatorizada observada, $\pi_{ij} = P(Y_i = j)$ mediante la ecuación:

$$\pi_{ij} = g\pi^*_{ij} + (1-g)q_j \quad (1)$$

donde $i = 1, \dots, n \quad j = 1, \dots, J$

El término g está relacionado con la probabilidad de responder a una pregunta directa y no debe ser cero ni uno. Si $g = 0$, solo se obtienen respuestas forzadas y no se obtiene información del atributo ordinal sensible, por otro lado, si $g = 1$ solo se utilizan preguntas directas sobre el atributo ordinal sensible y el entrevistado sentirá amenazada su privacidad y se negará a contestar, o lo hará de forma deshonesta.

Modelo de regresión probit ordinal

El modelo de regresión probit ordinal permite medir el efecto de las covariables no sensibles en una variable respuesta ordinal no sensible. La motivación del modelo es simple. Se supone que una variable ordinal *Z* es una discretización de una variable latente continua Z^L que es una combinación lineal de covariables no sensibles, $X_1, \dots, X_p$ a la que se agrega un error aleatorio con distribución normal estándar:

$$Z_i^L = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i,$$

donde $\varepsilon_i \sim N(0,1) \quad ; \quad i = 1, \dots, n, ; \; \boldsymbol{x}_i^T = (x_{i1}, \dots, x_{ip})$ es el vector de observaciones correspondientes a *p* covariables no sensibles para el i-ésimo individuo. Sea $\gamma_0, \dots, \gamma_J$ un conjunto de umbrales, la *j*-ésima categoría de *Z* se obtiene cuando $\gamma_{j-1} < Z^L \leq \gamma_j$ para *j=1,…,J*. Para evitar problemas de estimación se supone que $\gamma_0 = -\infty$ y $\gamma_J = \infty$, por lo que la probabilidad de la primera y la última categoría ordinal están bien definidas.

El modelo de regresión probit ordinal está dado por (Kruschke, 2014):

$$\pi^*_{ij} = P(Z_i = j|\boldsymbol{x}_i) = \Phi\left(\frac{\gamma_j - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1} - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right) \quad (2)$$

La Ecuación 2 usa el hecho de que Z^L tiene distribución normal con media $\boldsymbol{x}_i^T\boldsymbol{\beta}$, y desviación estándar σ.

Modelo de regresión probit ordinal aleatorizado

El objetivo de este trabajo es medir el efecto de las covariables no sensibles en la verdadera pero desconocida variable ordinal sensible *Z*, usando la variable ordinal aleatorizada observada *Y* y la relación entre sus probabilidades, π^*_{ij} y π_{ij}, dadas por las Ecuaciones 1 y 2.

El modelo de regresión probit ordinal aleatorizado propuesto es:

$$P(Y_i = j|\boldsymbol{x}_i) = gP(Z_i = j|\boldsymbol{x}_i) + (1-g)q_j$$

$$\pi_{ij} = g\pi_{ij} + (1-g)q_j$$

$$\pi_{ij} = g\left[\Phi\left(\frac{\gamma_j - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\gamma_{j-1} - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right] + (1-g)q_j \quad (3)$$

La función de verosimilitud para el Modelo 3 es:

$$L(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{j=1}^{J} \{I(Y_i = j) \times \pi_{ij}\} \quad (4)$$

donde I(·) es la función indicadora y $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_{J-1}, \gamma_J)^T$

En la revisión de la literatura no se encontraron artículos que aborden la estimación frecuentista o

Bayesiana de los parámetros del modelo de regresión probit ordinal aleatorizado.

Modelación Bayesiana

El teorema de Bayes establece que la distribución posterior del parámetro $\boldsymbol{\theta}$ dados los datos **D**, $\pi(\boldsymbol{\theta}|\mathbf{D})$, es proporcional al producto de la verosimilitud de los datos, $L(\boldsymbol{\theta}|\mathbf{D})$, y la distribución a priori del parámetro $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, es decir:

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) \propto L(\boldsymbol{\theta}|\boldsymbol{D})\pi(\boldsymbol{\theta})$$

Un estimador Bayesiano de $\boldsymbol{\theta}$ es el que minimiza la esperanza bajo la distribución posterior de la función de pérdida. Cuando la función de pérdida es el error cuadrado o lineal absoluta, entonces el estimador Bayesiano de $\boldsymbol{\theta}$ es la media o la mediana de la distribución posterior, según corresponda.

En el modelo de regresión probit ordinal aleatorizado $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma)^T$ y $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$, donde $\boldsymbol{X}$ es la matriz diseño que contiene la información de las covariables no sensibles. La distribución posterior es:

$$\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{y}) \propto L(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{y})\, \pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma})$$

La verosimilitud de los datos bajo el diseño de respuesta forzada se da en las Ecuaciones 3 y 4. Obtener la media de la distribución posterior no es analíticamente posible. En su lugar, la distribución posterior se aproxima por Cadenas de Markov Monte Carlo (MCMC, por sus siglas en inglés). Para mejorar la convergencia de las MCMC se estandariza $\boldsymbol{X}$. Sea $\boldsymbol{X}^*$ la matriz estandarizada, es decir, la k-ésima columna es dada por $x^*_{ik} = (x_{ik} - \bar{x}_k)/s_{x_k}$, donde $\bar{x}_k$ y s_{x_k} son la media muestral y la desviación estándar para la j-ésima covariable. Para evitar problemas de identificabilidad en la estimación de los umbrales, dos de ellos se fijan: $\gamma_1 = 1.5$ y $\gamma_{J-1} = J-0.5$ (Kruschke, 2014). Asociado con las covariables estandarizadas, se introduce un nuevo vector de parámetros $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$.

La distribución posterior se convierte en:

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma|\boldsymbol{X}^*, \boldsymbol{y}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma|\boldsymbol{X}^*, \boldsymbol{y}) \times \pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma)$$

La distribución a priori conjunta se da como un producto de distribuciones marginales *a prioris* independientes:

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma) = \pi(\alpha_0) \prod_{k=1}^{p} \pi(\alpha_k) \times \prod_{j=1}^{J-1} \pi(\gamma_j) \times \pi(\sigma) \qquad (5)$$

Las estimaciones de parámetros unidimensionales como son la prevalencia o media de variables sensibles después de usar alguna de las diferentes TRA tienen varianzas estimadas grandes. Un efecto similar ocurrirá al estimar vectores de parámetros como los efectos de covariables no sensibles en una variable respuesta ordinal aleatoria. Esto motiva a usar *a prioris* no informativas o *a prioris* que reduzcan las estimaciones de los parámetros α_j.

Se estudian cuatro *a prioris* conjuntas, todas ellas usan $\alpha_0 \sim N(\frac{1+J}{2}, J^2)$, $\sigma \sim N(\frac{J}{1000}, 10J)$, y $\gamma_j \sim N(j + 0.5, J^2)$; $j = 2,\ldots,J\text{-}2$. La diferencia entre las cuatro a prioris conjuntas es la distribución de los α_k, k = 1,…,p. Estas distribuciones *a prioris* son Normal (N), Doble exponencial (DE), *t* y Cauchy. La lista de las distribuciones a prioris utilizadas y los parámetros adicionales necesarios son:

$$\alpha_k \sim N(0, J^2) \qquad (6)$$

$$\alpha_k \sim DE(0, \lambda), \qquad \lambda \sim U(0.001, 10) \qquad (7)$$

$$\alpha_k \sim t\,(0,\ \sigma_t^2, v^* + 1), \quad v^* \sim Exp(1/29), \quad \sigma_t \sim U(0.001, 1000) \qquad (8)$$

$$\alpha_k \sim t\,(0,\ \sigma_t^2, 1), \qquad \sigma_t \sim U(0.001, 10 \qquad (9)$$

En la literatura, las *a prioris* para los α_j que se usan cuando se modela una variable respuesta ordinal no sensible tienen una distribución normal con media cero y varianzas 100, 0.5 o 0.1 (Van, 2017; Xie *et al.*, 2009). La distribución a priori $\alpha_k \sim N(0, J^2)$ es vaga en la escala estandarizada para el modelo de regresión probit ordinal aleatorizado y la varianza del hiperparámetro es totalmente definido por el número de categorías. La distribución a priori 7 se usa comúnmente en estimación Lasso Bayesiana para reducir las

pendientes estimadas porque su masa se concentra cerca de cero y el hiperparámetro λ toma cualquier valor en el intervalo (0.001, 10). La distribución a priori 9 es un caso especial de las distribuciones a priori 8 cuando los grados de libertad son 1. Ambas *a prioris* tienen media cero, de modo que σ_t^2 está en un rango amplio de 10^{-6} a 10^6.

Después de obtener los $\alpha's$ en las MCMC, los parámetros en la escala original se obtuvieron con las ecuaciones:

$$\beta_0 = \alpha_0 - \sum_{k=1}^{p} \frac{\alpha_k}{s_{x_k}} \bar{x}_k,$$
$$\beta_k = \frac{\alpha_k}{s_{x_k}} \qquad k = 1, \dots, p$$

Estudio de simulación

El desempeño de los estimadores Bayesianos bajo las distribuciones *a prioris* estudiadas se evaluó mediante simulación Monte Carlo. El dispositivo de aleatorización de Hopkins que induce el diseño de respuesta forzada para un atributo ordinal sensible se usó con $g = 2/3$ y $q_1 = \cdots = q_J = \frac{1}{J}$ lo que induce respuestas directas con probabilidad 2/3.

Factores estudiados. Muchos factores pueden afectar el rendimiento de los estimadores Bayesianos del modelo de regresión probit ordinal aleatorizado, pero en esta investigación solo se consideran tres. Los factores estudiados en la simulación Monte Carlo fueron el número de categorías de la variable dependiente, el tamaño de la muestra y el número de covariables no sensibles.

Número de categorías de la variable respuesta ordinal aleatorizada, J. Se consideraron tres y cuatro categorías ordinales, *J*=3,4.

Tamaño de la muestra, n. En el enfoque frecuentista cuando el tamaño de la muestra es grande, el estimador de máxima verosimilitud tiene propiedades óptimas. Además, en el enfoque Bayesiano cuando el tamaño de la muestra es pequeño, la distribución a priori domina la distribución posterior y cuando el tamaño de la muestra es grande, lo hace la verosimilitud de los datos. Un tamaño de la muestra pequeño determinó de acuerdo con el número de parámetros en el modelo, y el tamaño de la muestra grande como el triple del tamaño de muestra pequeño.

Número de covariables no sensibles, p. Se estudiaron una y dos covariables no sensibles, *p*=1,2.

Por lo tanto, cuando *Y* tiene *J*=3 categorías ordinales y el modelo *p*=1 covariable no sensible, los tamaños de muestra fueron *n*=125, 375 y si *p*=2 los tamaños de muestra fueron *n*=150, 450. Cuando *Y* tiene *J*=4 categorías ordinales y el modelo tenía *p*=1 covariable no sensible, los tamaños de muestra eran *n*=150, 450, y si *p*=2, los tamaños de muestra eran *n*=175, 525 (Tabla 1).

Generación de datos

Las *p* covariables no sensibles X_k se generaron de forma independiente de una distribución uniforme en (0.9, 1.9). Los parámetros verdaderos se eligieron de acuerdo con el número de covariables no sensibles en el modelo, si *p*=1

Tabla 1: Tamaños de muestra, parámetros y umbrales utilizados en la simulación.

Número de categorías, *J*	**Número de covariables no sensibles, *p***	
	1	2
3	*n* = 125,375 $\boldsymbol{\beta} = (-10,10)^T$ $\boldsymbol{\gamma} = (-\infty,3.7,6.2,\infty)^T$	*n* = 150,450 $\boldsymbol{\beta} = (-10,10,5)^T$ $\boldsymbol{\gamma} = (-\infty,2.6,5.0,6.7,\infty)^T$
4	*n* = 150,450 $\boldsymbol{\beta} = (-10,10)^T$ $\boldsymbol{\gamma} = (-\infty,2.6,5.0,6.7,\infty)^T$	*n* = 175,525 $\boldsymbol{\beta} = (-10,10,5)^T$ $\boldsymbol{\gamma} = (-\infty,9.6,11.8,13.5,\infty)^T$

entonces $\boldsymbol{\beta}$ = (-10,10)^T y si p=2 entonces $\boldsymbol{\beta}$ = (-10,10,5)^T. De esta manera, $\beta_1 = 10$ y $\beta_2 = 5$ están lejos de cero y se espera que los estimadores Bayesianos no sean afectados por el efecto de contracción producido por las *a prioris*. Los umbrales se definen apropiadamente para que todas las categorías tengan observaciones (Tabla 1).

La variable latente continua se obtuvo con $Z_i^L = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim N(0,1/2^2)$. Después la variable ordinal sensible Z se obtuvo utilizando los umbrales.

Para simular el proceso de asignar al entrevistado a responder la pregunta directa con probabilidad g, o a dar una respuesta forzada con probabilidad $1-g$, se generó una variable W con distribución Bernoulli con parámetro g. Es decir, para cada $Z_i = j$, se generó una variable auxiliar W distribuida Bernoulli con parámetro g. Si W=1 entonces $Y_i = j$, de lo contrario Y_i es una realización de una distribución multinomial con parámetros $\pi_1 = \cdots = \pi_J = \frac{1}{J}$.

Se estudiaron cuatro estimadores Bayesianos, definidos por la distribución a priori usada para α_j. El estimador Bayesiano EB.N se obtiene al usar la a priori Normal (6), el EB.DE se obtiene con la a priori Doble Exponencial (7), el EB.T se obtiene con la a priori t (8) y el EB.C con la a priori Cauchy (9).

Las distribuciones marginales posteriores se obtuvieron con el paquete R2jags (Su y Yajima, 2015) implementado en el lenguaje R (R Core Team, 2016). Las MCMC se obtuvieron con tres cadenas de 10000 iteraciones, un quemado de 5000, y un adelgazamiento de uno de cada cinco. Aunque no se muestran los resultados, la convergencia de MCMC se verificó mediante el factor de reducción de escala potencial multivariado de Gelman y Brooks implementado en el paquete coda (Plummer *et al.*, 2006). Los intervalos de credibilidad del 95% para β_k,

$$ICr\,[\beta_k] = [ICrI\,(\beta_k), \quad ICrS\,(\beta_k)], \qquad k=1,\ldots,p$$

donde $ICrI\,(\beta_k)$ es el cuantil 0.025 de la distribución posterior marginal de β_k y $ICrS\,(\beta_k)$ el cuantil 0.975. El proceso se repitió R = 1000 veces.

Desempeño de los estimadores

Un criterio para comparar estimadores es el error cuadrático medio (ECM) porque mide la dispersión de las estimaciones respecto al valor verdadero del parámetro y tiene la propiedad de que es la suma del cuadrado del sesgo y la varianza del estimador. Otros criterios usados son dos propiedades de los intervalos de credibilidad que son la longitud (L) y cobertura (C). Es deseable que los intervalos de credibilidad sean de la menor longitud para que sean más informativos y que la proporción de intervalos que contienen el parámetro sea aproximadamente del valor nominal, por ejemplo 95%. Estos criterios se definen como

$$\hat{ECM}(\beta) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{p}\sum_{k=1}^{p}(\beta_{k,r} - \beta_k)^2$$

$$L(\beta) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{p}\sum_{k=1}^{p}\{LS\,(\beta_{k,r}) - LI\,(\beta_{k,r})\}$$

$$C(\beta) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{p}\sum_{k=1}^{p}[LI\,(\beta_{k,r}) \leq \beta_k \leq \tilde{LS}\,(\beta_{k,r})]$$

donde $\beta_{k,r}$ es el r-ésimo valor del estimador β_k de β_k.

Para comparar un estadístico T bajo dos condiciones, A y B, se utilizó el porcentaje de ganancia relativa (PGR) del estadístico cuando se pasa de la condición A a la condición B:

$$PGR.T\,(A,B) = \frac{T\,(A) - T\,(B)}{T\,(B)} \times 100$$

donde T corresponde a los criterios $\hat{ECM}$, L y C.

Resultados

Los resultados de la estimación del intercepto no se informan porque interesa solo el efecto de las covariables no sensibles en la variable respuesta ordinal aleatorizada. La Tabla 2 muestra el ECM,

la longitud y la cobertura de los estimadores Bayesianos y sus intervalos de credibilidad.

Efecto del número de categorías

Todos los estimadores Bayesianos tienen ECM y longitud de intervalo creíble más pequeño cuando la variable respuesta ordinal aleatorizada tiene cuatro categorías (Tabla 2).

La Tabla 3 muestra el porcentaje de ganancia relativa para los tres criterios comparando J=3 y J=4 categorías de la variable respuesta ordinal aleatorizada. El efecto del número de categorías es muy fuerte en el ECM de los estimadores Bayesianos porque el porcentaje de ganancia relativa es mayor, entre 94% y 619%, cuando Y tiene tres categorías. También es fuerte el efecto en la longitud de los intervalos de credibilidad en los estimadores Bayesianos pues el porcentaje de ganancia relativa es mayor, entre 42% y 142%, cuando Y tiene tres categorías. Y finalmente, es débil en la cobertura de intervalos de credibilidad ya que el porcentaje de ganancia relativa suele ser mayor cuando Y tiene cuatro categorías.

Efecto del tamaño de muestra

Todos los estimadores Bayesianos tienen menor ECM, longitud y cobertura de intervalos de credibilidad cuando el tamaño de la muestra es grande (Tabla 2).

La Tabla 4 muestra el PGR para los tres criterios comparando el tamaño de muestra pequeño (TMP) con el tamaño de muestra grande (TMG). De acuerdo con la teoría asintótica, los estimadores Bayesianos muestran un mejor desempeño con tamaño de muestra grande. El efecto del tamaño de la muestra es muy fuerte en el ECM de los estimadores Bayesianos porque el porcentaje de ganancia relativa es entre 226% y 883% mayor que con el tamaño de muestra pequeño. El efecto del tamaño de la muestra es fuerte en la longitud de los intervalos de credibilidad de los estimadores Bayesianos porque el porcentaje de ganancia relativa es entre 87% y 154% mayor cuando el tamaño de la muestra es pequeño. Finalmente, el efecto del tamaño de la muestra es débil en la cobertura de los intervalos de credibilidad de los estimadores Bayesianos lo que

Tabla 2: Valores de los criterios utilizados para comparar los estimadores Bayesianos bajo los factores estudiados.

p	J	n	Estimador	$\hat{ECM}$	L	C
1	3	125	EB.N	3.66	6.91	0.93
			EB.DE	3.99	7.14	0.94
			EB.T	7.33	8.37	0.91
			EB.C	7.14	8.30	0.91
		375	EB.N	0.80	3.40	0.95
			EB.DE	0.76	3.40	0.95
			EB.T	0.90	3.50	0.95
			EB.C	0.90	3.50	0.95
	4	150	EB.N	0.96	3.42	0.94
			EB.DE	0.85	3.35	0.94
			EB.T	1.02	3.46	0.93
			EB.C	1.01	3.46	0.93
		450	EB.N	0.20	1.70	0.95
			EB.DE	0.20	1.70	0.95
			EB.T	0.20	1.70	0.95
			EB.C	0.20	1.70	0.95
2	3	150	EB.N	2.06	4.72	0.93
			EB.DE	1.89	4.66	0.94
			EB.T	3.35	5.42	0.92
			EB.C	3.44	5.47	0.92
		450	EB.N	0.35	2.15	0.95
			EB.DE	0.30	2.10	0.95
			EB.T	0.35	2.15	0.95
			EB.C	0.35	2.15	0.95
	4	175	EB.N	0.56	2.77	0.95
			EB.DE	0.51	2.73	0.95
			EB.T	0.56	2.78	0.95
			EB.C	0.56	2.78	0.95
		525	EB.N	0.16	1.49	0.95
			EB.DE	0.16	1.48	0.95
			EB.T	0.16	1.49	0.95
			EB.C	0.16	1.49	0.95

Tabla 3: Porcentaje de ganancia relativa en ECM, L y C al comparar tres contra cuatro categorías sensibles.

Estimador	***p***	***n***	$PGR.\hat{E}CM$	$PGR.L$	$PGR.C$
EB.N	1	125:150	281	102	-1
EB.DE			369	113	0
EB.T			619	142	-2
EB.C			607	140	-2
EB.N		375:450	300	100	0
EB.DE			280	100	0
EB.T			350	106	0
EB.C			350	106	0
EB.N	2	150:175	271	70	-3
EB.DE			273	71	-1
EB.T			503	95	-4
EB.C			520	97	-4
EB.N		450:525	119	45	0
EB.DE			94	42	1
EB.T			119	45	0
EB.C			119	45	0

$$\text{PGR}.T\ (3,4) = 100\ (T\ (3) - T\ (4))/T\ (4)$$

se evidencia en que la cobertura es mayor más frecuentemente con el tamaño de muestra grande, y el porcentaje de ganancia relativa varía entre 1 y 4%.

Efecto del número de covariables no sensibles

Todos los estimadores Bayesianos tienen mejor ECM, longitud y cobertura de intervalos de credibilidad cuando hay dos covariables no sensibles (Tabla 2).

La Tabla 5 muestra el porcentaje de ganancia relativa para los tres criterios comparando p=1 con p=2. El efecto del número de covariables no sensibles en el ECM de los estimadores Bayesianos es muy fuerte, pues es mayor cuando hay una covariable que cuando hay dos covariables, porque el porcentaje de ganancia relativa varía entre un 25 % y un 157 %. Este efecto del número de covariables no sensibles es fuerte en la longitud de los intervalos de credibilidad de los estimadores Bayesianos, pues es mayor cuando hay una covariable, ya que el porcentaje de ganancia relativa varía entre 14% y 63%. Y finalmente, es débil en la cobertura de los intervalos de credibilidad de los estimadores Bayesianos, puesto que es más frecuente que el porcentaje de ganancia relativa sea mayor con dos covariables no sensibles.

Comparación del estimador de máxima verosimilitud con los estimadores Bayesianos estudiados

Existen dos enfoques importantes para realizar inferencia, el frecuentista y el Bayesiano. El estimador de máxima verosimiltud (EMV) es uno de los métodos más importantes en el enfoque frecuentista. En el contexto del modelo de regresión probit ordinal aleatorizado es natural comparar el desempeño de los estimadores, de máxima verosimiltud y Bayesianos, en función de sus esperanzas, sus intervalos de confianza (IC) o de credibilidad (ICr), su longitud y su cobertura.

Tabla 4: Porcentaje de ganancia relativa en ECM, L y C cuando se compara el tamaño de muestra pequeño con el grande.

Estimador	***p***	***J***	$PGR.\hat{E}CM$	$PGR.L$	$PGR.C$
EB.N	1	3	358	103	-2
EB.DE			425	110	-1
EB.T			714	139	-4
EB.C			693	137	-4
EB.N		4	380	101	-1
EB.DE			325	97	-1
EB.T			410	104	-2
EB.C			405	104	-2
EB.N	2	3	489	119	-3
EB.DE			528	122	-1
EB.T			856	152	-4
EB.C			883	154	-4
EB.N		4	247	87	0
EB.DE			226	85	1
EB.T			247	87	0
EB.C			247	87	0

$$\text{PGR.}T\ (TMP, TMG) = 100\ (T\ (TMP) - T\ (TMG))/T\ (TMG)$$

Se realizó una simulación con R = 1000 repeticiones en el escenario en que la variable respuesta ordinal aleatorizada tiene 4 categorías que se explican con dos variables independientes y se usa un tamaño de muestra de 525 que es el mayor de los tamaños de muestra considerados en este trabajo. Este tamaño de muestra se puede considerar grande y de acuerdo con la teoría asintótica, el EMV tiene distribución normal con media el vector de parámetros verdadero y matriz de covarianzas la inversa de la matriz de información de Fisher, esto es, se espera que el EMV tenga un excelente desempeño.

La esperanza e intervalos de confianza (IC) del $100(1-\alpha)\%$ para el EMV de β_k son:

$$E\ (\beta_k) = \frac{1}{R}\sum_{r=1}^{R} \beta_{k,r}$$

$$\beta_k \pm\ Z_{1-\alpha/2}\, EE(\beta_k)$$

donde $Z_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución normal estándar y $EE(\beta_k)$ es el elemento jj de la inversa de la matriz de Fisher. Los límites de confianza inferior y superior de β_k son $LCI(\beta_k) = \beta_k - Z_{1-\alpha/2}\, EE(\beta_k)$ y $LCS(\beta_k) = \beta_k + Z_{1-\alpha/2}\, EE(\beta_k)$.

Aunque no se muestran, los valores del ECM de los estimadores Bayesianos estudiados son similares a los reportados en la Tabla 2. Los datos se generaron con $\beta_1 = 10$ y $\beta_2 = 5$ por lo que la esperanza de los EMV, 13.97 y 6.38, sobreestiman a los valores verdaderos, y muestra que los EMV de los parámetros del modelo de regresión probit ordinal aleatorizado no son insesgados. Por el contrario, los estimadores Bayesianos son más cercanos a los valores verdaderos y el mejor de ellos es EB.DE, que usa la distribución a priori doble exponencial.

Tabla 5: Porcentaje de ganancia relativa en ECM, L y C al comparar una contra dos covariables sensibles.

Estimador	***J***	***n***	$PGR.\hat{E}CM$	$PGR.L$	$PGR.C$
EB.N	3	125:150	78	47	1
EB.DE			112	53	0
EB.T			119	55	-1
EB.C			108	52	-1
EB.N		375:450	129	58	1
EB.DE			153	62	0
EB.T			157	63	1
EB.C			157	63	1
EB.N	4	150:175	73	23	-1
EB.DE			68	23	-1
EB.T			84	25	-2
EB.C			82	25	-2
EB.N		450:525	25	14	1
EB.DE			29	15	1
EB.T			25	14	1
EB.C			25	14	1

$$\text{PGR}.T\ (1,2) = 100\ (T\ (1) - T\ (2))/T\ (2)$$

La sobreestimación de los EMV causa que la media del $LCI(\beta_1)$ sea 12.10 que es mayor que el valor verdadero $\beta_1 = 10$ y como consecuencia la cobertura del IC para β_1 es apenas de 0.04, muy lejos del valor nominal de 0.95. Por el contrario, los estimadores Bayesianos tienen ICr con medias del $LCrI(\beta_1)$ y $LCrS(\beta_1)$ que permiten que el valor verdadero $\beta_1 = 10$ esté en el ICr lo que se refleja en que su cobertura sea de 0.93 o 0.94 que son más cercanos al valor nominal de 0.95.

Las medias de $LCI(\beta_2)$ y $LCS(\beta_2)$ son 4.74 y 8.03 que permiten que el valor verdadero $\beta_2 = 5$ y esté contenido en el IC. Estos IC tienen cobertura de 0.62 que sigue siendo menor al valor nominal de 0.95. Por el contrario, los estimadores Bayesianos tienen ICr con medias de $LCrI(\beta_2)$ y $\tilde{LCrS}(\beta_2)$ que permiten que el valor verdadero $\beta_2 = 5$ esté en el ICr lo que se refleja en que su cobertura sea de 0.94 o 0.95 que coinciden con el valor nominal de 0.95.

El otro criterio usado para comparar los estimadores fue la longitud de los IC e ICr. Los EMV tienen longitud aproximadamente el doble de la longitud de los ICr para β_1 y de aproximadamente 2.6 veces mayores para β_2.

Aún en el escenario más favorable para el EMV por tener un tamaño de muestra grande, se observa que los estimadores Bayesianos son mejores pues son aproximadamente insesgados, con menores longitudes y cobertura de los ICr.

Conclusiones

Los estimadores Bayesianos propuestos estiman correctamente los parámetros del modelo de regresión probit ordinal aleatorizado aun cuando se usa el diseño de respuesta forzada inducido por el dispositivo de aleatorización de Hopkins para producir una variable respuesta ordinal aleatorizada. El valor nominal de cobertura del 95% de los intervalos de credibilidad se alcanzó

Tabla 6. Medias de las esperanzas, límites inferiores y superiores de los intervalos de confianza y de credibilidad, longitud y cubrimiento de los estimadores de máxima verosimilitud y Bayesianos del modelo de regresión probit ordinal aleatorizado.

	$E(\beta_1)$	$E(\beta_2)$	$LCI(\beta_1)$	$LCS(\beta_1)$	$LCI(\beta_2)$	$LCS(\beta_2)$	$L(\beta_1)$	$L(\beta_2)$	$C(\beta_1)$	$C(\beta_2)$
MV	**13.97**	**6.38**	**12.10**	**15.85**	**4.74**	**8.03**	**3.75**	**3.29**	**0.04**	**0.62**
	$E(\beta_1)$	$E(\beta_2)$	$LCrI(\beta_1)$	$LCrS(\beta_1)$	$LCrI(\beta_2)$	$LCrS(\beta_2)$	$L(\beta_1)$	$L(\beta_2)$	$C(\beta_1)$	$C(\beta_2)$
EB.N	10.08	5.06	9.27	10.99	4.47	5.70	1.72	1.23	0.93	0.95
EB.DE	10.04	5.03	9.24	10.94	4.45	5.67	1.70	1.22	0.94	0.95
EB.T	10.08	5.05	9.27	10.98	4.47	5.70	1.72	1.23	0.93	0.94
EB.C	10.08	5.05	9.27	10.98	4.47	5.70	1.72	1.23	0.93	0.95

generalmente con los tamaños de muestra más grandes, lo que es consistente con la teoría asintótica. Un resultado adicional es que todas las covariables no sensibles incluidas en el modelo de regresión probit ordinal aleatorizado rechazan que $\beta_k \neq 0$, j=1,…, p porque las longitudes de los intervalos de credibilidad indican que las estimaciones β_k están lejos de cero.

Los estimadores Bayesianos de los β_k de las covariables no sensibles en el modelo de regresión ordinal aleatorizado tienen óptimo error cuadrático medio, longitud y cobertura de los intervalos de credibilidad cuando se utilizan más categorías de la variable respuesta ordinal aleatorizada, tamaño de muestra más grande y mayor número de covariables no sensibles.

Sorprendentemente, los estimadores Bayesianos tienen menor error cuadrático medio usando la a priori doble exponencial para cada coeficiente de las covariables no sensibles. Por lo tanto, el mejor estimador Bayesiano en el modelo de regresión probit ordinal aleatorizado se obtiene utilizando la a priori doble exponencial.

Referencias

Abul-Ela, A.L.A., Greenberg, G.G., Horvitz, D.G. (1967). A Multiproportions Randomized Response Model. Journal of the American Statistical Association, 62, 990-1008. https://doi.org/10.2307/2283687

Ardah, H.I. Oral, E. (2017). Model Selection in Randomized Response Techniques for Binary Responses. Communications in Statistics - Theory and Methods, 47, 3305-3323. https://doi.org/10.1080/03610926.2017.1353626

Bar-Lev, S.K., Bobovitch, E., Boukai, B. (2004). A Note on Randomized Response Models for Quantitative Data. Metrika, 60, 225-250. https://doi.org/10.1007/s001840300308

Barabesi, L., Franceschi, S., Marcheselli, M. (2012). A Randomized Response Procedure for Multiple Sensitive Questions. Statistical Papers, 53, 703-718.

Blair, G., Imai, K., Zhou, Y.-Y. (2015). Design and Analysis of the Randomized Response Technique. Journal of the American Statistical Association, 110, 1304-1319. https://doi.org/10.1080/01621459.2015.1050028

Boruch, R.F. (1971). Assuring Confidentiality of Responses in Social Research: a Note on Strategies. The American Sociologist, 6, 308-311.

Cruyff, M.J.L.F., van den Hout, A., van der Heijden, P.G.M. (2008). The Analysis of Randomized Response Sum Score Variables. Journal of the Royal Statistical Society, Series B, 71, 21-30.

http://dx.doi.org/10.1111/j.1467-9868.2007.00624.x

Drane, W. (1976). N the Theory of Randomized Responses to Two Sensitive Questions. Communications in Statistics - Theory and Methods, 5, 565-574. https://doi.org/10.1080/03610927608827375

Eichhorn, B.H., Hayre, L.S. (1983). Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. Journal of Statistical planning and Inference, 7, 307–316.

Eriksson, S.A. (1973). A New Model for Randomized Response. International Statistical Review, 41, 101-113. https://doi.org/10.2307/1402791

Ewemooje, O.S., Amahia, G.N. (2015). Improved Randomized Response Technique for Two Sensitive Attributes. Afrika Statistika, 10, 839-853. https://doi.org/10.16929/as/2015.639.78

Greenberg, B., Abul-Ela, A., Simmons, W., Horvitz, D.G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association, 64, 520-539. https://doi.org/10.2307/2283636

Kim, J.-M., Warde, W.D. (2005). Some New Results on the Multinomial Randomized Response Model. Communications in Statistics. Theory and Methods, 847-856. https://doi.org/10.1081/STA-200054378

Kruschke, J.K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Academic Press.

Kuk, A.Y.C. (1990). Asking Sensitive Question Indirectly. Biometrika, 77, 436-438. https://doi.org/10.1093/biomet/77.2.436

Lee, C.S., Sedory, S.A., Singh, S. (2013). Estimating at Least Seven Measures of Qualitative Variables from a Single Sample using Randomized Response Technique. Statistics and Probability Letters, 83, 399-409. https://doi.org/10.1016/j.spl.2012.10.004

Lensvelt-Mulders, G.J., van der Heijden, P.G., Laudy, O., van Gils, G. (2006). A Validation of a Computer-Assisted Randomized Response Survey to Estimate the Prevalence of Fraud in Social Security. Journal of the Royal Statistical Society A., 169, 305-318. A Validation of a Computer-Assisted Randomized Response Survey to Estimate the Prevalence of Fraud in Social Security on JSTOR

Liu, P.T., Chow, L.P. (1976). A new discrete quantitative randomized response model. ACM SIGSIM Simulation Digest, 7, 30-31. https://doi.org/10.1080/01621459.1976.10481479

Plummer, M., Best, N., Cowles, K., Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. R News, 6, 7-11. CODA: Convergence diagnosis and output analysis for MCMC

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Scheers, N., Dayton, C. (1988). Covariate Randomized Response Models. Journal of the American Statistical Association, 83, 969-974. https://doi.org/10.1080/01621459.1988.10478686

Su, Y.-S., Yajima, M. (2015). R2jags: Using R to run JAGS.

Tamhane, A. (1981). Randomized Response Techniques for Multiple Sensitive Attributes. Journal of the American Statistical Association, 76, 916-923. https://doi.org/10.1080/01621459.1981.10477741

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. https://doi.org/10.1016/j.csda.2006.12.002

van der Heijden, P., van Gils, G. (1996). Some logistic regression models for randomized response data. In Proceedings of the 11th International Workshop on Statistical Modeling, 341-348.

Van, M.H. (2017). Determinants of the Levels of Development Based on the Human Development Index: Bayesian Ordered Probit Model. International Journal of Economics and Financial Issues, 7, 425.

Warner, S. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, 63-69. https://doi.org/10.1080/01621459.1965.10480775

Xie, Y., Zhang, Y., Liang, F. (2009). Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. Journal of Transportation Engineering, 135, 18-25.