Statistica Sinica Preprint No: SS-2015-0338R3	
Title	Regression Analysis with Response-selective Sampling
Manuscript ID	SS-2015-0338R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0338
Complete List of Authors	Yuan Yao Kani CHEN Yuanyuan Lin and Chaoxu Zhou
Corresponding Author	Yuan Yao
E-mail	yaoyuan@hkbu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Regression Analysis with Response-selective Sampling

Kani Chen

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Yuanyuan Lin

Department of Statistics, The Chinese University of Hong Kong, Hong Kong

Yuan Yao

Department of Mathematics, Hong Kong Baptist University, Hong Kong

Chaoxu Zhou

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Abstract: Response-selective sampling, in which samples are drawn from a population according to the values of the response variable, is common in biomedical, epidemiological, economic and social studies. This paper proposes to use transformation models, the generalized accelerated failure time models in econometrics, for regression analysis with response-selective sampling. With unknown error distribution, the transformation models are broad enough to cover linear regression models, Cox's model, and the proportional odds model as special cases. To the best of our knowledge, except for the case-control logistic regression, there is presently no prospective estimation approach that can work for biased sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

without modification. We prove that the maximum rank correlation estimation is valid for response-selective sampling and establish its consistency and asymptotic normality. Unlike inverse probability methods, the proposed method of estimation does not involve sampling probabilities, which are often difficult to obtain in practice. Without the need of estimating the unknown transformation function or the error distribution, the proposed method is numerically easy to implement with the Nelder-Mead simplex algorithm that does not require convexity or continuity. We propose an inference procedure using random weighting to avoid the complication of density estimation when using the plug-in rule for variance estimation. Numerical studies with supportive evidence are presented. Application is illustrated with the Forbes Global 2000 data.

Key words and phrases: General transformation model; Maximum rank correlation; Random weighting; Response-selective sampling.

1. Introduction

Response-selective sampling is commonly used in biomedical, epidemiological, financial, and social studies. Specifically, let (Y^*, X^*) and (Y, X) represent the pair of response and covariates in the population and in the sample, respectively. As defined in Lawless (1997), sampling schemes that depend on the value of Y are called response selective or response biased. The response-selective sampling assumes that, for any y, the conditional distribution of X given $Y = y$ is the same as that of X^* given $Y^* = y$. For example, in case-control studies, the conditional distribution of X^* given

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

$Y^* = 1(0)$ is the population distribution of the covariates for all cases (controls), which is the same as that of the covariates of cases (controls) in the case-control sample. Throughout the paper, we denote the observations as $(Y_i, X_i), i = 1, ..., n$, which are independent and identically distributed.

The response-selective sampling schemes are likely to contain more information relevant to one's interest. In general, compared to prospective sampling which selects samples during the study period and watches for outcomes, retrospective sampling like response-selective sampling selects samples with outcomes established at the start of the study, hence it is useful in clinical studies for its effectiveness and its saving duration and costs. Thus, in a study of possible dependence of levels of hypertension (response) on those of sodium intake (covariate), sampling from patients in a hospital, which can be regarded as response-selective sampling, would be more effective than from the general public as the latter has much smaller proportion of people with hypertension. An example in economic and social studies is that wage is only observed for employed people.

The statistical analysis of biased sampling has received considerable attention in the past decades. Case-control or choice-based sampling, which is a special case of response-selective sampling, has been extensively studied; see Anderson (1972), Manski and Lerman (1977), Prentice and Pyke

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

(1979), Breslow and Day (1980), Cosslet (1981), Scott and Wild (1986, 1997), Manski (1993), etc. There are other studies on biased sampling data, involving semiparametric and parametric models; see Hausman and Wise (1981), Jewell (1985), Bickel and Ritov (1991), Wang (1996), Lawless et al. (1999), Chen (2001), Tsai (2009), Luo and Tsai (2009), Luo et al. (2009), among others. In statistical analysis of biased sampling, one of the celebrated findings is that the prospective estimating equation is still valid for case-control logistic regression; see Anderson (1977) and Prentice and Pyke (1979). However, in general, estimating equations based on prospective sampling is invalid for biased sampling and modifications using, for example, inverse probability methods is necessary. This paper shows, for a general transformation model, a rank estimation method based on prospective sampling still applies, without any modification, to response-selective sampling.

Regression analysis with response-selective sampling is generally associated with the fitted model. In particular, the estimation of the parameter of interest with biased sampling usually relies on the model assumptions, such as the inverse probability method and the pseudo-likelihood method; see Binder (1992), Lin (2000), Wang (1996), and Tsai (2009). Recently, nonparametric tests and estimation for right censored data with biased

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

sampling can be found in Ning, Qin, and Shen (2010) and Huang and Qin (2011). A novel approach to analyze length-selective data with semiparametric transformation and accelerated failure time models has been developed by Shen, Ning, and Qin (2009). We consider a class of transformation models with response-selective sampling, under which an unknown monotonic transformation of the response is linearly related to the covariates with an unspecified error distribution. The transformation models are called generalized accelerated failure time (GAFT) models in econometrics. They include many popular models, such as the proportional hazards model, the proportional odds model, and the accelerated failure time models or linear models. The response-selective sampling that we consider can be viewed as a special case of the Heckman model; see Heckman (1977, 1979). This model assumes an outcome linear regression model and a probit selection model. We consider more general transformation models and assume the "selectivity/observability" solely depends on the value of the response variable. In the case analysis of wage, we assume the chance that a potential job is taken only depends on the wage offered. The proposed estimating method does not depend on the specification of the sampling probabilities, unlike the Heckman correction. There is a rich literature on linear transformation models with a known error distribution; see, for ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

ample, Dabrowska and Doksum (1988), Cheng et al. (1995, 1997), Chen et al. (2002), and Zeng and Lin (2007). However, their methods cannot be directly applied to transformation models with an unknown error distribution. Similarly, the case-control logistic regression method in Anderson (1977) and Prentice and Pyke (1979) cannot be generalized directly.

In view of the importance of the response-selective sampling designs as well as transformation models, an easy-to-implement estimation methodology, with an advantage over the existing methods in terms of generality, is worth pursuing. Conventional methods, such as least squares (LS) or least absolute deviations (LAD) cannot be directly applied to response-selective sampling. The maximum rank correlation (MRC) estimate, originated from Han (1987) for prospective studies, is based on the rank correlation (Kendall's τ) between two variables. For illustration, consider a simple linear regression model

$$Y_i = \beta' X_i + \epsilon_i, \quad 1 \le i \le n,$$

where the (Y_i, X_i, ϵ_i) are independent and identically distributed ($i.i.d.$) copies of (Y, X, ϵ). The idea of MRC estimation is to maximize the rank correlation between Y_i and $\beta' X_i$ with respect to β. Heuristically, given that $\beta' X_i > \beta' X_j$, it is more likely that $Y_i > Y_j$ than otherwise, that the rank of Y_i and the rank of $\beta' X_i$ are positively correlated. A number of studies on

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

MRC have been conducted. Sherman (1993) proved its $\sqrt{n}$-consistency and asymptotic normality, and Khan and Tamer (2004) extended this method to semiparametric models with censoring by proposing the partial rank (PR) estimator. A smoothed partial rank (SPR) estimator was considered in Song et al. (2007) for transformation models with censoring.

We offer an estimator based on MRC for transformation models with response-selective sampling that does not rely on any further model assumption. It works equally well regardless of what the monotonic transformation is, as the MRC estimate only depends on the ranks of responses. The estimation of the transformation function, which is likely to be quite complex and computationally burdensome, is not required. The proposed method is easy to implement and computationally straightforward with the help of Nelder-Mead simplex direct search. The Nelder-Mead simplex algorithm does not require convexity or continuity; see Nelder and Mead (1965). The MRC objective function is a U-statistic. To avoid estimating the covariance matrix, we propose to use a random weighting resampling scheme for inference. In addition, since prospective sampling can be regarded as a special case of response-selective sampling, the proposed estimation is valid for prospective sampling.

We describe the model in Section 2. The proposed estimation and its

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

inference with theoretical justification are presented in Section 3. A simulation study with supportive evidence is given in Section 4. In Section 5, our method is applied to the Forbes Global 2000 data set. The paper concludes with a remark in Section 6. Proofs are deferred to the supplementary materials.

2. Model description

Let (Y^*, W^*) be a response and a $(d+1)$-dimensional vector of covariates in the population. Assume that

$$H(Y^*) = \theta_0' W^* + \epsilon^*, \tag{2.1}$$

where $H(\cdot)$ is an unknown monotonically increasing function, ϵ^* is the error, independent of W^*, with unspecified distribution, and θ_0 is a $(d+1)$-dimensional vector of regression coefficients. When $H(\cdot)$ is the identity function, (2.1) is a linear regression model. When ϵ^* follows the extreme-value distribution or the standard logistic distribution, the resulting model is the proportional hazards model or the proportional odds model, respectively. As θ_0 in (2.1) is not uniquely defined, one can take $\|\theta_0\| = 1$. For convenience, we choose to fix the first component of θ_0, $|\theta_{0,1}| = 1$. Then, $\theta_0 = (\pm 1, \beta_0')'$, where β_0 denotes the rest components.

Accordingly, W^* can be decomposed as $W^* = (Z^*, X^*)$, where Z^* is the covariate corresponding to the fixed regression coefficient and X^* is the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

other d-dimensional covariate. As we can replace Z^* by $-Z^*$ if $\theta_{0,1} = -1$, we can further assume $\theta_{0,1} = 1$ and then (2.1) can be rewritten as

$$H(Y^*) = Z^* + \beta_0' X^* + \epsilon^*. \tag{2.2}$$

We point out that the parameter estimation does not vary with different decompositions of covariates regardless of a positive constant multiplication.

Let (Y, Z, X) be the response and covariates following the distribution P of response-selective sampling. The nature of response-selective sampling implies that, for any y, the conditional distribution of (X, Z) given $Y = y$ is the same as that of (X^*, Z^*) given $Y^* = y$. An alternative but equivalent definition of response-selective sampling uses a sampling index Δ. Suppose the response Y^* is observed if and only if the covariates (X^*, Z^*) are observed and the response-covariates pair (Y^*, X^*, Z^*) is observed if and only if $\Delta = 1$. Then the response-selective sampling is defined by the conditional independence of Δ and (X^*, Z^*) given Y^*. And we can denote the observations as $(Y_i^*, X_i^*, Z_i^*, \Delta_i)$ where $\Delta_i = 1$ for $i = 1, ..., n$. With biased-sampling, Wang (1996) provided a novel pseudo-likelihood method for Cox's proportional hazards model. A pseudo-partial likelihood approach can be found in Tsai (2009). The existing methods for Cox's model with biased-sampling are conceptually appealing and have clear interpretation. The pseudo-likelihood approach relies on the specification of proportional

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

hazards model, hence it cannot apply to all transformation models. We know of no specific construction of regression analysis based on transformation models with response-selective sampling. We propose a general estimation and inference procedure based on MRC for model (2.2) with response-selective sampling. Our approach applies to all transformation models with or without knowing the error distribution, and is therefore more robust.

3. Estimation and inference

With response-selective sampling, the observations are (Y_i, Z_i, X_i), $1 \leq i \leq n$, *i.i.d.* copies of (Y, Z, X). Throughout, $I(\cdot)$ is the indicator function. Similar to Han (1987), the rank correlation for response-selective sampling is defined as

$$U_n(\beta) = \sum_{i \neq j} I(Z_i + \beta' X_i > Z_j + \beta' X_j) I(Y_i > Y_j). \tag{3.1}$$

The MRC estimate maximizes the rank correlation $U_n(\beta)$, say $\hat{\beta}_n$. Han (1987) and Sherman (1993) established the consistency and asymptotic normality of $\hat{\beta}_n$ with data from prospective sampling. However, with response-selective sampling, it is not clear whether the large sample properties still hold.

Theorem 1. *Under regularity conditions C1-C6 given in the Appendix,*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \to N\left(0, A^{-1}B(A^{-1})'\right)$$

in distribution, where the explicit forms of A *and* B *are given in the supplementary materials.*

The limiting covariance matrix of $\hat{\beta}_n$ involves the derivative of conditional expectation of the objective function, which can be difficult to estimate. To circumvent this, we propose a distributional approximation based on a random weighting method by externally generating *i.i.d.* random variables. Let $e_1, \cdots, e_n$ be a sequence of *i.i.d.* nonnegative random variables with mean 1 and variance 1. Take

$$\tilde{U}_n(\beta) = \sum_{i \neq j} e_i e_j I(Z_i + \beta' X_i > Z_j + \beta' X_j) I(Y_i > Y_j) \tag{3.2}$$

and $\tilde{\beta}_n = \arg\max_{\beta \in \mathcal{B}} \tilde{U}_n(\beta)$. The distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ can be approximated by the resampling distribution of $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$ when fixing the data (Y_i, Z_i, X_i), $1 \leq i \leq n$.

Proposition. *Given* $\{(Y_i, Z_i, X_i), 1 \leq i \leq n\}$, *under regularity conditions C1-C6 in the Appendix, as* $n \to \infty$,

$$\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n) \to N\left(0, A^{-1}B(A^{-1})'\right)$$

in distribution, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

The resampling method based on random weighting for the U-statistic objective function is well established in Jin et al (2001). We omit the proof here.

Remark 1. For computation, the numerical minimization is straightforward with the Nelder-Mead simplex algorithm that does not require convexity or continuity. In simulations, we use the "fminsearch" function in Matlab directly to search over a wide range of starting values in case there exist local maximizers. Matlab code is available upon request. In addition, another slight problem is that, with a large sample size or a large dimension of covariates, the computation is slower in simulations due to many replications. Algorithm proposed by Abrevaya (1999) improves the complexity of computation for MRC from $O(n^2)$ to $O(n \log n)$; it is available for large sample sizes. A smoothed approximation of the indicator function considered by Song et al. (2007) can be applied for large dimensions of covariates.

Remark 2. Our objective function $U_n(\beta)$ only depends on the responses through their orders which are not changed by the unknown monotonically increasing transformation $H(\cdot)$. Thus our estimate of β_0 is invariant of the transformation and estimating the unknown transformation $H(\cdot)$ can be avoided.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Remark 3. For data including cases with missing all covariates, the complete observations can be regarded as drawn from a response-selective sampling provided the missing mechanism is missing-at-random.

Remark 4. Condition C3 in the Appendix is imposed to facilitate the proof of consistency. We assume that the error distribution has a twice differentiable density function with log-concavity. Although it looks somewhat restrictive, it includes a number of widely used distributions, for example, $N(0, \sigma^2)$ and Pareto family. Thus linear models with normal errors, Cox's model, and the proportional odds model are included. With increasing technicalities, this condition can be loosened or dropped, as evidenced in our simulation results in Section 4.

4. Simulation studies

Extensive simulation studies were conducted to examine the finite sample performance of the proposed method. In the first, we consider the linear model

$$Y = Z + X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{4.1}$$

where $(\beta_1, \beta_2) = (1, -1)$, $Z \sim N(0, 1)$, and X_1 and X_2 follow a bivariate

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

normal distribution with mean $(1, -0.5)$ and variance

$$\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

We also consider the linear model

$$Y = Z + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \epsilon, \tag{4.2}$$

where $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, -1, 1, -1)$, $Z \sim N(0, 1)$, and $X = (X_1, X_2, X_3, X_4)$ follows a multivariate normal distribution with mean $(1, -0.5, 1, 0.5)$ and unit variance and equal correlation coefficients $\rho = 0.2$. Response-selective sampling were conducted with the four different schemes. In schemes 1 and 2, the samples were restricted to $Y < -2$ or $Y > 4$ and $Y < -2$ or $Y > 2.5$, respectively. Scheme 3 is simply prospective sampling. In scheme 4, the sampling probability was fixed as $\Phi(y - 2)$ for response value y.

Distributions for the error were the double exponential distribution with parameter 1, the standard normal distribution, and the standard extreme value distribution. The sample size was 200 and simulation results were based on 100 replications. The external random weights were generated from the standard exponential distribution with 500 replications.

We only present the simulation results with the inverse probability weighted method (IPW) for scheme 4 for comparison, in which the least

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

squares estimating equations are weighted by the correctly specified or misspecified sampling probabilities. Here, for different sampling schemes, the IPW method with correctly specified weights refers to the least squares estimation with the true sampling probability as weights, and the IPW method with misspecified weights refers to the least squares estimation with a unit vector as weights. The inverse probability weighted cannot be directly applied to schemes 1-2 due to null observations for some intervals in the range of the response and the lack of specification of the sampling mechanism for the estimation of sampling probability. Sampling scheme 3 was prospective sampling.

In Tables 1-2, we present the bias of the estimates of the regression parameters β_1 and β_2 (BIAS), the empirical standard error (SE), the average of the estimated standard errors (SEE), and the 95% coverage probabilities (CP) for the proposed method. We present the estimation results for scheme 4 with the proposed method and the inverse probability weighted (IPW) method in Table 3.

INSERT TABLES 1-3 HERE

From Tables 1-3, the proposed method works well in all configurations. The estimated standard errors based on random weighting are generally close to the empirical standard errors. The proposed method gives more

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

robust results than the inverse probability weighted method (IPW) for sampling scheme 4. The IPW with the true sampling probability as weight (correctly weighted) gives reasonable results for normal error and double exponential error, but provides inaccurate estimates with extreme value error. The IPW with misspecified weights gives generally biased estimates. Overall, the first simulation contains supportive evidence of the superiority of the proposed method in terms of both generality and flexibility.

The second simulation was intended to show that condition C3 may just be technical. Consider the model

$$Y = Z + \beta' X + \epsilon,$$

where $\beta = 1$, $Z \sim N(0,1)$, $X \sim N(0,1)$, ϵ follows the mixture of the standard normal distribution and a Bernoulli distribution with probability of success 0.5 and the mixture probabilities are $(0.5, 0.5)$. The error distribution is not log-concave and thus does not satisfy C3. Samples with values of the response less than -1.5 or greater than 2.5 were drawn. The bias of the estimate was 0.0302. The empirical and estimated standard deviations were 0.1591 and 0.1479, respectively. The proposed method may still work without assuming the log-concavity of the error distribution.

The third simulation used a rather extreme example to demonstrate that a biased sampling could be much more efficient than prospective sam-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

pling. Consider the model

$$Y = Z + \beta' X + \epsilon,$$

where $\beta = 1$, $\epsilon \sim N(0, 10^{-4})$, $Z \sim N(0, 1)$, and X follows a distribution with density function

$$f_X(x) = \begin{cases} 5 * 10^{-5}, & \text{if } -105 \leq x \leq -5; \\ 9.99, & \text{if } -0.05 \leq x \leq 0.05; \\ 5 * 10^{-5}, & \text{if } 5 \leq x \leq 105. \end{cases}$$

For a response-selective sampling that only takes observations with responses larger than 4.5 or smaller than -4.5, the mean and standard deviation of the estimates of β were 1.0004 and 0.0038, respectively. For prospective sampling, the mean and standard deviation of the estimates of β were 1.0108 and 0.0576, respectively. The relative efficiency for the response-selective sampling versus the prospective sampling was 230. This indicates the possibility that response-selective sampling can be designed more efficiently than prospective sampling in terms of parameter estimation.

We conducted simulations with non-trivial transformation functions to show the robustness of the proposed method. We generated data from the model:

$$H_2(Y) = Z + X_1\beta_1 + X_2\beta_2 + \epsilon,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

where $H_2(t) = \{|t|^\lambda \text{sgn}(t) - 1\}/\lambda$ with $\lambda = 3$, and (β_1, β_2), Z, X_1, X_2 the same as in (4.1). We also generated data from the same model as (4.2) except with the transformation function $H_2(\cdot)$. The samples were obtained with the sampling schemes 2 and 3 of the first simulation. The sample size was 200 and the simulation results were based on 100 replications. We summarize the results in Table 4, which contains strong evidence that the proposed method is indeed robust/insensitive to the transformation function.

INSERT TABLE 4 HERE

Overall, the results of simulation studies agree with the theory, and the consistency and asymptotic normality established in Theorem 1 might hold in more general scenarios without the technical conditions.

5. Application

We applied the proposed method to the Forbes Global 2000 data published in 2012. The data set contains the profits, assets, and market value for companies of the Forbes Global 2000. It is commonly known that profits measure the financial performance of the companies and assets indicate their size. The purpose here was to analyze the relationship among market value, profits, and assets of companies with the existing Forbes Global 2000 data. The companies on the Forbes Global 2000 list, the biggest and most

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

powerful companies in the world, are in fact a biased sampling from the general population. We deleted four records with zero values for market value, profits, and assets. The sample size was $n = 1996$. We fit the transformation model to the data with covariates $X_1 = assets/250$, $Z = profits$, and response Y=*market value* with the proposed method. For identifiability, we set the coefficient of Z to 1. The random weights were generated from the standard exponential distribution with resampling times $N = 500$. The estimate of the coefficient of X_1 was 0.2912 and the estimated standard error was 0.0503. For comparison, we also fit the same model with known logarithmic transformation function with the IPW method, in which the least squares estimating equation was weighted by the reciprocal of the market value divided by the maximum observed value of market value. We estimated the standard error by the random weighting resampling method. The IPW gave the estimate of the coefficient of X_1 as 0.4872 with the estimated standard error 0.6878.

6. Concluding remarks

This paper gives a general method of regression analysis based on the method of MRC for transformation models with response-selective sampling. For identifiability of the model, the first component of the parameter is fixed as ± 1, and then the comparison between the proposed method and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

other methods is essentially to compare the relative ratios of the coefficients when the data is simulated with more general set of coefficients, say, the first component not necessarily ± 1.

Consistency and asymptotic normality of the proposed estimator are shown. Simulation studies show that response-selective sampling gives a more efficient estimation than prospective sampling in certain situations, and the proposed estimator works well for a variety of sampling schemes and models. In addition, the nature of the MRC method implies that the estimation does not vary with different monotonic transformations, avoiding the estimation of the transformation functions. Furthermore, this method can be applied to general models of the form

$$Y^* = D \cdot F(\theta_0' W^*, \epsilon^*), \tag{6.1}$$

where Y^*, W^*, θ_0 and ϵ^* are defined as in Section 2. $D\colon \mathbb{R} \to \mathbb{R}$ is a non-degenerate, monotonic function and $F\colon \mathbb{R}^2 \to \mathbb{R}$ is strictly monotonic in each of the variables. Though we cannot separate the covariate term and the error term in this model, our estimation and inference procedure can still be applied as long as the monotonicity assumptions of the composite transformation $D \cdot F$ are valid.

Supplementary Materials

Supplementary materials include proof of Theorem 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Acknowledgment

The authors are thankful to the Editor, an associate editor and two referees for their constructive suggestions. Kani Chen's research is supported by the Hong Kong Research Grants Council (Grant No. 600509, 601011, 600612, 600813 and 16300714). Yuanyuan Lin's research is supported by the Hong Kong Research Grants Council (Grant No. 509413 and 14311916) and Direct Grants for Research at the Chinese University of Hong Kong (Grant No. 4053136 and 3132754).

References

Abrevaya, J. (1999). Computation of the maximum rank correlation estimator. *Economics Letters* **62**, 279–285.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Bickel, P. J. and Ritov, Y. (1991). Large sample theory of estimation in biased sampling regression models. *Ann. Statist.* **19**, 797–816.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–147.

Breslow, N. E. and Day, N. E. (1980). *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Chen, K. (2001). Parametric models for response-biased sampling. *J. R. Statist. Soc. B.* **63**, 775–789.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation model with censored data. *Biometrika* **89**, 659–668.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Prediction of survival probabilities with semi-parametric transformation models. *J. Am. Statist. Assoc.* **92**, 227–235.

Cosslet, S. R. (1981). Maximum likelihood estimate for choice-based samples. *Econometrica* **49**, 1289–1316.

Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in the two-sample generalized odds-rate model. *J. Am. Statist. Assoc.* **83**, 744–749.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. Econometrics* **35**, 303–316.

Hausman, J. A. and Wise, D. A. (1981). Stratification on endogenous variables and estimation: the Gary Income Maintenance Experiment. In

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Structural Analysis of Discrete Data: with Econometric Applications (eds C. Manski and D. McFadden), 364–391. Massachusetts Institute of Technology Press, Cambridge.

Heckman, J. J. (1977). Sample selection bias as a specification error with an application to the estimation of labor supply functions. NBER working paper #172.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–685.

Huang, C. Y. and Qin, J. (2010). Nonparametric estimation for length-biased and right-censored data. *Biometrika* **98**, 177–186.

Jewell, N. (1985). Regression from stratified samples of dependent variables. *Biometrika* **72**, 11–21.

Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. *J. Econometrics* **136**, 251–280.

Lawless, J. F. (1997). Likelihood and Pseudo Likelihood Estimation Based on Response-Biased Observation. *Lecture Notes-Monograph Series*, Vol. **32**, Selected Proceedings of the Symposium on Estimating Functions, 43-55.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Statist. Soc. B.* **61**, 413–438.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.

Luo, X. and Tsai, W. Y. (2009). Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. *Biometrika* **96**, 873–886.

Luo, X., Tsai, W. Y., and Xu, Q. (2009). Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika* **96**, 617–633.

Manski, C. F. (1993). The selection problem in econometrics and statistics.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Econometrika (eds G. S. Maddala, C. R, Rao and H. D. Vinod), 73–84. North-Holland, Amsterdam.

Manski, C. F. and Lerman, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* **45**, 1977–1988.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7**, 308-313.

Ning, J., Qin, J., and Shen, Y. (2010). Nonparametric tests for right-censored data with biased sampling. *J. R. Statist. Soc. B.* **72**, 609-630.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models with case-control studies. *Biometrika* **66**, 403–411.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc. B.* **48**, 170–182.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123–137.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Sherman, R. (1994). Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators. *Ann. Statist.* **22**, 439–459.

Shen, Y., Ning, J., and Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *J. Am. Statist. Assoc.* **104**, 1192–1202.

Song, X., Ma, S., Huang, J., and Zhou, X. H. (2007). A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* **8**, 197–211.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

Tsai, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**, 601–615.

Van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Wang, M. C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–354.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist. Soc. B.* **69**, 507–564.

Kani Chen, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

E-mail: makchen@ust.hk

Yuanyuan Lin, Department of Statistics, The Chinese University of Hong Kong, Hong Kong

E-mail: ylin@sta.cuhk.edu.hk

Yuan Yao, Department of Mathematics, Hong Kong Baptist University, Hong Kong

E-mail: yaoyuan@hkbu.edu.hk

Chaoxu Zhou, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

E-mail: chaoxu.zhou@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Appendix: Regularity conditions

We assume the regularity conditions hold:

C1) The unknown parameter β lies in a bounded space $\mathbf{B} \subset \mathcal{R}^d$;

C2) Both of Z^* and X^* have continuously differentiable density functions to the second order;

C3) f_{ϵ^*} is log-concave (i.e., $\log f_{\epsilon^*}$ is concave);

C4) (Identifiability condition) $\xi(\beta) := (\beta - \beta_0)'(X_2^* - X_1^*) = 0$ almost surely if and only if $\beta = \beta_0$.

Let $\Omega = (Y, Z, X)$ denote an observation from the distribution P on the set $S \subseteq \mathbb{R} \otimes \mathbb{R} \otimes \mathbb{R}^d$. For each $\omega = (y, z, x)$ in S and each β in $\mathbf{B}$, define

$$\tau(\omega, \beta) = E[I\{Y < y\}I\{\beta X + Z < \beta x + z\}] + E[I\{Y > y\}I\{\beta X + Z > \beta x + z\}].$$

Write ∇_m for the mth-partial derivative operator of a function σ with respect to $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, and

$$|\nabla_m|\sigma(\theta) = \sum_{i_1, \ldots, i_m} \left| \frac{\partial^m}{\partial\theta_{i_1} \cdots \partial\theta_{i_m}} \sigma(\theta) \right|.$$

The next two conditions guarantee a Taylor expansion of $\tau(\omega, \cdot)$ about β_0:

C5) $E|\nabla_1 \tau(\cdot, \beta_0)|^2 < \infty$;

C6) $E|\nabla_2|\tau(\cdot, \beta_0) < \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Table 1. *Simulation results for sampling schemes 1-3 with $p = 2$.*

	Scheme 1		Scheme 2		Scheme 3	
	β_1	β_2	β_1	β_2	β_1	β_2
$\epsilon \sim$ Double exponential distribution						
BIAS	0.0238	0.0030	0.0142	0.0216	0.0070	0.0206
SE	0.2213	0.2337	0.1887	0.1783	0.2447	0.2571
SEE	0.2286	0.2320	0.1816	0.1821	0.2464	0.2496
CP	0.9600	0.9600	0.9600	0.9700	0.9600	0.9600
$\epsilon \sim$ Normal distribution						
BIAS	0.0159	0.234	0.0020	0.0006	0.0262	0.0402
SE	0.1338	0.1430	0.1469	0.1442	0.4074	0.4028
SEE	0.1368	0.1356	0.1443	0.1431	0.3964	0.3907
CP	0.9600	0.9500	0.9500	0.9700	0.9600	0.9600
$\epsilon \sim$ Extreme value distribution						
BIAS	0.0390	0.0234	0.0169	0.0149	0.0451	0.0081
SE	0.1443	0.1339	0.0933	0.0963	0.2478	0.2227
SEE	0.1632	0.1638	0.0968	0.0967	0.2332	0.2381
CP	0.9700	0.9700	0.9500	0.9500	0.9500	0.9600

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Table 2. *Simulation results for sampling schemes 1-3 with $p = 4$.*

		$\epsilon \sim$ Double exponential				$\epsilon \sim$ Normal				$\epsilon \sim$ Extreme value			
		BIAS	SE	SEE	CP	BIAS	SE	SEE	CP	BIAS	SE	SEE	CP
Scheme 1	β_1	0.0014	0.1226	0.1365	0.9600	0.0098	0.1036	0.1007	0.9500	0.0138	0.0829	0.0798	0.9500
	β_2	0.0036	0.1453	0.1337	0.9600	0.0322	0.1047	0.0958	0.9500	0.0029	0.0821	0.0790	0.9600
	β_3	0.0240	0.1244	0.1311	0.9500	0.0174	0.1022	0.1000	0.9500	0.0069	0.0888	0.0810	0.9600
	β_4	0.0068	0.1313	0.1366	0.9500	0.0055	0.1035	0.0992	0.9400	0.0050	0.0852	0.0808	0.9600
Scheme 2	β_1	0.0132	0.1064	0.0996	0.9600	0.0083	0.0905	0.0822	0.9400	0.0061	0.0894	0.0745	0.9400
	β_2	0.0076	0.1050	0.1005	0.9600	0.0070	0.0864	0.0829	0.9400	0.0169	0.0693	0.0761	0.9600
	β_3	0.0002	0.1115	0.0996	0.9600	0.0169	0.0819	0.0805	0.9500	0.0147	0.0798	0.0758	0.9500
	β_4	0.0100	0.1125	0.1045	0.9600	0.0070	0.0774	0.0805	0.9500	0.0155	0.0749	0.0746	0.9500
Scheme 3	β_1	0.0087	0.0789	0.0727	0.9500	0.0079	0.0695	0.0652	0.9500	0.0090	0.0698	0.0707	0.9600
	β_2	0.0085	0.0896	0.0738	0.9500	0.0077	0.0702	0.0652	0.9600	0.0027	0.0892	0.0720	0.9400
	β_3	0.0096	0.0660	0.0707	0.9600	0.0107	0.0711	0.0659	0.9400	0.0057	0.0742	0.0706	0.9600
	β_4	0.0020	0.0816	0.0739	0.9600	0.0028	0.0611	0.0656	0.9500	0.0166	0.0791	0.0703	0.9400

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Table 3. *Simulation results for sampling scheme 4 with the proposed method and IPW.*

		Proposed				IPW correctly weighted		IPW mis-weighted	
		BIAS	SE	SEE	CP	BIAS	SE	BIAS	SE
		$\epsilon \sim$ Double exponential distribution							
$p=2$	β_1	0.0302	0.1106	0.1142	0.9500	0.1080	0.1386	0.2630	0.0692
	β_2	0.0124	0.1083	0.1154	0.9600	0.0096	0.2723	0.1165	0.0861
$p=4$	β_1	0.0119	0.0995	0.0880	0.9500	0.0354	0.1498	0.0977	0.0917
	β_2	0.0089	0.0875	0.0889	0.9600	0.0030	0.1312	0.0853	0.0781
	β_3	0.0109	0.0840	0.0882	0.9600	0.0357	0.1432	0.0951	0.0892
	β_4	0.0062	0.0881	0.0897	0.9500	0.0068	0.1708	0.1553	0.1126
		$\epsilon \sim$ Normal distribution							
$p=2$	β_1	0.0022	0.1092	0.0985	0.9400	0.0578	0.0849	0.1518	0.0487
	β_2	0.0253	0.1023	0.0958	0.9600	0.0118	0.1299	0.0782	0.0592
$p=4$	β_1	0.0126	0.0667	0.0694	0.9500	0.0227	0.1135	0.0629	0.0584
	β_2	0.0061	0.0829	0.0707	0.9400	0.0050	0.1454	0.0550	0.0623
	β_3	0.0054	0.0712	0.0715	0.9500	0.0061	0.1284	0.0440	0.0655
	β_4	0.0215	0.0676	0.0714	0.9600	0.0879	0.1388	0.0851	0.0758
		$\epsilon \sim$ Extreme value distribution							
$p=2$	β_1	0.0073	0.0729	0.0853	0.9600	0.2398	0.1577	0.3679	0.0523
	β_2	0.0098	0.0951	0.0845	0.9500	0.1814	0.1798	0.2580	0.0728
$p=4$	β_1	0.0012	0.0699	0.0721	0.9600	0.1451	0.1258	0.1898	0.0685
	β_2	0.0018	0.0751	0.0702	0.9500	0.1109	0.2250	0.1769	0.0659
	β_3	0.0211	0.0707	0.0715	0.9600	0.1440	0.1256	0.2033	0.0690
	β_4	0.0050	0.0773	0.0730	0.9500	0.1079	0.1755	0.0863	0.0886

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0338

Table 4. *Simulation results with the non-trivial transformation function* $H_2(\cdot)$.

	Sampling scheme		Proposed BIAS	SE	SEE	CP
			$\epsilon \sim$ Standard extreme value distribution			
$p=2$	Scheme 2	β_1	0.0014	0.1025	0.0948	0.9500
		β_2	0.0094	0.1016	0.0928	0.9400
$p=4$	Scheme 2	β_1	0.0072	0.0830	0.0733	0.9600
		β_2	0.0101	0.0869	0.0737	0.9400
		β_3	0.0045	0.0882	0.0748	0.9600
		β_4	0.0093	0.0861	0.0743	0.9700
$p=2$	Scheme 3	β_1	0.0073	0.0847	0.0807	0.9600
		β_2	0.0103	0.0892	0.0863	0.9400
$p=4$	Scheme 3	β_1	0.0072	0.0707	0.0723	0.9500
		β_2	0.0063	0.0732	0.0723	0.9500
		β_3	0.0153	0.0751	0.0711	0.9400
		β_4	0.0012	0.0738	0.0707	0.9500
			$\epsilon \sim$ Standard normal distribution			
$p=2$	Scheme 2	β_1	0.0106	0.1684	0.1758	0.9600
		β_2	0.0101	0.1670	0.1623	0.9600
$p=4$	Scheme 2	β_1	0.0080	0.0992	0.0995	0.9700
		β_2	0.0100	0.1003	0.1024	0.9500
		β_3	0.0080	0.1035	0.1005	0.9500
		β_4	0.0090	0.0975	0.1003	0.9600
$p=2$	Scheme 3	β_1	0.0055	0.0874	0.0817	0.9500
		β_2	0.0048	0.0849	0.0796	0.9400
$p=4$	Scheme 3	β_1	0.0080	0.0761	0.0682	0.9400
		β_2	0.0111	0.0716	0.0677	0.9500
		β_3	0.0024	0.0720	0.0685	0.9500
		β_4	0.0077	0.0757	0.0698	0.9600