

Semiparametric Regression in Size-Biased Sampling

Ying Qing Chen

Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

email: yqchen@scharp.org

SUMMARY. Size-biased sampling arises when a positive-valued outcome variable is sampled with selection probability proportional to its size. In this article, we propose a semiparametric linear regression model to analyze size-biased outcomes. In our proposed model, the regression parameters of the covariates are of major interest, while the distribution of random errors is unspecified. Under the proposed model, we discover that the regression parameters are invariant regardless of size-biased sampling. Following this invariance property, we develop a simple estimation procedure for inferences. Our proposed methods are evaluated in simulation studies and applied to two real data analyses.

KEY WORDS: Biased sampling; Linear Regression model; Log transformation; Size-biased probability of selection.

1. Introduction

Muttalak (1988) presented a study to estimate vegetation coverage in an area of Laramie, Wyoming. This area was an old limestone quarry dominated by regrowth of mountain mahogany (*Cercocarpus Montanus*). As described in the study, a line-intercept sampling method (Canfield, 1941) was used: a straight baseline was first established, and then parallel transect lines were drawn perpendicular to the baseline. Those mountain mahogany shrubs intercepted by the transect lines were measured for their widths. See Figure 1 for an illustration. As shown in Figure 1, a shrub width is defined by the maximum distance between the shrub tangents that are parallel to the transect line. Apparently, shrub widths collected this way are not random samples, but subject to size-biased sampling, i.e., their probabilities of being sampled are proportional to the widths themselves. In statistical literature, this study has been a classical example to motivate methods development for size-biased sampling (Muttalak and McDonald, 1990; Jones, 1991; Wang, 1996).

Size-biased sampling arises frequently in other studies as well, for example, when tumor size is measured in cancer-screening trials to study tumor biology and progression (Kimmel and Flehinger, 1991). Because of lead-time bias in cancer-screening trials, a tumor may be detected according to its individual size (Ghosh, 2008). More examples of size-biased sampling have been observed in industrial fiber testing (Cox, 1969), family studies of rare genetic diseases (Patil and Rao, 1978; Davidov and Zelen; 2001), etiological studies (Simon, 1980), and chronic/early disease modeling (Zelen, 2005).

In general, consider an outcome variable $X > 0$ with distribution function $F_X(x)$. When X is subject to size-biased sampling, its size-biased outcome, Y , then follows the distribution function of

$$G_Y(y) = \frac{1}{\mu_X} \int_0^y x dF_X(x),$$

where $0 < \mu_X = EX = \int_0^\infty x dF_X(x) < \infty$. Cox (1969) proposed a one-sample empirical estimator of $F_X(\cdot)$ based on observed Y 's. Vardi (1982, 1985) later showed that it is indeed the nonparametric maximum likelihood estimator (NPMLE) of $F_X(\cdot)$. A kernel density estimator by smoothing the NPMLE was developed in Jones (1991).

Usually, covariates, say Z , are also collected to study potential predictors or risk factors in association with X . In Muttalak (1988), information was collected on maximum shrub height and total number of shrub stems, both of which are important predictors to estimate the percentage of an area's vegetation coverage. In Kimmel and Flehinger (1991), because tumors of local lesions may grow to shed cancer cells into the lymphatic system and/or blood stream, secondary cancers known as lymph node or distant metastases may develop and lead to rapid disease progression or death. Therefore it is important to capture the association between tumor size and the metastasis status to understand the natural history of cancer progression. In fact, information was collected on tumor metastasis status and cancer types in Kimmel and Flehinger (1991). In either example, regression becomes an important tool to measure the association.

Parametric models can be used in regression. It is however important that the assumed parametric distributions are correctly specified, otherwise serious bias may arise in inferences (Duan, 1983). Nonparametric methods have been developed to estimate regression functions of size-biased outcomes. For example, the kernel estimator developed by Jones (1991) was extended to multivariate setting (Ahmad, 1995); Wu (2000) developed a class of local polynomial estimators; and Cristóbal and Alcalá (2000) proposed several regression function estimators by way of modified local polynomials. Because X tends to be positive, semiparametric proportional hazards models (Cox, 1972) have also been developed, for example, to model shrub width and tumor size in Wang (1996) and Ghosh (2008), respectively. In these models, regression parameters are based on hazard functions. When X is not

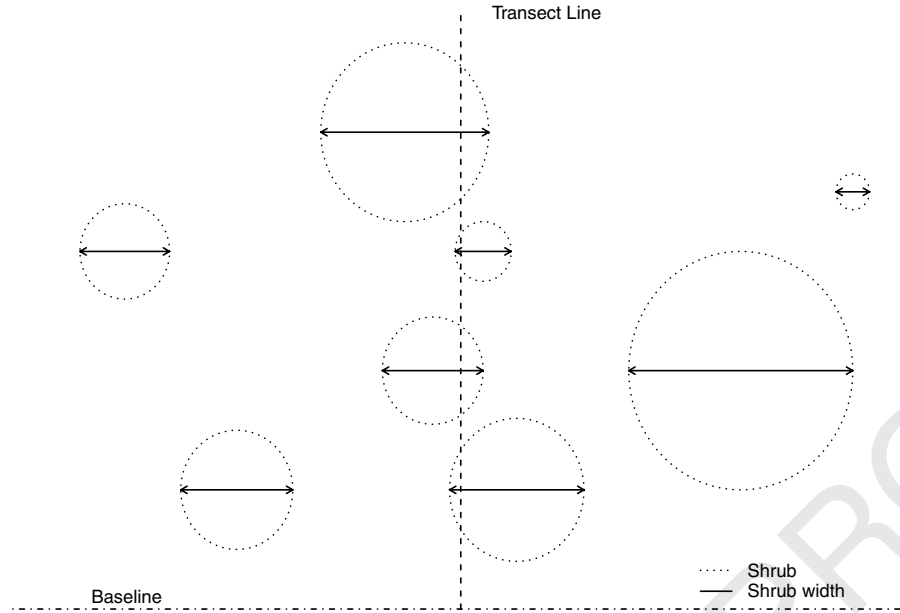


Figure 1. Schematic intercept sampling of shrub widths by one transect line.

time to event, such as shrub width or tumor size, regression models of X 's hazard functions are yet to be seen practical.

For general positive outcomes, a useful alternative to the semiparametric proportional hazards model is the log-linear regression model (Kalbfleisch and Prentice, 2002, p. 218). When X is time to event, it is the so-called accelerated failure time model (Wei, Ying, and Lin, 1990). It is also a transformation model in Tsiatis (1990). In this article, we aim to develop a semiparametric version of this model in size-biased sampling. In the rest of this article, we first introduce a semiparametric linear model and its invariance property in size-biased sampling. Then we develop a simple estimation procedure for inferences on regression parameters. We assess our method's validity and performance by Monte Carlo simulations, and further demonstrate it in actual data analysis.

2. The Method

2.1 A Linear Regression Model

Let $Z \in \mathcal{R}^p$ be the covariates vector associated with outcome variable X . Assume that the distribution of (X, Z) is determined by a joint density function of $f_{X,Z}(x, z)$. Consider the following linear regression model:

$$\log X = -\beta^T Z + \varepsilon, \quad (1)$$

where $\beta \in \mathcal{B} \subset \mathcal{R}^p$ is the regression parameter. Here, τ denotes the transpose of a vector or a matrix. In this model, $\xi \equiv \exp(\varepsilon)$ are assumed to have an unspecified density function of $f_\xi(\cdot)$. Thus an equivalent form of model (1) is

$$f_{X|Z}(x|z) = f_\xi(xe^{\beta^T z})e^{\beta^T z}, \quad (2)$$

where $f_\xi(\cdot)$ is nuisance parameter and β is the parameter of interest. Similar to the semiparametric proportional hazards model, such a semiparametric model formulation shall allow a wide range of error distributions.

Now consider a size-biased sampling indicator S , being 1 when X is selected and 0 otherwise, such that $\text{pr}\{S = 1 | X = x, Z = z\} = \text{pr}\{S = 1 | X = x\} \propto x$. Note that this selection probability depends on X only. Then the size-biased density function becomes

$$\begin{aligned} f_{X,Z|S}(x, z | S = 1) &= \frac{\text{pr}\{S = 1 | X = x, Z = z\} f_{X,Z}(x, z)}{\text{pr}\{S = 1\}} \\ &= \frac{x f_{X,Z}(x, z)}{\mu_X}, \end{aligned} \quad (3)$$

which is identical to those in Cristóbal and Alcalá (2000) and Wu (2000). Furthermore, under the assumed log-linear regression model (1), we have

$$f_{X|Z,S}(x|z, S = 1) = \frac{x f_{X|Z}(x|z)}{\mu(X|z)} = \frac{x f_\xi(xe^{\beta^T z})e^{\beta^T z}}{\mu_\xi e^{-\beta^T z}},$$

where $\mu(X|z) = \int_0^\infty x f_{X|Z}(x|z) dx$ and $\mu_\xi = E\xi$. As a result, the regression parameter β is invariant regardless of size-biased sampling (3), as summarized in the following property:

PROPERTY 1: Assume that there exists a random variable $\eta = \exp(\varepsilon)$ with the density function $f_\eta(\cdot)$, where $f_\eta(y) = y f_\xi(y)/\mu_\xi$. Under size-biased sampling, model (1) satisfies that

$$f_{X|Z,S}(x|z, S = 1) = f_\eta(xe^{\beta^T z})e^{\beta^T z}. \quad (4)$$

According to equation (4), the size-biased (X, Z) , say (Y, Z) , satisfies $\log Y = -\beta^T Z + \varepsilon$. Comparing it with equation (2), we find that (Y, Z) would follow a model almost identical to that of (X, Z) for the same β , except for the nuisance parameters. Property 1 is hence called an "invariance property" of model (1) in presence of size-biased sampling. An illustrative example is provided in the Web Supplementary Materials.

2.2 Estimation and Inferences

In random sampling, a collected dataset would usually consist of n independent and identically distributed (i.i.d.) copies of (X, Z) , (X_i, Z_i) , $i = 1, 2, \dots, n$. Then the likelihood function of β in model (1) would be proportional to

$$L_1(\beta; X, Z) = \prod_{i=1}^n f_{X,Z}(X_i, Z_i) \propto \prod_{i=1}^n f_{\xi}(X_i e^{\beta^T Z_i}) e^{\beta^T Z_i}. \quad (5)$$

Now suppose that data are instead collected subject to size-biased sampling (3). That is, the actual collected data consist of n i.i.d. copies of $\{(X, Z) | S = 1\}$, $\{(X_i, Z_i) | S_i = 1\}$, $i = 1, 2, \dots, n$. Because under equation (3),

$$\begin{aligned} f_{Z|S}(z | S = 1) &= \frac{\int_{x \in \mathcal{X}} \text{pr}\{S = 1 | X = x, Z = z\} f_X(x) dx \cdot f_Z(z)}{\int_{x \in \mathcal{X}, z \in \mathcal{Z}} \text{pr}\{S = 1 | X = x, Z = z\} f_X(x) f_Z(z) dx dz} \\ &= f_Z(z), \end{aligned}$$

where \mathcal{X} and \mathcal{Z} are the supports of X and Z , respectively, then by Property 1, the likelihood function becomes proportional to

$$\begin{aligned} L_2(\beta; X, Z | S = 1) &= \prod_{i=1}^n f_{X,Z}(X_i, Z_i | S_i = 1) \\ &= \prod_{i=1}^n f_{X|Z,S}(X_i | Z_i, S_i = 1) f_{Z|S}(Z_i | S_i = 1) \\ &\propto \prod_{i=1}^n f_{X|Z,S}(X_i | Z_i, S_i = 1) = \prod_{i=1}^n \frac{X_i f_{\xi}(X_i e^{\beta^T Z_i}) e^{\beta^T Z_i}}{\mu_{\xi} e^{-\beta^T Z_i}} \\ &= \prod_{i=1}^n f_{\eta}(X_i e^{\beta^T Z_i}) e^{\beta^T Z_i}. \end{aligned}$$

If $f_{\xi}(\cdot)$ is parametric, the usual maximum likelihood estimation (MLE) can be used to estimate β . Nevertheless, $L_1(\beta)$ and $L_2(\beta)$ are identical in β under the semiparametric model (1), regardless of size-biased sampling. For this reason, we hence aim at developing a semiparametric estimation procedures for β , without resort to any parametric form of $f_{\xi}(\cdot)$ or $f_{\eta}(\cdot)$.

To simplify our notations, we drop the notation S and denote the size-biased data by (Y_i, Z_i) , $i = 1, 2, \dots, n$. Let $\lambda(y) = -d \log \bar{F}(y) / dy$ denote a hazard function, where $\bar{F}(y) = 1 - F(y)$. By Property 1, it is true that

$$\lambda(y | Z_i) = \lambda_{\eta}(y e^{\beta^T Z_i}) e^{\beta^T Z_i}. \quad (6)$$

Let $N_i(y) = I(Y_i \leq y)$ and $\Delta_i(y) = I(Y_i \geq y)$, $i = 1, 2, \dots, n$. Thus the score function for β based on (Y_i, Z_i) is

$$\begin{aligned} l_{\beta}(\beta) &= \frac{\partial L_2(\beta)}{\partial \beta} \\ &= \sum_{i=1}^n \frac{\partial}{\partial \beta} \left\{ \log \lambda(Y_i | Z_i; \beta) - \int_0^{Y_i} \lambda(y | Z_i; \beta) dy \right\} \\ &= \sum_{i=1}^n \int_0^{\infty} \left\{ \frac{\partial \log \lambda(y | Z_i; \beta)}{\partial \beta} \right\} \\ &\quad \times \{ dN_i(y) - \Delta_i(y) \lambda(y | Z_i; \beta) dy \}. \end{aligned}$$

Apparently, because $\lambda_{\eta}(\cdot)$ is not specified and hence unknown, $l_{\beta}(\cdot)$ cannot be used directly to solve for an estimator of β as in the usual MLE method.

For model (1), however, $\lambda_{\eta}(\cdot)$ itself is an infinite-dimensional nuisance parameter. To estimate β , we adapt a quasipartial estimating equation approach in Chen and Jewell (2001). This approach would first construct a consistent nonparametric estimator for $\lambda_{\eta}(\cdot)$, and then develop appropriate estimating functions to estimate β .

Suppose that β_0 is the true value of β . Because $EN_i(y) = F(y | Z_i)$, $i = 1, 2, \dots, n$, therefore

$$\begin{aligned} E\{dN_i(y e^{-\beta_0^T Z_i})\} &= dF(y e^{-\beta_0^T Z_i} | Z_i) \\ &= f(y e^{-\beta_0^T Z_i} | Z_i) e^{-\beta_0^T Z_i} dy \\ &= \bar{F}(y e^{-\beta_0^T Z_i} | Z_i) \lambda(y e^{-\beta_0^T Z_i} | Z_i) e^{-\beta_0^T Z_i} dy \\ &= \bar{F}(y e^{-\beta_0^T Z_i} | Z_i) d\Lambda_{\eta}(y), \end{aligned} \quad (7)$$

where $\bar{F}(y | Z_i) = 1 - F(y | Z_i)$ and $\Lambda_{\eta}(y) = \int_0^y \lambda_{\eta}(u) du$. In addition, because $E\{\Delta_i(y e^{-\beta_0^T Z_i})\} = \bar{F}(y e^{-\beta_0^T Z_i} | Z_i)$, we consider the following unbiased estimating equation for $\Lambda_{\eta}(\cdot)$ when $\beta = \beta_0$,

$$\sum_{i=1}^n \{dN_i(y e^{-\beta_0^T Z_i}) - \Delta_i(y e^{-\beta_0^T Z_i}) d\Lambda_{\eta}(y)\} = 0. \quad (8)$$

By solving equation (8) we obtain a nonparametric estimator for $\Lambda_{\eta}(\cdot)$,

$$\hat{\Lambda}_{\eta}(y; \beta_0) = \int_0^y \frac{\sum_i dN_i(u e^{-\beta_0^T Z_i})}{\sum_i \Delta_i(u e^{-\beta_0^T Z_i})}. \quad (9)$$

Let $M_i(y) = N_i(y) - \int_0^y \Delta_i(u) d\Lambda(u | Z_i)$, $i = 1, 2, \dots, n$. Then $\{M_i(y e^{-\beta_0^T Z_i})\}$ are martingales with respect to the filtration defined by $\mathcal{F}_y = \sigma\{N_i(y e^{-\beta_0^T Z_i}), \Delta_i(y e^{-\beta_0^T Z_i}), Z_i\}$. As a result, $\hat{\Lambda}_{\eta}(y; \beta_0)$ is consistent, because $\hat{\Lambda}_{\eta}(y; \beta_0) - \Lambda_{\eta}(y) = \int_0^y \sum_i dM_i(u e^{-\beta_0^T Z_i}) / \sum_i \Delta_i(u e^{-\beta_0^T Z_i})$. Moreover, by the Martingale central limit theorem, $n^{1/2}\{\hat{\Lambda}_{\eta}(y; \beta_0) - \Lambda_{\eta}(y)\}$ goes to a mean zero Gaussian process as $n \rightarrow \infty$.

Similar to equation (7), we also notice that $E\{Z_i dN_i \times (y e^{-\beta_0^T Z_i})\} = Z_i \bar{F}(y e^{-\beta_0^T Z_i} | Z_i) d\Lambda_{\eta}(y)$. Thus with the consistent estimator of $\hat{\Lambda}_{\eta}(\cdot)$ obtained in equation (9), we propose to use the following estimating equations to estimate the regression parameter β ,

$$\sum_{i=1}^n \int_0^{\infty} Z_i \{dN_i(ye^{-\beta^T Z_i}) - \Delta_i(ye^{-\beta^T Z_i}) d\hat{\Lambda}_\eta(y; \beta)\} = 0, \quad (10)$$

to solve for β . By some algebraic manipulation, the above equations become equivalent to $U_n(\beta) = 0$, where $U_n(\beta) = U_n(\infty; \beta)$, and

$$U_n(y; \beta) = n^{-1/2} \sum_{i=1}^n \int_0^y \{Z_i - \bar{Z}(y; \beta)\} dN_i(ye^{-\beta^T Z_i}). \quad (11)$$

Here, $\bar{Z}(y; \beta) = \mathcal{E}^{(1)}(y; \beta) / \mathcal{E}^{(0)}(y; \beta)$ with $\mathcal{E}^{(k)}(y; \beta) = n^{-1} \times \sum_i Z_i^{\otimes k} \Delta_i(ye^{-\beta^T Z_i})$, $k = 0, 1, 2$, where \otimes is the Kronecker matrix product that defines $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$ and $v^{\otimes 2} = vv^T$ for a vector v . In general, because $U_n(\beta)$ is not a continuous function of β , a unique solution to the estimating equations in (11) may not always be plausible. We thus define a solution $\hat{\beta}_n$ as a zero-crossing of $U_n(\beta)$ such that $U_n(\hat{\beta}_n -)U_n(\hat{\beta}_n +) \leq 0$, as in Tsiatis (1990), or the minimizer of the Euclidean norm of $\|U_n(\beta)\|$, as in Wei et al. (1990).

Assume that $\lim_{n \rightarrow \infty} \mathcal{E}^{(k)}(y; \beta) = e^{(k)}(y; \beta)$. We have the following asymptotic results:

THEOREM 2. *Under the regularity conditions specified in the Appendix, $\hat{\beta}_n$ is consistent, and*

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} \mathcal{N}\{0, D^{-1}V(D^{-1})^T\},$$

where

$$D = \int_0^{\infty} \frac{\lambda'_\eta(y)}{\lambda_\eta(y)} \left\{ e^{(2)}(y) - \frac{e^{(1)}(y)^{\otimes 2}}{e^{(0)}(y)} \right\} f_\eta(y) dy, \text{ and}$$

$$V = \int_0^{\infty} \left\{ e^{(2)}(y) - \frac{e^{(1)}(y)^{\otimes 2}}{e^{(0)}(y)} \right\} f_\eta(y) dy.$$

Details of proof can be found in the Web Supplementary Materials.

To apply the asymptotic results, we need to find consistent estimators for $D^{-1}V(D^{-1})^T$. A straightforward way is to find consistent estimators for V and D , respectively. For example, $\hat{V} = n^{-1} \sum_i \int_0^{\infty} \{\mathcal{E}^{(2)}(y; \hat{\beta}_n) - \mathcal{E}^{(1)}(y; \hat{\beta}_n)^{\otimes 2} / \mathcal{E}^{(0)}(y; \hat{\beta}_n)\} dN_i(ye^{-\hat{\beta}_n^T Z_i})$. For D , it is less straightforward because of the unknown $\lambda_\eta(\cdot)$. One approach was suggested in Tsiatis (1990) by a smoothing kernel of $\lambda_\eta(\cdot)$. As pointed out by one reviewer, however, such an estimator is usually sensitive to the choice of smoothing parameter, although it may have less impact on the variance estimator of $\hat{\beta}_n$ in the proposed semiparametric model. Other alternative approaches to estimating the asymptotic variance of $\hat{\beta}_n$ may involve computer-intensive resampling to approximate the variance-covariance matrix, such as in Parzen, Wei, and Ying (1994). A less computer-intensive sample-based algorithm can be used as well. See the Web Supplementary Materials for details.

In addition to the estimation of regression parameter β , we may also be interested in estimating the distribution function of $F_\eta(\cdot)$ in model (6), and further the distribution function of $F_\xi(\cdot)$ in model (1). In order to do so, for a given value of β , consider $\hat{\eta}_i(\beta) = Y_i e^{\beta^T Z_i}$, $i = 1, 2, \dots, n$. By model (6), an estimate for the distribution function $F_\eta(\cdot)$ is straightforward,

which is simply the empirical distribution function based on $\hat{\eta}_i(\beta)$'s:

$$\hat{F}_\eta(y; \beta) = n^{-1} \sum_{i=1}^n I\{\hat{\eta}_i(\beta) \leq y\} = n^{-1} \sum_{i=1}^n N_i(ye^{-\beta^T Z_i}).$$

Moreover, according to Property 1, we know that $F_\xi(x) = \mu_\xi \int_0^x y^{-1} dF_\eta(y)$. To estimate μ_ξ , we consider the harmonic mean of $\hat{\eta}_i(\beta)$ by Cox (1969),

$$\hat{\mu}_\xi(\beta) = \frac{n}{\sum_i \hat{\eta}_i(\beta)^{-1}} = \frac{n}{\sum_i Y_i^{-1} e^{-\beta^T Z_i}}.$$

Therefore, an estimate of $F_\xi(x)$ is

$$\begin{aligned} \hat{F}_\xi(x, \beta) &= \hat{\mu}_\xi(\beta) \int_0^x y^{-1} d\hat{F}_\eta(y; \beta) \\ &= \frac{n}{\sum_i Y_i^{-1} e^{-\beta^T Z_i}} \cdot \int_0^x y^{-1} \left\{ \frac{\sum_i dN_i(ye^{-\beta^T Z_i})}{n} \right\} \\ &= \frac{n}{\sum_i Y_i^{-1} e^{-\beta^T Z_i}} \cdot n^{-1} \sum_{i=1}^n \frac{N_i(xe^{-\beta^T Z_i})}{Y_i e^{\beta^T Z_i}} \\ &= \sum_{i=1}^n \frac{N_i(xe^{-\beta^T Z_i})}{Y_i e^{\beta^T Z_i}} \bigg/ \sum_{i=1}^n \frac{1}{Y_i e^{\beta^T Z_i}}, \end{aligned}$$

which is a weighted average of rescaled $N_i(\cdot)$'s. Thus, $F_\eta(\cdot)$ and $F_\xi(\cdot)$ can be estimated by $\hat{F}_\eta(\cdot; \hat{\beta}_n)$ and $\hat{F}_\xi(\cdot; \hat{\beta}_n)$, respectively, where $\hat{\beta}_n$ are the estimates obtained earlier. In fact, when $\beta = 0$, our proposed models reduce to a one-sample problem, and $\hat{F}_\xi(\cdot; 0)$ becomes the exactly same NPMLE as studied in Vardi (1982).

Let $G_\eta(y) = n^{1/2}\{\hat{F}_\eta(y; \hat{\beta}) - F_\eta(y)\}$. Given the fact that $G_\eta(y) = n^{1/2}\{\hat{F}_\eta(y; \hat{\beta}) - \hat{F}_\eta(y; \beta_0)\} + n^{1/2}\{\hat{F}_\eta(y; \beta_0) - F_\eta(y)\}$, it is true that $G_\eta(\cdot)$ converges weakly to a mean zero Gaussian process with covariance function of $\sigma(y_1, y_2)$, similar to Theorem 8.3.3 of Fleming and Harrington (1991, p. 299). To construct a confidence interval (CI) of $F_\eta(y)$, we consider the logit transformation of $\hat{F}_\eta(y; \hat{\beta})$. That is, a 95% CI is

$$\left(\frac{\phi_l(y)}{1 + \phi_l(y)}, \frac{\phi_u(y)}{1 + \phi_u(y)} \right),$$

where

$$\phi_l(y) = \hat{F}_\eta(y) / \hat{F}_\eta(y) \exp[-1.96\hat{\sigma}(y, y) / \{\hat{F}_\eta(y)\hat{F}_\eta(y)\}]$$

and

$$\phi_u(y) = \hat{F}_\eta(y) / \hat{F}_\eta(y) \cdot \exp[1.96\hat{\sigma}(y, y) / \{\hat{F}_\eta(y)\hat{F}_\eta(y)\}].$$

Here, $\hat{\sigma}(y, y)$ is an estimated standard error of $G(y)$ and $\hat{F}_\eta(y) = 1 - \hat{F}_\eta(y)$. Similar procedure can be applied to $G_\xi(y) = n^{1/2}\{\hat{F}_\xi(y; \hat{\beta}) - F_\xi(y)\}$ to construct CIs for $F_\xi(y)$.

Our proposed estimating functions can be extended to include more general weight functions, $w_n(y; \beta)$ say, with

$\lim_n w_n(y; \beta_0) = w(y)$. Specifically, consider the weighted estimating equations for β :

$$\begin{aligned} U_n^w(\beta) &= n^{-1/2} \sum_{i=1}^n \int_0^\infty w_n(y; \beta) \{Z_i - \bar{Z}(y; \beta)\} dN_i(ye^{-\beta^T Z_i}) = 0. \end{aligned} \quad (12)$$

Let $D_w = \int_0^\infty w(y) \lambda'_\eta(y) / \lambda_\eta(y) \{e^{(2)}(y) - e^{(1)}(y)^{\otimes 2} / e^{(0)}(y)\} \times f_\eta(y) dy$ and $V_w = \int_0^\infty w(y)^2 \{e^{(2)}(y) - e^{(1)}(y)^{\otimes 2} / e^{(0)}(y)\} \times f_\eta(y) dy$. Denote $\hat{\beta}_n^w$ the solution in equation (12). Under the assumed regularity conditions, $\hat{\beta}_n^w$ is consistent and $n^{1/2}(\hat{\beta}_n^w - \beta_0) \rightarrow_D \mathcal{N}\{0, D_w^{-1} V_w (D_w^{-1})^T\}$. The sample-based method can be similarly used to estimate the variance of $n^{1/2}(\hat{\beta}_n^w - \beta_0)$. In addition, by an application of Cauchy–Schwarz inequality, the optimal weight function for the weighted estimating functions in equation (12) should be proportional to $w_{\text{opt}}(y) = \lambda'_\eta(y) / \lambda_\eta(y)$. It is indeed shown in Ritov (1990) that $\hat{\beta}_n^w$ of $w_{\text{opt}}(y)$ is the most efficient estimator among all of the semiparametric estimators under model (6). It is however yet to see a broader application of $w_{\text{opt}}(y)$ in practice given the challenge in its estimation.

Another prominent choice of weight function is $w_g(y) = n^{-1} \sum_i \Delta_i (ye^{-\beta^T Z_i})$ of the Gehan type (Tsiatis, 1990), for which $U_n^w(\beta)$ reduces to

$$\begin{aligned} U_n^g(\beta) &= n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n (Z_i - Z_j) I\{\log Y_i - \log Y_j \leq -\beta^T (Z_i - Z_j)\}. \end{aligned}$$

Therefore, solving $U_n^g(\beta) = 0$ amounts to minimizing $n^{-1} \sum_{i,j} \min\{\log Y_i - \log Y_j + \beta^T (Z_i - Z_j), 0\}$. This minimization is achieved by a linear programming of maximizing $\sum_{i,j} \delta_{ij}$ in β subject to $\delta_{ij} \leq \min\{\log Y_i - \log Y_j + \beta^T (Z_i - Z_j), 0\}$ (Koenker and Bassett, 1978). Denote $\hat{\beta}_n^g$ the minimizer. Then $\hat{\beta}_n^g$ is consistent and asymptotically normal (Jin et al., 2003), and can further assist to augment an algorithm for the general weighted estimating equations in (12). Specifically, consider the iterative algorithm proposed by Jin et al. (2003): $\hat{\beta}_{(0)} = \hat{\beta}_n^g$; in the k th step,

$$\begin{aligned} \hat{\beta}_{(k)} &= \arg \min_{\beta} \sum_{i,j} \bar{w} \{ \log Y_i + \hat{\beta}_{(k-1)}^T Z_i; \hat{\beta}_{(k-1)} \} \\ &\quad \times \min\{\log Y_i - \log Y_j + \beta^T (Z_i - Z_j), 0\}, \end{aligned}$$

where $\bar{w}(\beta) = w(y; \beta) / w_g(y; \beta)$. As k grows, $\hat{\beta}_{(k)}$ is deemed to converge when $\|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\|$ is less than a prespecified convergence criterion. In practice, this algorithm works sufficiently well by $k = 3$. To estimate the variance–covariance of $\hat{\beta}_n^g$, we can use the resampling approach developed in Rao and Zhao (1992), Parzen et al. (1994), and Jin et al. (2003). A computer software package called `rankreg` is available in R-language (<http://cran.r-project.org/doc/packages/rankreg.pdf>). Instruction and sample programs can be found in the package manual. This package generally works well with moderate sample sizes. However, it may be time consuming and take up to 3–4 hours on a personal computer of 1G RAM

to estimate a model of four covariates, when sample size becomes 1000.

One use of the weighted estimating equations is in model adequacy assessment. As proposed in Gill and Schumacher (1987) for the proportional hazards model, the rationale is that the estimates based on different weighted estimating equations should be reasonably close to each other if the assumptions of the proportional hazards model do hold, and should differ otherwise. This same rationale can be applied to the adequacy assessment of model (1). Specifically, consider the estimating equations in (12) for two different weight functions, $w_{n,1}(y)$ and $w_{n,2}(y)$ say. Their associated estimates of the regression parameter are $\hat{\beta}_{n,1}^w$ and $\hat{\beta}_{n,2}^w$, respectively. Should model (1) be true, then a Wald-type of statistic based on $\hat{\beta}_{n,1}^w - \hat{\beta}_{n,2}^w$ is seemingly straightforward to use, given the asymptotic normality of $n^{1/2}(\hat{\beta}_{n,1}^w - \hat{\beta}_{n,2}^w)$ similar to that in Lin (1991).

Due to the difficulty in estimating $\lambda_\eta(\cdot)$ and $\lambda_\xi(\cdot)$ in the asymptotic variance of $n^{1/2}(\hat{\beta}_{n,1}^w - \hat{\beta}_{n,2}^w)$, however, we would instead use the method proposed in Wei et al. (1990) to derive a goodness-of-fit test for model (1). That is, we consider $(U_{n,1}^w(\beta_0)^T, U_{n,2}^w(\beta_0)^T)^T$, which is asymptotically joint normal with the variance–covariance matrix that is the limit of

$$\begin{bmatrix} V_{11}(\beta_0) & V_{12}(\beta_0) \\ V_{21}(\beta_0) & V_{22}(\beta_0) \end{bmatrix},$$

where

$$V_{11}(\beta) = n^{-1} \sum_i \int_0^\infty w_{n,1}^2(y; \beta) \{Z_i - \bar{Z}(y; \beta)\}^{\otimes 2} dN_i(ye^{-\beta^T Z_i}),$$

$$\begin{aligned} V_{12}(\beta) &= V_{21}(\beta) = n^{-1} \sum_i \int_0^\infty w_{n,1}(y; \beta) w_{n,2}(y; \beta) \\ &\quad \times \{Z_i - \bar{Z}(y; \beta)\}^{\otimes 2} dN_i(ye^{-\beta^T Z_i}), \end{aligned}$$

and

$$V_{22}(\beta) = n^{-1} \sum_i \int_0^\infty w_{n,2}^2(y; \beta) \{Z_i - \bar{Z}(y; \beta)\}^{\otimes 2} dN_i(ye^{-\beta^T Z_i}).$$

Thus, the following statistic can be used in the goodness-of-fit assessment of model adequacy:

$$\begin{aligned} T &= \min_{\beta \in U(\hat{\beta}_{n,1}^w)} \left\{ \begin{bmatrix} U_{n,1}^w(\beta) \\ U_{n,2}^w(\beta + \hat{\beta}_{n,1}^w - \hat{\beta}_{n,2}^w) \end{bmatrix}^T \right. \\ &\quad \times \begin{bmatrix} V_{11}(\beta_0) & V_{12}(\beta_0) \\ V_{21}(\beta_0) & V_{22}(\beta_0) \end{bmatrix}^{-1} \\ &\quad \left. \times \begin{bmatrix} U_{n,1}^w(\beta) \\ U_{n,2}^w(\beta + \hat{\beta}_{n,1}^w - \hat{\beta}_{n,2}^w) \end{bmatrix} \right\}, \quad (13) \end{aligned}$$

which is asymptotically χ_p^2 -distributed as shown in Wei et al. (1990).

3. Numerical Studies

3.1 Simulations

Simulation studies are conducted to assess both semiparametric methods and the MLE methods for the proposed linear

regression model. In our simulations, we consider the linear regression model (1) of $\log X = -\beta^T Z + \varepsilon$, where two distributions are chosen for the random variable ε , i.e.,

- (i) a standard Normal distribution with the density function of $(2\pi)^{-1/2} \exp(-s^2/2)$, and
- (ii) an extreme-value distribution with the density function of $2 \exp(2s - e^{2s})$.

Hence for $\xi = \exp(\varepsilon)$, the density function $f_\xi(\cdot)$'s are log normal $(2\pi)^{-1/2} s^{-1} \exp\{-\log s)^2/2\}$ and Weibull $2s \exp(-s^2)$, respectively. According to Property 1, for $\eta = \exp(\varepsilon)$ in $\log Y = -\beta^T X + \varepsilon$, then the respective density function $f_\eta(\cdot)$'s are $(2\pi)^{-1/2} \exp\{-\log s)^2/2\}/\mu_L$ and $2s^2 \exp(-s^2)/\mu_W$. Numerical integration shows that $\mu_L = \int_0^\infty (2\pi)^{-1/2} \exp\{-\log y)^2/2\} dy \approx 1.6487$ and $\mu_W = \int_0^\infty 2y^2 \exp(-y^2) dy \approx 0.8862$.

In each simulation, we generate a sample of n i.i.d. copies of size-biased (Y, Z) according to the model of $\log X = -\beta_0^T Z + \varepsilon$. Here, β_0 is the true value of β , and Z are simulated according to a uniform distribution $U[0, 1]$ and hence continuous. We use four methods to estimate β in model (1):

- (i) Para-L: MLE method with underlying log-normal distribution,
- (ii) Para-W: MLE method with underlying Weibull distribution,
- (iii) Semi-L: semiparametric method with $w(y) \equiv 1$, and

- (iv) Semi-G: semiparametric method with the Gehan weight function $w_g(\cdot)$.

In particular, when the underlying distribution is log normal, Para-W represents an incorrect MLE method using the Weibull distribution. Similarly, when the underlying distribution is Weibull, Para-L represents an incorrect MLE method using the log-normal distribution.

Simulation results are tabulated in Table 1. In this table, n is chosen to be 50, 200, and 500, representing small, moderate, and large sample sizes, respectively. The true value β_0 is chosen to be 0 and 1, representing the null and a specific alternative hypotheses, respectively. Each cell in the table is calculated with 10,000 simulated samples. A bias is calculated as the average of 10,000 $(\hat{\beta} - \beta_0)$'s. A coverage probability is calculated as the percentage of 10,000 95% nominal CIs containing β_0 . The sample standard error (SSE) of 10,000 $\hat{\beta}$'s and the mean of 10,000 estimated standard errors (MSE) are also calculated.

As shown in Table 1, the bias of $\hat{\beta}$ for each method in this simulation setup is mostly close to zero, even though the MLE methods may specify incorrect underlying distributions for $f_\xi(\cdot)$. The coverage probabilities for a correct Para-L or Para-W, i.e., an MLE method with correctly assumed underlying distributions, and Semi-L or Semi-W are generally close to the nominal level of 95%, while an incorrect Para-L or Para-W deviates from 95% notably. Similarly, the SSE and MSE are generally close to each other when the correct MLE methods

Table 1

Summary of simulation results on parametric and semiparametric estimation of β in model (1) of $\log X = -\beta^T Z + \varepsilon$

β	n	Estimation ^b	$f_\xi(\cdot)$ ^a : Log normal				$f_\xi(\cdot)$ ^a : Weibull			
			Bias	95% CP	SE	Mean SE	Bias	95% CP	SE	Mean SE
0	50	Para-L	-0.0003	0.959	0.5031	0.5037	0.0092	0.831	0.3387	0.2437
		Para-W	0.0123	0.853	0.6017	0.4911	0.0002	0.954	0.2297	0.2294
		Semi-L	0.0006	0.951	0.7657	0.7651	0.0001	0.950	0.7899	0.7901
		Semi-G	0.0002	0.955	0.6533	0.6534	-0.0003	0.951	0.5084	0.5085
	200	Log normal	0.0002	0.945	0.2540	0.2539	0.0096	0.837	0.1779	0.1271
		Weibull	0.0125	0.859	0.3152	0.2436	0.0002	0.955	0.1096	0.1095
		Semi-L	0.0003	0.947	0.3897	0.3899	-0.0006	0.948	0.3973	0.3974
		Semi-G	0.0004	0.952	0.3261	0.3261	-0.0002	0.950	0.2603	0.2601
	500	Log normal	-0.0001	0.949	0.1565	0.1564	0.0098	0.832	0.1195	0.0814
		Weibull	0.0121	0.857	0.2049	0.1537	-0.0002	0.951	0.0716	0.0718
		Semi-L	0.0002	0.950	0.2504	0.2503	-0.0002	0.950	0.2472	0.2470
		Semi-G	-0.0001	0.950	0.2477	0.2479	-0.0003	0.949	0.1619	0.1619
1	50	Log normal	-0.0045	0.942	0.5023	0.5021	0.0091	0.831	0.3391	0.2471
		Weibull	0.0122	0.851	0.6297	0.4731	-0.0001	0.947	0.2348	0.2347
		Semi-L	0.0005	0.951	0.5451	0.5453	-0.0007	0.952	0.5411	0.5409
		Semi-G	-0.0002	0.954	0.5227	0.5225	0.0001	0.951	0.4603	0.4605
	200	Log normal	0.0002	0.951	0.2511	0.2510	0.0093	0.830	0.1903	0.1275
		Weibull	0.0120	0.856	0.3111	0.2419	-0.0005	0.951	0.1198	0.1198
		Semi-L	0.0009	0.950	0.2751	0.2750	-0.0001	0.951	0.2755	0.2754
		Semi-G	-0.0003	0.951	0.2603	0.2602	-0.0001	0.949	0.2443	0.2445
	500	Log normal	-0.0000	0.951	0.1628	0.1627	0.0096	0.833	0.1168	0.0809
		Weibull	0.0124	0.852	0.2081	0.1541	0.0003	0.950	0.0716	0.0716
		Semi-L	-0.0001	0.949	0.1689	0.1690	0.0001	0.949	0.1697	0.1698
		Semi-G	0.0002	0.951	0.1654	0.1653	-0.0000	0.951	0.1447	0.1448

^a $f_\xi(\cdot)$: true density function of $\xi = \exp(\varepsilon)$.

^bEstimation: Para-L, parametric estimation using log-normal distribution; Para-W, parametric estimation using Weibull distribution; Semi-L, semiparametric estimation without weight function; and Semi-G, semiparametric estimation using Gehan weight function.

and the semiparametric methods are used, but otherwise for the incorrect MLE methods.

As far as the semiparametric methods and the correct MLE methods are compared, the correct MLE methods generally yield smaller SSE and MSE, which means the semiparametric methods are yet to be fully efficient. As shown in the table, however, the efficiency of semiparametric methods may be potentially improved by choosing different weight functions. For example in the current simulation setup, the Gehan weight function may yield smaller variances, compared with those of $w_g(\cdot) \equiv 1$. As a summary, the correct MLE methods generally outperform the semiparametric methods. The validities of MLE methods however depend on whether or not the assumptions are correct. The semiparametric methods are generally more reliable with respect to different underlying distributions. Their efficiencies may be improved by choosing appropriate weight functions.

3.2 Real Data Examples

In this section, we use our proposed methods to analyze the real data examples that are introduced in §1: one is the classi-

cal dataset of shrub widths in Muttlak (1988), and the other is the tumor size dataset in Kimmel and Flehinger (1991). To save space, we present major analysis results in this article.

Example 1. The original dataset of shrub widths contains data on 89 shrub samples. In addition to the measurement of shrub widths (Width), two more attributes of mountain mahogany, maximum height (Height), and number of stems (Stem), were measured. Both attributes are important predictors of shrub width in estimating vegetation coverage of mountain mahogany. To ensure uniform coverage over the study area, two independent replications (Replicate), each with three systematically placed parallel transects (Transect), were established (Muttlak, 1988).

As shown in the scatter plots of Figure 2, when either attribute of Height or Stem increases, Width tends to increase. There appears, however, a quadratic relationship between Width and Stem. It is plausible that there may exist systematic deviation in Replicate and Transect. As shown in the boxplots of Figure 2, there appears some distributional

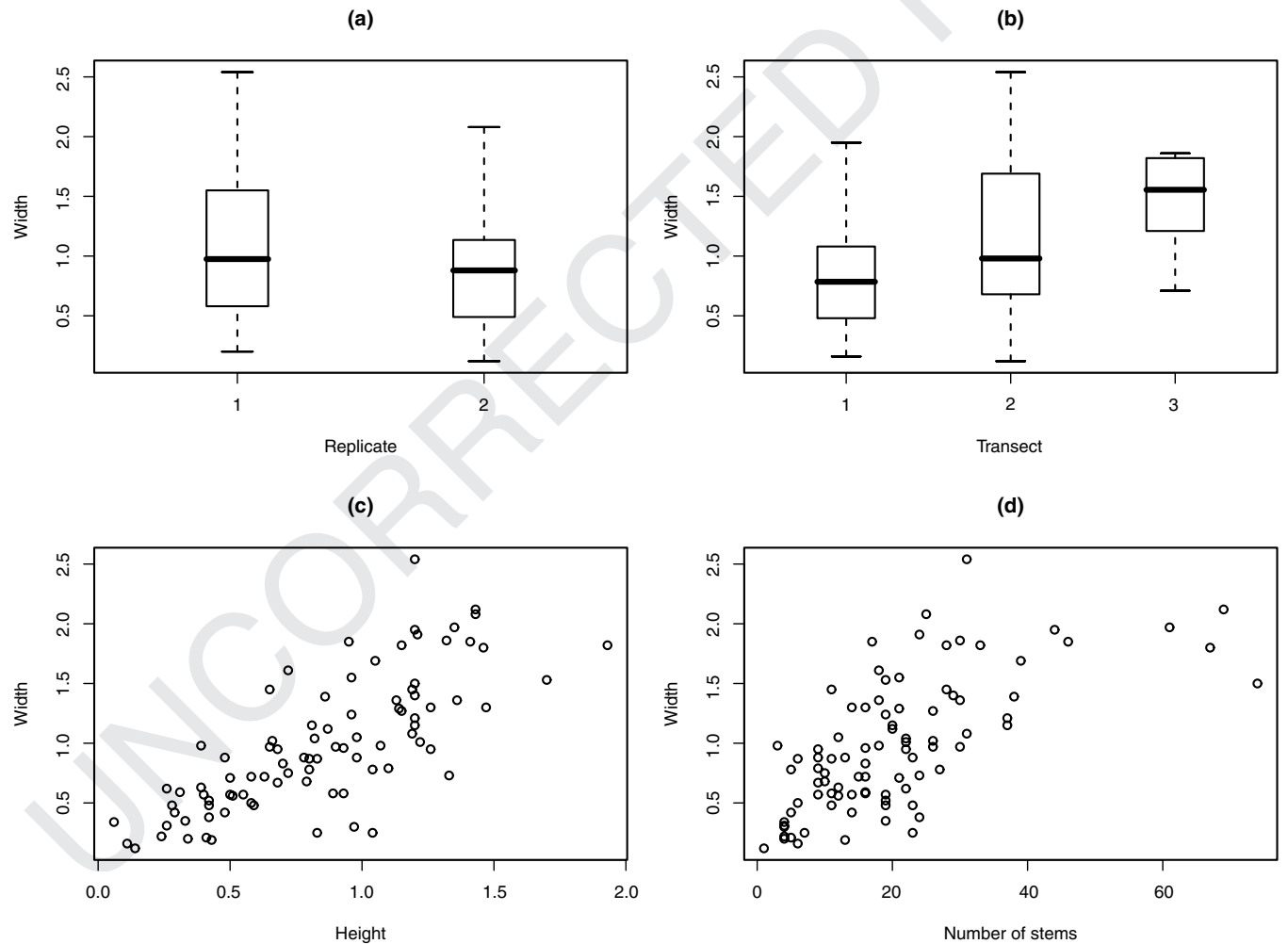


Figure 2. Exploratory plots of shrub data: (a) boxplot of Width by Replicate; (b) boxplot of Width by Transect; (c) scatter plot of Width versus Height; and (d) scatter plot of Width versus Stem.

Table 2
Summary statistics for shrub widths data

Variable	Mean ^a	Min.	25%-tile	Median	75%-tile	Max.
Width	0.98	0.12	0.57	0.88	1.36	2.54
Replicate 1	51.69%					
Replicate 2	48.31%					
Transect I	56.18%					
Transect II	37.08%					
Transect III	6.73%					
Height	0.84	0.06	0.50	0.83	1.19	1.93
Stem	20.42	1.00	11.00	19.00	26.00	74.00

^aMean, sample mean for continuous variables and frequency percentage for categorical variables.

difference in Width between different Replicates, although this difference is not seemingly prominent. There also appear distributional differences in Width between Transects I and II, and between Transects I and III. Some univariate summary statistics of the outcome variable and four covariates are in Table 2.

To examine the association between Width and the two attributes adjusting for Replicate and Transect, we consider a linear regression model $\log X = -\beta^T Z + \varepsilon$, where X is the outcome variable Width, and Z is the covariate vector including Height, Stem, Replicate, and Transect, in regression analysis. Estimates of regression parameters are tabulated in Table 3. As shown in the table, all four methods, i.e., Para-L, Para-W, Semi-L, and Semi-W, yield that Height and Stem are significant predictors, adjusting for Replicate and Transect, by their 95% CIs. This implies that taller shrubs with more stems are associated with greater width spread and hence more vegetation coverage, and per unit increase in Height and Stem would lead to about 118% and 2.5% increase in Width, respectively.

For the seemingly quadratic association with Stem, we include an additional squared Stem in the model, and find that Width is not significantly associated with the added quadratic term. It is however still significantly associated with Height and Stem, with or without Replicate and Transect. Applying

the goodness-of-fit test in equation (13) to the model with all covariates, we obtain $T = 9.13 \sim \chi^2$ ($p = 0.10$, $df = 5$), which does not reject the model's adequacy.

Example 2. In Kimmel and Flehinger (1991), a lung cancer dataset was used to examine the relationship between the occurrence of metastases and the size of primary cancers. In the dataset, lung cancers were diagnosed in a population of male smokers over 45 years old who enrolled voluntarily in a randomly controlled trial to evaluate the use of sputum cytology. Two types of lung cancer were detected: adenocarcinomas by radiologic screening or patient's symptoms, and epidermoid cancers by sputum cytology, chest X-ray or patient's symptoms. The diagnosis of metastases was based on the then best available staging, clinical, surgical, or pathological readings. Among the total 228 patients, there were 141 adenocarcinomas and 87 epidermoid cancers. Tumor sizes were determined by the geometric means of recorded dimensions: resected cancers were measured directly, and nonresected cancers were mainly measured on radiography. More descriptive details can be found in Ghosh (2008).

As pointed out in Ghosh (2008), a complication in analyzing tumor size in this dataset is the presence of size-biased sampling, i.e., tumors detected in the screening program tend to depend on their growth and hence the sizes themselves. In Ghosh (2008), tumor sizes were treated as time-to-event variables. Their hazard functions were modeled by the proportional hazards model. Therefore, the regression parameters would be interpreted in relative risk of tumor sizes. Additional summary statistics and exploratory analysis results of this dataset can be found in Ghosh (2008).

To examine the association between tumor size and the occurrence of metastases in size-biased sampling, we consider the proposed linear regression model. We first fit a linear regression model with the presence/absence of metastases only. The regression parameter estimate of $\hat{\beta}$ equals 0.45 with a 95% CI (0.12, 0.78). This means that tumor size is about 56% significantly greater in presence of metastases. However, when including the cancer types of adenocarcinomas and epidermoid in the model, $\hat{\beta}$ becomes 0.24 with a 95% CI (-0.07, 0.57), which is not statistically significant. On the other hand,

Table 3
Parametric and semiparametric estimation of β in model (1) of $\log X = -\beta^T Z + \varepsilon$ for shrub widths data

Covariate	Est.	SE	95% CI	Est.	SE	95% CI
			Para-L			Para-W
Replicate	0.0701	0.0949	(-0.1159, 0.2562)	-0.0100	0.0814	(-0.1697, 0.1496)
Transect I						
Transect II	-0.0341	0.0987	(-0.2276, 0.1593)	-0.0531	0.0849	(-0.2195, 0.1132)
Transect III	-0.0238	0.1930	(-0.4021, 0.3545)	0.2056	0.1533	(-0.0951, 0.5062)
Height	-0.9993	0.1280	(-1.2501, 0.7484)	-0.8895	0.1233	(-1.1313, -0.6478)
Stem	-0.0131	0.0038	(-0.0206, -0.0056)	-0.0115	0.0038	(-0.0190, -0.0040)
			Semi-L			Semi-G
Replicate	-0.0240	0.0804	(-0.1817, 0.1337)	-0.0600	0.0814	(-0.2175, 0.0976)
Transect I						
Transect II	-0.0758	0.0953	(-0.2625, 0.1110)	-0.0652	0.0849	(-0.2520, 0.1216)
Transect III	-0.1719	0.2054	(-0.5744, 0.2306)	-0.1608	0.1533	(-0.5634, 0.2418)
Height	-0.8132	0.1336	(-1.0751, -0.5514)	-0.7485	0.0904	(-1.0103, -0.4866)
Stem	-0.0142	0.0042	(-0.0226, -0.0059)	-0.0146	0.0028	(-0.0228, -0.0064)

the regression parameter estimate for cancer types is 0.65 with a 95% CI (0.45, 0.86). This implies that tumor size tends to be 91% greater in epidermoid cancers than that in adenocarcinomas. Our finding is consistent with the relative risk calculation in Ghosh (2008), although larger sample size may be needed to confirm a significant positive association in relative risk. Applying the goodness-of-fit test in equation (13) to the model with both covariates, we obtain $T = 4.97 \sim \chi^2$ ($p = 0.08$, $df = 2$), which does not reject the model's adequacy, either.

4. Discussion

Our proposed regression model is essentially a semiparametric linear model. In the statistical literature, semiparametric linear models have been extensively studied. For example, the accelerated failure time model is a classical example of semiparametric linear model in time-to-event analysis (Kalbfleisch and Prentice, 2002, p. 218). Additional examples include semiparametric linear models in curve regression analysis (Hardle and Marron, 1990), generalized linear regression (Chen, 1995) and partial linear regression (Bhattacharya and Zhao, 1997).

One prominent assumption in our proposed model is the choice of log transformation. Because outcomes in size-biased sampling are mostly positive, a log-transformation would relax the restriction on β to allow a wide range of covariates Z . In addition, for log-transformed X , our proposed model essentially assumes a multiplicative association between X and Z . It is the multiplicative association that leads to the critical invariance property, which greatly simplifies our estimation. Other transformations, such as identity, are yet to be seen to have these advantages.

However, our proposed model can be generalized to, for example, $\log X = h(X, \beta) + \varepsilon$ for some function $h(X, \beta)$. Under this generalized model, the invariance property still holds in size-biased sampling. This shall lead to further work. Specifically, an alternative class of regression models can assume that

$$\log X = h(\beta^T X) + \varepsilon,$$

where $h(\cdot)$ is unknown, but ε follows a known distribution, such as a standard Normal distribution, to avoid potential identifiability issue. This model shall further relax model assumptions and assist with model adequacy assessment.

As pointed by an associate editor, the use of hazard functions for outcomes other than censored time to event seems unnatural. This is in fact why semiparametric linear regression model may be more appealing to model outcomes, such as shrub width or tumor size, than the usual proportional hazards model. Nevertheless, we extensively use (cumulative) hazards functions in this article to develop our estimation procedure, because of the simple representation of baseline cumulative hazard function in equation (9). This simple representation ultimately facilitates a straightforward estimation of β . The advantage of hazard functions is also seen in developing the asymptotic theory of Theorem 2. Otherwise the expression of D would be more complex.

When X is time to event, regression models based on hazard functions are appealing. When time to event is size biased, it is usually called length biased. Researchers have been studying censored length-biased time to event in the statistical literature. For example, Vardi (1989) estimated the life-

time distribution under multiplicative censorship; Asgharian, M'LAN, and Wolfson (2002) developed an unconditional approach to studying length-biased lifetimes; and a recent work by Cristóbal, Alcalá, and Ojeda (2007) proposed some nonparametric estimation from backward recurrence times. More work is needed to extend the proposed linear regression model to censored time to event.

5. Supplementary Materials

Details of those referenced in Section 2 are available in the Web Supplementary Materials under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

ACKNOWLEDGEMENTS

The author thanks an associate editor and two reviewers for their constructive comments, and also thanks Professor Debashis Ghosh for his references to an earlier work and the tumor size dataset. This research is supported in part by the National Institutes of Health grants.

REFERENCES

- Ahmad, I. A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statistics & Probability Letters* **22**, 121–129.
- Asgharian, M., M'LAN, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association* **97**, 201–209.
- Bhattacharya, P. K. and Zhao, P. L. (1997). Semiparametric inference in a partial linear model. *Annals of Statistics* **25**, 244–262.
- Canfield, R. H. (1941). Application of the line intercept method in sampling range vegetation. *Journal of Forestry* **39**, 388–394.
- Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Annals of Statistics* **23**, 1102–1129.
- Chen, Y. Q. and Jewell, N. P. (2001). On a general class of hazards regression models. *Biometrika* **88**, 687–702.
- Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, N. L. Johnson and H. Smith, Jr. (eds.), 506–527. New York: Wiley-Interscience.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society, Series B* **34**, 187–220.
- Cristóbal, J. A. and Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. *Journal of Statistical Planning and Inferences* **89**, 145–168.
- Cristóbal, J. A., Alcalá, J. T., and Ojeda, J. L. (2007). Nonparametric estimation of a regression function from backward recurrence times in a cross-sectional sampling. *Lifetime Data Analysis* **13**, 273–293.
- Davidov, O. and Zelen, M. (2001). Referent sampling, family history and relative risk: The role of length-biased sampling. *Biostatistics* **2**, 173–181.
- Duan, N. (1983). Smearing estimate: A nonparametric retransformation methods. *Journal of the American Statistical Association* **78**, 605–610.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Ghosh, D. (2008). Proportional hazards regression for cancer studies. *Biometrics* **64**, 141–148.
- Gill, R. D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.
- Hardle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics* **18**, 63–89.
- Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.

- Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–353.
- Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511–519.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. New York: Wiley.
- Kimmel, M. and Flehinger, B. J. (1991). Nonparametric estimation of size-metastasis relationship in solid cancers. *Biometrics* **47**, 987–1004.
- Koenker, R. and Bassett, G. S. (1978). Regression quantiles. *Econometrica* **46**, 33–50.
- Lin, D. Y. (1991). Goodness of fit for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association* **86**, 725–728.
- Muttlak, H. A. (1988). Some aspects of ranked set sampling with size biased probability of selection. Ph.D. Dissertation. University of Wyoming, Laramie, Wyoming.
- Muttlak, H. A. and McDonald, L. L. (1990). Ranked set sampling with size-based probability of selection. *Biometrics* **46**, 435–446.
- Parzen, M. I., Wei, L. J., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350.
- Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-based sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179–189.
- Rao, C. R. and Zhao, L. C. (1992). Approximation to the distribution of M -estimates in linear models by randomly weighted bootstrap. *SankhyāA* **54**, 323–331.
- Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics* **18**, 303–328.
- Simon, R. (1980). Length biased sampling in etiologic studies. *American Journal of Epidemiology* **111**, 444–452.
- Tsiatis, A. A. (1990). Estimating regression parameter using linear rank tests for censored data. *Annals of Statistics* **18**, 354–372.
- Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics* **10**, 616–620.
- Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics* **13**, 178–203.
- Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76**, 751–761.
- Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–354.
- Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–851.
- Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statistica Sinica* **10**, 789–817.
- Zelen, M. (2005). Forward and backward recurrence times and length biased sampling: Age specific models. *Lifetime Data Analysis* **10**, 325–334.

Received xxxx xxxx. Revised xxxx xxxx.
Accepted xxxx xxxx.

Q3

Queries

- Q1** Author: The meaning of the sentence in the 'Supplementary Materials' section is not clear; please rewrite or confirm that the sentence is correct.
- Q2** Author: Reference Jin et al. (2001) has not been cited in the text. Please indicate where it should be cited; or delete from the list.
- Q3** Wiley-Blackwell: Please provide history date of this article.