CORE
Metadata, citation and similar papers at core.ac.uk
Provided by New York University Faculty Digital Archive

Tobit Model Estimation and Sliced Inverse Regression

Lexin Li
Department of Statistics
North Carolina State University
E-mail: li@stat.ncsu.edu

Jeffrey S. Simonoff
Leonard N. Stern School of Business
New York University
E-mail: jsimonof@stern.nyu.edu

Chih-Ling Tsai
Graduate School of Management
University of California at Davis
E-mail: cltsai@ucdavis.edu

Abstract

It is not unusual for the response variable in a regression model to be subject to censoring or truncation. Tobit regression models are a specific example of such a situation, where for some observations the observed response is not the actual response, but rather the censoring value (often zero), and an indicator that censoring (from below) has occurred. It is well-known that the maximum likelihood estimator for such a linear model (assuming Gaussian errors) is not consistent if the error term is not homoscedastic and normally distributed. In this paper we consider estimation in the Tobit regression context when those conditions do not hold, as well as when the true response is an unspecified nonlinear function of linear terms, using sliced inverse regression (SIR). The properties of SIR estimation for Tobit models are explored both theoretically and based on Monte Carlo simulations. It is shown that the SIR estimator has good properties when the usual linear model assumptions hold, and can be much more effective than other estimators when they do not. An example related to household charitable donations demonstrates the usefulness of the estimator.

Key words: Dimension reduction; Heteroscedasticity; Nonnormality; Single-index model.

1 Introduction

A common occurrence in many regression models is the existence of truncation or censoring in the response variable. Tobin (1958) pioneered the study of such models in economics, analyzing household expenditures on durable goods while taking into account the fact that expenditures cannot be negative. That is, for some observations the observed response is not the actual response, but rather the censoring value (often zero), and an indicator that censoring (from below) has occurred. More specifically, the so-called Type I Tobit model (Amemiya, 1984) is defined as follows. Given a univariate response y^* and a p-dimensional predictor vector X, the model is defined as

$$\begin{aligned} y^* &= \eta_1^\intercal X + \varepsilon, \quad \varepsilon \sim \text{Normal}(0, \sigma^2), \\ y &= \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}, \end{aligned} \tag{1}$$

where $\eta_1 \in \mathbb{R}^p$ is the unknown parameter. While the true response is y^*, only the left censored version y of y^* is observable. Additionally, a censoring indicator δ is defined, with $\delta = 1$ if $y^* > 0$ and $\delta = 0$ otherwise. Generally, the left censoring does not have to be fixed at 0, or at any constant value. In that case, we define an appropriate censoring variable C, such that $y = \max(y^*, C)$.

The parameters η_1 and σ^2 can be estimated consistently using maximum likelihood, but it is well-known that the MLE of η_1 is not consistent if the error term ε is not homoscedastic and normally distributed. Several estimators have been proposed that are consistent under more general situations, and will be described in more detail in Section 3.

Generalizations of this model are also possible. Following Amemiya (1984), a Type II Tobit model is defined as

$$\begin{aligned} y_1^* &= \eta_1^\intercal X + \varepsilon_1, \quad \varepsilon_1 \sim \text{Normal}(0, \sigma_1^2), \\ y_2^* &= \eta_2^\intercal X + \varepsilon_2, \quad \varepsilon_2 \sim \text{Normal}(0, \sigma_2^2), \\ y_2 &= \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0 \end{cases}. \end{aligned}$$

A Type III Tobit model is

$$\begin{aligned} y_1^* &= \eta_1^\intercal X + \varepsilon_1, \quad \varepsilon_1 \sim \text{Normal}(0, \sigma_1^2), \\ y_2^* &= \eta_2^\intercal X + \varepsilon_2, \quad \varepsilon_2 \sim \text{Normal}(0, \sigma_2^2), \\ y_1 &= \begin{cases} y_1^* & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0 \end{cases} \\ y_2 &= \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0 \end{cases}. \end{aligned}$$

For the Type II model, what is observed are y_2 and the indicator variable δ; $\delta = 1$ if $y_1^* > 0$ and $\delta = 0$ otherwise, which reflects the sign of y_1^*. For the Type III model, y_1, y_2, and δ are observed.

In this article, we generalize the Type I Tobit model to be of the form

$$\begin{aligned} y^* &= f(\eta_1^\intercal X, \ldots, \eta_d^\intercal X, \varepsilon), \\ y &= \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}, \end{aligned} \tag{2}$$

where f is an unknown differentiable function, $d \leq p$, the error ε is stochastically independent of X but unspecified. In this model, the mean $E(y^*|X)$ may be nonlinear, and the variance $Var(y^*|X)$ may be heteroscedastic, and both may depend on X through linear terms. We also briefly consider the generalization of the Type II and Type III models in a similar fashion, where the linear terms are replaced by functions of linear terms in a way corresponding to (1) and (2).

Section 2 describes the principle of sufficient dimension reduction, and how such an approach can be used to estimate the η_i vectors ($i = 1, \cdots, d$) in the Tobit models described in this section. The specific dimension reduction estimator sliced inverse regression (SIR) is the focus of the discussion. In Section 3 Monte Carlo simulations are used to compare the performance of SIR to maximum likelihood estimation, as well as to other Tobit model estimators that have been proposed to address the presence of nonnormal and/or heteroscedastic errors. A real data example is analyzed in Section 4. Section 5 concludes the paper with a discussion of further potential work.

2 Sufficient Dimension Reduction and the Tobit Model

2.1 Sufficient dimension reduction

Sufficient dimension reduction considers the following regression structure for a univariate response y and a $p \times 1$ predictor vector X:

$$y \perp\!\!\!\perp X \,|\, \eta^\intercal X, \tag{3}$$

where $\perp\!\!\!\perp$ represents independence, and $\eta = (\eta_1, \ldots, \eta_d)$ is a $p \times d$ matrix with $d \leq p$. In practice, d is often far less than p; thus, we can replace the p-dimensional X with the d-dimensional $\eta^\intercal X$, and dimension reduction is achieved. More importantly, such a dimension reduction loses no regression information of y given X because of (3), so it is called *sufficient dimension reduction* (SDR; Cook, 1998).

Such an η always exists (by trivially taking η as the identity matrix). Since any basis of the subspace spanned by the columns of η, Span(η), leads to (3),

we call Span(η) a dimension reduction subspace. We further define the *central subspace*, denoted by $\mathcal{S}_{y|X}$, as the intersection of all dimension reduction subspaces. By definition, $\mathcal{S}_{y|X}$ is a unique and parsimonious population parameter that contains all regression information of $y \,|\, X$, and thus is the main object of interest in our dimension reduction inquiry. Its dimension, $d = \dim(\mathcal{S}_{y|X})$, is called the structural dimension of the regression.

The structure of $\mathcal{S}_{y|X}$ covers many known regression representations, such as the transformed linear regression model $h(y) = \eta_1^\intercal X + \varepsilon$, where h is a transformation function; the single-index model $y = f(\eta_1^\intercal X) + \varepsilon$, where f is a smooth link function; the heteroscedastic model $y = f(\eta_1^\intercal X) + g(\eta_2^\intercal X) \times \varepsilon$, where f and g are both univariate functions; and the nonparametric additive model $y = \sum_{j=1}^{d} f_j(\eta_j^\intercal X) + \varepsilon$, where f's are univariate functions. In all of these models ε is a random error independent of predictors with no restriction on its distribution.

Depending on available data and study-specific goals, regression analysis may focus more on the conditional mean $\mathrm{E}(y \,|\, X)$, and less on other aspects of the conditional distribution of $y \,|\, X$. In these situations, a dimension reduction inquiry hinges on finding a $p \times d$ matrix γ, with $d \leq p$, such that,

$$y \perp\!\!\!\perp \mathrm{E}(y \,|\, X) \,|\, \gamma^\intercal X.$$

That is, $\gamma^T X$ contains all of the information about y that is available through $\mathrm{E}(y \,|\, X)$. We call the subspace Span(γ) a mean dimension reduction subspace. Subsequently, the intersection of all such mean dimension reduction subspaces is called the *central mean subspace*, and is denoted as $\mathcal{S}_{E(y|X)}$. We assume the existence of $\mathcal{S}_{y|X}$ and $\mathcal{S}_{E(y|X)}$ throughout this article.

A dimension reduction subspace is always a mean dimension reduction subspace, and $\mathcal{S}_{E(y|X)} \subseteq \mathcal{S}_{y|X}$, since $y \perp\!\!\!\perp X \,|\, \eta^T X$ implies that $y \perp\!\!\!\perp \mathrm{E}(y \,|\, X) \,|\, \eta^T X$. In some cases, the central subspace and the central mean subspace coincide. For instance, for the transformed linear model, the single-index model, and the additive model that are discussed above, $\mathcal{S}_{E(y|X)} = \mathcal{S}_{y|X}$.

There are a number of numerical methods for estimating the central subspace and the central mean subspace, for instance, sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), principal Hessian directions (PHD; Li, 1992), and iterative Hessian transformation (IHT; Cook and Li, 2002). In this article, we focus on sliced inverse regression, one of the first and perhaps the most popular method for sufficient dimension reduction.

2.2 Sliced inverse regression

It is known that, under the linearity condition that is to be discussed below, the inverse mean vector, $\Sigma_x^{-1}\mathrm{E}(X \,|\, Y)$, resides in the central subspace (Li, 1991,

Cook, 1998). Thus, the population solution of SIR amounts to the following eigen-decomposition

$$\Sigma_{x|y}\ \upsilon_j = \lambda_j\ \Sigma_x\ \upsilon_j. \tag{4}$$

Here Σ_x denotes the covariance matrix of X, and $\Sigma_{x|y}$ denotes the covariance matrix of the inverse mean $\mathrm{E}(X \mid y)$. The eigenvectors, $\upsilon_1, \ldots, \upsilon_d$, that correspond to the d nonzero eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$, consist of a basis for the central subspace.

Given n independent realizations $\{(X_i, y_i), i = 1, \ldots, n\}$ of (X, y), SIR first partitions the range of y into h slices so that each y_i belongs to one of the slices. The sample estimate of $\mathrm{E}(X \mid y)$ is then obtained by averaging over all the X_i's whose corresponding y_i's belong to the same slice. The usual sample covariance matrices $\hat{\Sigma}_{x|y}$ and $\hat{\Sigma}_x$ are then computed and substituted in (4), resulting in the SIR sample estimates. The number of slices h is a tuning parameter in SIR, but it has been shown by various studies that the choice of h does not usually affect the SIR estimates, as long as $h > d$ and n is large enough for the asymptotics to provide useful approximations (Li, 1991, Cook, 1998). Under the linearity condition, the SIR estimates are known to be $\sqrt{n}$-consistent up to a multiplicative constant (Li, 1991).

The structural dimension d is determined by a sequence of tests of hypotheses, $d = k$ versus $d > k$, $k = 0, \ldots, p-1$ (Li, 1991). For a given k, the statistic, $\hat{\Lambda}_k = \sum_{j=k+1}^{p} \lambda_j$, where λ_j's are the eigenvalues in (4), has an asymptotic chi-squared distribution with $(p-k)(h-k-1)$ degrees of freedom, when X is multivariate normal, and is distributed as a linear combination of chi-squared variables, where the weights can be estimated consistently, when the linearity condition is met. Based on this statistic, we conclude that the structural dimension d is greater than k if the null hypothesis $d = k$ is rejected at a given nominal level. We then increment k by 1, and repeat the asymptotic test. The estimate of d is taken as the minimum k such that the null hypothesis $d = k$ is not rejected.

Note that SIR does not impose any traditional assumptions on the conditional distribution of $y \mid X$. Instead, it requires the linearity condition, an assumption placed on the marginal distribution of X, which states that, for any $b \in \mathbb{R}^p$,

$$\mathrm{E}(b^{\mathsf{T}} X \mid \eta_1^{\mathsf{T}} X, \ldots, \eta_d^{\mathsf{T}} X) = c_0 + c_1 \eta_1^{\mathsf{T}} X + \ldots + c_d \eta_d^{\mathsf{T}} X, \tag{5}$$

for some constants $c_0, c_1, \ldots, c_d$. Elliptical symmetry of the marginal distribution of X is sufficient for (5) to hold (Eaton, 1986), and in particular, (5) holds when X is multivariate normal. The linearity condition is not a severe restriction, since it holds to a reasonable approximation as p increases with d fixed (Hall and Li, 1993). In addition, the condition may be induced by predictor transformation, re-weighting (Cook and Nachtsheim, 1994), or clustering (Li, Cook, and Nachtsheim, 2004).

2.3 The SIR estimator for the tobit model

In this section we discuss the application of sliced inverse regression to different Tobit models. We begin with the Type I Tobit model.

Proposition 1. For the Type I Tobit model, $\mathcal{S}_{(y,\delta)\,|\,X} \subseteq \mathcal{S}_{y^*|X}$.

Proof: Let η be a basis for the central subspace $\mathcal{S}_{y^*|X}$ of regression of y^* on X. By definition, $y^* \perp\!\!\!\perp X \,|\, \eta^\tau X$. Since (y, δ) is a function of y^*, by Proposition 4.5 of Cook (1998), $(y, \delta) \perp\!\!\!\perp X \,|\, \eta^\tau X$. Thus Span($\eta$) is a dimension reduction subspace for regression of (y, δ) on X. Because $\mathcal{S}_{(y,\delta)\,|\,X}$ is the smallest dimension reduction subspace for (y, δ) on X, $\mathcal{S}_{(y,\delta)\,|\,X} \subseteq \mathcal{S}_{y^*|X}$.

Remark 1: Proposition 1 implies that one can gain information on the central subspace of interest, $\mathcal{S}_{y^*|X}$, by estimating the central subspace $\mathcal{S}_{(y,\delta)\,|\,X}$. Moreover, it is expected that equality between $\mathcal{S}_{(y,\delta)\,|\,X}$ and $\mathcal{S}_{y^*|X}$ will normally hold in practice, since proper containment (such that $\mathcal{S}_{(y,\delta)\,|\,X} \subset \mathcal{S}_{y^*|X}$) may often be the exception rather than the rule in practice (Cook, 1994).

Remark 2: Proposition 1 is applicable to *all* estimation methods for the central subspace. Here we focus on SIR as an example. For the Type I Tobit model, operationally, one only need to modify the usual slicing procedure of SIR to the so-called double slicing (Li, Wang, and Chen, 1999). That is, we first partition the response based on the indicator variable δ, and then slice y in each subgroup. Proposition 1 then implies that the SIR estimates are consistent estimates for the vectors $\eta_1, \ldots, \eta_d$ in (2).

Remark 3: The conclusion of Proposition 1 holds in more general cases. For instance, when $C \perp\!\!\!\perp X \,|\, (\eta^\tau X, y^*)$, or, when $C \perp\!\!\!\perp (X, y^*)$, one can show that $\mathcal{S}_{(y,\delta)\,|\,X} \subseteq \mathcal{S}_{y^*|X}$, where C is the censoring variable as defined in Section 1. See Cook (2003) for a more detailed discussion.

The next proposition deals with the Type II and Type III Tobit models.

Proposition 2 (i) For the Type II Tobit model, $\mathcal{S}_{E(\delta\,|\,X)} \subseteq \mathcal{S}_{E(y_1^*|X)}$, and $\mathcal{S}_{E((y_2,\delta)\,|\,X)} \subseteq \mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$. (ii) For the Type III Tobit model, $\mathcal{S}_{E((y_1,\delta)\,|\,X)} \subseteq \mathcal{S}_{E(y_1^*|X)}$, and $\mathcal{S}_{E((y_2,\delta)\,|\,X)} \subseteq \mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$. (iii) $\mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)} = \mathcal{S}_{E(y_1^*|X)} + \mathcal{S}_{E(y_2^*|X)}$.

Proof: (i) and (ii) can be shown by noting that both δ and (y_1, δ) are functions of y_1^*, and (y_2, δ) is a function of (y_1^*, y_2^*). (iii) holds following Proposition 4 of Cook and Setodji (2004).

Remark 4: For the Type II and Type III Tobit models, Proposition 2 suggests a way of estimating the central subspaces, $\mathcal{S}_{E(y_1^*|X)}$, $\mathcal{S}_{E(y_2^*|X)}$, and $\mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$,

respectively. For Type II models, Proposition 2 (i) indicates that one is able to estimate $\mathcal{S}_{E(y_1^*|X)}$ via estimating $\mathcal{S}_{E(\delta\,|\,X)}$ by applying sliced inverse regression on the observed binary response δ for the given predictor vector X. Moreover, one can estimate $\mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$ through $\mathcal{S}_{E((y_2,\delta)\,|\,X)}$ by applying SIR on bivariate responses (y_2, δ). For Type III models, both y_1 and δ are observed simultaneously. Hence, Proposition 2 (ii) gives similar results to Proposition 2 (i), except that $\mathcal{S}_{E(\delta\,|\,X)}$ in (i) is replaced by $\mathcal{S}_{E((y_1,\delta)\,|\,X)}$. Finally, Proposition 2 (iii) shows that the information from $\mathcal{S}_{E(y_2^*|X)}$ can be obtained by "subtracting" $\mathcal{S}_{E(y_1^*|X)}$ from $\mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$.

Remark 5: Proposition 2 is focused on the central mean subspace. If we extend our consideration to the central subspace, conclusions (i) and (ii) still hold, but (iii) should be changed to $\mathcal{S}_{y_1^*|X} + \mathcal{S}_{y_2^*|X} \subseteq \mathcal{S}_{(y_1^*,y_2^*)\,|\,X}$. As a result, we may obtain extra irrelevant information by subtracting $\mathcal{S}_{E(y_1^*|X)}$ from $\mathcal{S}_{E((y_1^*,y_2^*)\,|\,X)}$ when we seek the estimate of $\mathcal{S}_{E(y_2^*|X)}$.

Chen and Li (1998) proposed an approximate formula for standard deviations of SIR estimates. With $\lambda_1 \geq \ldots \geq \lambda_d$ denoting the d non-zero eigenvalues in SIR decomposition (4), the j-th SIR direction υ_j can be associated with the vector of the square root of the diagonal elements from the matrix,

$$\frac{1-\lambda_j}{\lambda_j}\; n^{-1}\; \Sigma_x^{-1}, \tag{6}$$

as the estimated standard deviations. Applying this result, one can obtain t-ratios for the elements of SIR estimators. The properties of such t-ratios are examined in the next section.

3 Comparison of SIR to Other Tobit Estimators

In this section we compare the performance of SIR for the Type I Tobit model to that of other estimators using Monte Carlo simulations, including the possibilities of nonnormal and/or heteroscedastic errors. The maximum likelihood estimator (MLE) is defined as the maximizer of the Tobit likelihood function

$$\prod_{\delta_i=0}[1-\Phi(\eta_1^{\intercal}x_i/\sigma)]\prod_{\delta_i=1}\sigma^{-1}\phi[(y_i-\eta_1^{\intercal}x_i)/\sigma)],$$

where Φ and ϕ are the distribution and density function, respectively, for the standard normal. Amemiya (1973) demonstrated consistency and asymptotic normality of the MLE under model (2) but Goldberger (1980) and Arabmazar and Schmidt (1982) showed that these properties are critically dependent on

the assumption of normality, and Arabmazar and Schmidt (1981) showed that the assumption of homoscedasticity is also required.

Powell (1984) addressed these difficulties by proposing that η_1 be estimated based on least absolute deviations (LAD), rather than least squares. The estimator minimizes

$$\sum_{i=1}^{n} |y_i - \max(C_i, \eta_1^{\intercal} x_i)|$$

over all η_1. Powell (1984) demonstrated that under various regularity conditions, but without assuming normality or homoscedasticity, the LAD estimator is consistent and asymptotically normal, and also described how the covariance matrix of $\hat{\eta}_1$ can be estimated in practice.

It is known that the LAD estimation algorithm sometimes suffers from convergence difficulties, particularly when the censoring proportion is high (Fitzenberger, 1997). Chernozhukov and Hong (2002) proposed a computationally simple three-step estimation procedure for quantile regression estimation of η_1. The estimator is based on (1) estimating a parametric model for δ, (2) fitting the (uncensored) quantile estimator to a subset of the observations with estimated probability of not being censored high enough based on the parametric model, and then (3) fitting the uncensored quantile estimator to the subset of observations with estimated fitted values (based on step (2)) larger than the observed censoring values. This results in a censored quantile regression (CQR) estimator with asymptotic properties identical to those of the LAD estimator when estimating the 50% quantile.

Censored regression data is common in survival data, although in that context censoring is typically from the right rather than from the left. Heller and Simonoff (1990) investigated the performance of various regression estimators under right censoring, and found that the method of Buckley and James (1979) performed best. This estimator is based on modifying the usual least squares normal equations to account for censoring using the Kaplan-Meier (Kaplan and Meier, 1958) product limit estimator of the error distribution. It can be adapted to left-censored data by substituting $-y$ for y, and then reversing the signs of all of the regression coefficients.

Each of the estimators described in this section thus far is designed for the linear model (1), rather than the more general model summarized in (2). An alternative to SIR for model (2) with $d = 1$ is the density weighted average derivative estimator (WADE; Powell, Stock, and Stoker, 1989), which is based on the result that

$$\mathrm{E}\left[g(x_i)\frac{\partial f(x_i)}{\partial x_i}\right] = -2\mathrm{E}\left[\frac{\partial g(x_i)}{\partial x_i} y_i\right]$$

is proportional to η_1, where $g(\cdot)$ is the density function of the predictors. This quantity is estimated using a kernel density estimator, and Powell and Stoker

(1996) described strategies for choosing the bandwidth for the kernel estimator.

We now describe the results of Monte Carlo simulation comparisons of the properties of different censored regression estimators. The first set of simulations examine the situation where the true regression model is the Type I Tobit model (1), implying that the estimators designed for this linear model should work best. The issue here, then, is whether the SIR estimator (which is designed for the more general model (2)) is competitive with estimators designed for the linear model. The situations examined correspond to several samples sizes n and number of predictors p: $(n = 50, p = 5)$, $(n = 200, p = 5)$, $(n = 200, p = 10)$, and $(n = 500, p = 20)$. The data were generated from a linear model with parameter vector $\eta_1^{\mathsf{T}} = \beta_1^{\mathsf{T}}/\sqrt{30}$ for $p = 5$, where $\beta_1^{\mathsf{T}} = (0, 1, 2, 3, 4)$, $\eta_1^{\mathsf{T}} = \beta_2^{\mathsf{T}}/\sqrt{60}$ for $p = 10$, where β_2 is β_1 twice, and $\eta_1^{\mathsf{T}} = \beta_3^{\mathsf{T}}/\sqrt{120}$ for $p = 20$, where β_3 is β_2 twice. The intercept term was then determined so as to yield censoring rates of 25% and 50%, based on a censoring distribution that was fixed at zero (random censoring was also examined, but the results were similar to those for fixed censoring). The errors ε were distributed as either Gaussian or Cauchy random variables with unit scale parameters (simulations based on double exponential and t_3 errors were also examined, but the comparative performance of the different estimators was very similar to that for Gaussian errors for those error distributions, and so are not given). The effectiveness of the estimators is measured by the average absolute correlation between the true index $\eta_1^{\mathsf{T}} X$ and estimated index $\hat{\eta}_1^{\mathsf{T}} X$ (recall that SIR and WADE only estimate the slope coefficients up to a multiplicative constant, since the model (2) is only identifiable up to a constant; this measure removes that scaling effect). There were 1000 simulation replications for each set of runs.

Table 1 summarizes the results of these simulations. The performance of the MLE, Buckley-James estimator (BJ), LAD estimator, weighted average derivative estimator (WADE), and SIR estimators are given. Results for the CQR estimator of Chernozhukov and Hong (2002) were virtually identical to those of the LAD estimator, so they are not given. As would be expected, the Gaussian-based MLE and least squares-based BJ are best under Gaussian errors, but the LAD and SIR estimators have performance that is close to that of the MLE. The WADE estimator lags behind slightly. Under Cauchy errors, the MLE and BJ estimators perform noticeably worse (particularly BJ), and WADE fails completely. On the other hand, the performance of the SIR estimator is similar to that of LAD, except for the smallest sample size (where it is still better than MLE and BJ). Thus, the SIR estimator is competitive with the estimators designed for the linear model, under both Gaussian and non-Gaussian errors.

Of course, the benefit of the SIR estimator comes for the more general model (2), when estimators such as MLE, BJ, and LAD are not necessarily appropriate. Figure 1 summarizes results for a set of nonlinear models that are consistent with model (2). The four models are defined as follows:

$$
\begin{aligned}
y_1 &= (1-\tau) \times 0.5\eta_1^{\mathsf{T}} X + \tau \times \exp(-\eta_1^{\mathsf{T}} X) \times \sin(0.5\pi\eta_1^{\mathsf{T}} X) + 0.5\varepsilon_1 \\
y_2 &= 0.25\eta_1^{\mathsf{T}} X + [(1-\tau) \times 0.2 + \tau \times \exp(-\eta_1^{\mathsf{T}} X)] \times \varepsilon_2 \\
y_3 &= \eta_1^{\mathsf{T}} X + 6[(1-\tau) \times 0.2\varepsilon_3 + \tau \times \varepsilon_4] \\
y_4 &= (1-\tau) \times 0.5\eta_1^{\mathsf{T}} X + \tau \times \exp(-0.75\eta_1^{\mathsf{T}} X) \times \varepsilon_5.
\end{aligned}
$$

We examine the case with $n = 200$, and $p = 10$. The predictors X are independent standard normal predictors, and the true direction is $\eta_1^{\mathsf{T}} = (1, 1, 1, 0, \ldots, 0)$. Censoring is fixed at zero, and all of the error terms ε_i are independent of X. The errors are standard normal, with the exception of ε_4, which is χ^2 on one degree of freedom.

In each model, the parameter τ controls the "distance" between the model and a Gaussian linear model (with linear mean, constant variance, and Gaussian error). With τ ranging from 0 to 1, model 1 is a deviation from a linear mean function, model 2 from constant variance, model 3 from a Gaussian error, and model 4 evolves into a model with no mean effect but only a variance component. Average absolute correlations of the true and estimated indices are given for $\tau = 0(.1)1$, which are then connected by lines, based on 100 simulation replications in each case.

It is apparent from Figure 1 that SIR provides consistently effective performance for all four models, just as it is designed to do, and is usually the most effective estimator. For models 1 (nonlinear mean function) and 3 (non-Gaussian error) the LAD and CQR estimators are second-best, with their performance deteriorating as the model becomes more nonlinear; the other estimators fare much worse. In models 2 (nonconstant variance) and 4 (variance-only) the pattern is similar, except that MLE, BJ, and WADE become a bit better as τ approaches 1 (although still behind SIR). The superiority of SIR over MLE, BJ, LAD, and CQR is not surprising, since the latter estimators are not designed for nonlinear mean functions, but the poor performance of WADE is striking; clearly the SIR estimator is a better choice for potentially nonlinear relationships.

In any regression application, inference about the statistical significance of slope parameters is an important consideration. As noted earlier, equation (6) provides a way of estimating the standard error of SIR slope estimates, and thereby constructing approximate t-ratios for significance testing. Table 2 summarizes the properties of such statistics in the linear model situation. The table gives the observed average size for the t-tests for all of the slope coefficients in a model under different conditions, based on 1000 simulation replications, when the nominal size is .05. In addition to the approximate standard errors based on (6), the table also gives results based on bootstrap estimates of the standard errors (Efron, 1979). The bootstrap is based on 100 replications, and corresponds to resampling all of the observations, rather than residuals. It is

apparent that the bootstrap works well in all of the situations, but the approximate standard errors also lead to reasonable (although slightly anticonservative) performance, except when the sample is small ($n = 50$); for $n = 500$, the approximate standard errors and bootstrap standard errors lead to t-ratios with virtually identical performance.

We close this section with a brief discussion of results when the structural dimension of the regression relationship d is greater than one. The other Tobit estimators, naturally, are not designed for this situation, but SIR estimation allows for the identification and estimation of multiple linear terms in (2). The models examined are similar, but not identical, to those examined in Li (1991). The first model is

$$y = x_1(x_2 + x_3 + 1) + \sigma\varepsilon$$

(corresponding to the two indices x_1 and $x_2 + x_3 + 1$). The predictors are generated to be normally distributed and uncorrelated with each other. The second model is

$$y = x_1/[.5 + (x_2 + 1.5)^2] + \sigma\varepsilon$$

(corresponding to two indices x_1 and $x_2+1.5$). All simulations have $n = 400$ and $p = 10$, and are based on 1000 replications. In all cases the resultant censoring proportion was roughly 50%, based on fixed-at-zero censoring.

Table 3 summarizes the results. The entries in the table are the absolute correlations between the actual and estimated index values. The methods other than SIR are not designed to estimate more than one index, of course, but it is still meaningful to see how correlated the fits implied by the single set of coefficient estimates that they produce are with the actual index values implied by the true coefficients. Since SIR will not identify the indices in a systematic way (the first column in the output does not necessarily correspond to the same index from simulation replication to replication), we determined the absolute correlations for all four possibilities (actual first index and index based on first SIR component, actual first index and index based on second SIR component, actual second index and index based on first SIR component, and actual second index and index based on second SIR component), and assigned the SIR components such that the average absolute correlation with the fits of the true indices was maximized. It can be seen that for all of the estimators other than SIR the estimated coefficients correspond to the first index, being much less correlated with the second index. The first estimated SIR index is only slightly less correlated with the first actual index than are those for MLE and BJ, but SIR also provides a second index that is highly correlated with the second actual index. LAV and CQR seem to mix the two components together more in their single set of estimated coefficients, as they have the lowest correlation with the first component, but a higher correlation than do MLE and BJ with the second component (although much lower than SIR's second component). As would be expected, performance is weaker for $\sigma = 1$ than for $\sigma = .5$.

We are also interested in whether the SIR test for the number of indices is effective. Table 4 gives the average p-values for the test of k versus at least $k+1$ components for $k = 0$ through 4; for both models, the first two tests should be statistically significant, while the last three should not be. The existence of at least one component is clear, and for the lower σ cases, the test for a second component is on average marginally statistically significant as well. As expected, none of the tests for more than two components are close to statistical significance on average. It is easier to identify the existence of the second component in the first model, which is consistent with the higher absolute correlations in Table 3.

4 Example

Yen (2002) studied data relating to donations of U.S. households to charitable organizations, using data from the 1995 Consumer Expenditure Survey. Of the 5085 households in the data, only 395 made any donations to charities, implying that (as Yen notes) a Tobit regression analysis is appropriate. In this paper we focus on the total amount donated by each household, with predictors including income, household size, the number of vehicles in the household, the number of wage earners, and the age of the head of the household. Yen (2002) also included several indicator variables as predictors, but since the consistency of SIR requires the linearity condition, it is less-suited for models with such predictors, so we omit them here. Several of the remaining predictors are long right-tailed, but we do not transform them, in order to stay consistent with the analysis in Yen (2002). Table 5 gives the Tobit maximum likelihood and SIR estimates, along with asymptotic standard errors. When analyzing the donations, according to the MLE (the first column of the table), household income, number of vehicles, and the age of the head of the household are all statistically significant predictors at a .05 level, and are all (as expected) directly related to the amount of donations.

Unfortunately, a cursory examination of the data shows that the donation amount is extremely long right-tailed, with increasing variability as the level of donations increases. The LAD estimator is insensitive to this heteroscedasticity, of course, but the high censoring proportion of over 92% means that neither the LAD nor CQR estimators are calculable. This suggests analyzing the data using the log of the donation amount (when it is positive) as the response, as was done in Yen (2002). This changes the inferential implications of the model, since the MLE for this response (second column) no longer finds the number of vehicles statistically significant.

Analysis using SIR is identical whether or not the response is logged, as is clear from model (2). The test of zero versus more than zero dimensions is highly significant (239.2 on 25 degrees of freedom), and the test of one ver-

sus more than one dimension is also significant (26.8 on 16 degrees of freedom, $p = .044$). None of the other tests are close to significance, indicating two dimensions. Note that this test is based on assuming a multivariate normal distribution for the predictors, so it can only be considered a rough guideline here. The coefficients for the two indices are given in the last two columns of Table 5. All of the coefficients are statistically significant for the first index, and this is clearly (based on the coefficients) a location index. Income, household size, number of vehicles, and age of the head of the household are all directly related to donation amount (given the other predictors). Interestingly, the number of earners is inversely related to donation amount, which might seem surprising, but is actually to be expected; it is easy to imagine that given the total household income, an increase in the number of earners in the household actually corresponds to a more difficult financial situation for a family, and lower levels of charitable donations.

The second index only includes two statistically significant predictors, number of earners and age of the head of household, both coefficients having the same sign. It appears that this index could be related to nonconstant variance. Figure 2 is a plot of the standardized residuals from a linear regression with the observed logged total donations as the response and the fitted first SIR index as the predictor versus each of these two variables, for those observations with positive donations. The pictures can only be suggestive, since all of the zero-donation households are not included, but it can be seen that for both variables (number of earners particularly) the variability decreases as the variable increases, suggesting that a standard Tobit fit to these data could be problematic because of heteroscedasticity.

5 Discussion and Conclusions

In this paper we have adapted the sliced inverse regression approach in sufficient dimension reduction to the Tobit (censored) regression situation. The resultant estimator is $\sqrt{n}$-consistent for true index coefficients up to a multiplicative constant under a wide range of circumstances, has good small-sample properties, and is easy to compute. Although we have focused here on the Type I Tobit model, it also can be adapted to Type II and Type III Tobit models, and further work in that direction is warranted. Since heteroscedasticity, nonnormality, and nonlinear relationships are common in econometric or business data for which the Tobit model is appropriate, it would seem that the SIR estimator is a useful alternative to consider when fitting such models.

References

Amemiya, T. (1973), "Regression analysis when the dependent variable is truncated normal," *Econometrica*, **41**, 997–1016.

Amemiya, T. (1984), "Tobit models: a survey," *Journal of Econometrics*, **24**, 3–61.

Arabmazar, A. and Schmidt, P. (1981), "Further evidence on the robustness of the Tobit estimator to heteroscedasticity," *Journal of Econometrics*, **17**, 253–258.

Arabmazar, A. and Schmidt, P. (1982), "An investigation of the robustness of the Tobit estimator to non-normality," *Econometrica*, **50**, 1055–1063.

Buckley, J. and James, I. (1979), "Linear regression with censored data," *Biometrika*, **66**, 429–436.

Chernozhukov, V. and Hong, H. (2002), "Three-step censored quantile regression and extramarital affairs," *Journal of the American Statistical Association*, **97**, 872–882.

Chen, C.H. and Li, K.C. (1998), "Can SIR be as popular as multiple linear regression?," *Statistica Sinica*, **8**, 289–316.

Cook, R.D. (1994), "On the interpretation of regression plots," *Journal of the American Statistical Association*, **89**, 177–190.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.

Cook, R.D. (2003), "Dimension reduction and graphical exploration in regression including survival analysis," *Statistics in Medicine*, **22**, 1399–1413.

Cook, R.D. and Li, B. (2002), "Dimension reduction for the conditional mean in regression," *Annals of Statistics*, **30**, 455–474.

Cook, R.D. and Nachtsheim, C.J. (1994), "Re-weighting to achieve elliptically contoured covariates in regression," *Journal of the American Statistical Association*, **89**, 592–600.

Cook, R.D. and Setodji, C.M. (2004), "A model-free test for reduced rank in multivariate regression," *Journal of the American Statistical Association*, **98**, 340–351.

Cook, R.D. and Weisberg, S. (1991), "Discussion of Li (1991)," *Journal of American Statistical Association*, **86**, 328–332.

Eaton, M. (1986), "A characterization of spherical distributions," *Journal of Multivariate Analysis*, **20**, 272–276.

Efron, B. (1979), "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, **7**, 1-26.

Fitzenberger, B. (1997), "Computational aspects of censored quantile regression," in *Proceedings of the 3rd International Conference on Statistical Data Analysis Based on the L1-Norm and Related Methods*, ed. Y. Dodge, Institute of Mathematical Statistics, Hayward, CA, 171–186.

Goldberger, A.S. (1980), "Abnormal selection bias," *SSRI Workshop Series no. 8006 University of Wisconsin, Madison.*

Hall, P. and Li, K.C. (1993), "On almost linearity of low dimensional projections from high dimensional data," *Annals of Statistics*, **21**, 867–889.

Heller, G. and Simonoff, J.S. (1990), "A comparison of estimators for regression with a censored response variable," *Biometrika*, **77**, 515–520.

Kaplan, E.L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, **53**, 457–481.

Li, K.C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, **86**, 316–327.

Li, K.C. (1992), "On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma," *Annals of Statistics*, **87**, 1025–1039.

Li, L., Cook, R.D., and Nachtsheim, C.J. (2004), "Cluster-based estimation for sufficient dimension reduction," *Computational Statistics and Data Analysis.* **47**, 175-193.

Li, K.C., Wang, J.L., and Chen, C.H. (1999), "Dimension reduction for censored regression data," *Annals of Statistics*, **27**, 1–23.

Powell, J.L. (1984), "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, **25**, 303-325.

Powell, J.L., Stock, J.H., and Stoker, T.M. (1989), "Semiparametric estimation of index coefficients," *Econometrica*, **57**, 1403–1430.

Powell, J.L. and Stoker, T.M. (1996), "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics*, **75**, 291–316.

Tobin, J. (1958), "Estimation of relationships for limited dependent variables," *Econometrica*, **26**, 24–36.

Yen, S.T. (2002), "An econometric analysis of household donations in the USA," *Applied Economics Letters*, **9**, 837–841.

	25% censoring proportion		50% censoring proportion	
	Gaussian	*Cauchy*	*Gaussian*	*Cauchy*
	$n = 50, p = 5$			
MLE	.953	.680	.944	.696
BJ	.952	.625	.943	.611
LAD	.926	.870	.897	.817
WADE	.898	.573	.889	.525
SIR	.933	.776	.907	.893
	$n = 200, p = 5$			
MLE	.989	.843	.986	.879
BJ	.989	.636	.986	.617
LAD	.981	.974	.971	.965
WADE	.949	.553	.947	.494
SIR	.980	.958	.984	.957
	$n = 200, p = 10$			
MLE	.975	.707	.971	.782
BJ	.975	.525	.971	.558
LAD	.958	.935	.938	.892
WADE	.898	.433	.889	.376
SIR	.971	.911	.965	.905
	$n = 200, p = 20$			
MLE	.979	.710	.975	.800
BJ	.979	.442	.975	.503
LAD	.963	.944	.947	.910
WADE	.880	.327	.870	.274
SIR	.976	.930	.972	.926

Table 1: Average absolute correlations of the true and estimated linear predictors for different estimators under linear model conditions.

	25% censoring proportion		50% censoring proportion	
	Gaussian	*Cauchy*	*Gaussian*	*Cauchy*
$n = 50, p = 5$	.138	.202	.126	.235
	.028	.065	.032	.075
$n = 200, p = 5$	.067	.072	.081	.075
	.059	.034	.061	.035
$n = 200, p = 10$	.076	.086	.068	.089
	.057	.046	.048	.045
$n = 200, p = 20$	.058	.056	.056	.060
	.055	.045	.054	.054

Table 2: Average empirical size of .05 level SIR t-tests for slope coefficients. The first line corresponds to using approximate standard errors, while the second line corresponds to using bootstrap standard errors.

Model 1: $y = x_1(x_2 + x_3 + 1) + \sigma\varepsilon$

	$\sigma = .5$		$\sigma = 1$	
	Component 1	*Component 2*	*Component 1*	*Component 2*
SIR	.923	.774	.924	.650
MLE	.952	.227	.949	.189
BJ	.949	.243	.949	.191
LAV	.869	.475	.862	.462
CQR	.867	.478	.863	.456

Model 2: $y = x_1/[.5 + (x_2 + 1.5)^2] + \sigma\varepsilon$

	$\sigma = .5$		$\sigma = 1$	
	Component 1	*Component 2*	*Component 1*	*Component 2*
SIR	.933	.772	.909	.510
MLE	.947	.243	.929	.180
BJ	.947	.246	.930	.179
LAV	.875	.426	.850	.386
CQR	.875	.424	.857	.359

Table 3: Average absolute correlations of estimated indices and actual indices for two-index models.

Model	σ	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
1	.5	2.9×10^{-11}	.062	.531	.747	.805
	1	1.1×10^{-7}	.184	.589	.780	.822
2	.5	1.7×10^{-9}	.075	.563	.772	.825
	1	6.1×10^{-4}	.339	.666	.808	.841

Table 4: Average p-values for test of k versus at least $k + 1$ components for two-index models.

	MLE		SIR	
	Total donations	*Logged total donations*	*First index*	*Second index*
Intercept	−5543.36	−17.287		
	(322.08)	(1.090)		
Income	0.126	0.350	0.172	−0.013
	(0.012)	(0.038)	(0.002)	(0.009)
Household size	0.269	0.284	0.507	−0.321
	(0.289)	(0.883)	(0.052)	(0.200)
Number of vehicles	0.544	0.698	0.484	−0.270
	(0.247)	(0.770)	(0.048)	(0.184)
Number of earners	−0.773	−0.373	−0.665	0.906
	(0.511)	(1.554)	(0.094)	(0.361)
Age of household head	0.137	0.414	0.192	0.049
	(0.023)	(0.071)	(0.004)	(0.016)

Table 5: Tobit regression results for charitable donation data.

Figure 1: Average absolute correlations of the true and estimated linear predictors for different estimators under four nonlinear model conditions. MLE: long dashed-and-dotted line; BJ: short dashed-and-dotted line; LAD: long dashed line; CQR: short dashed line; WADE: dotted line; SIR: solid line.

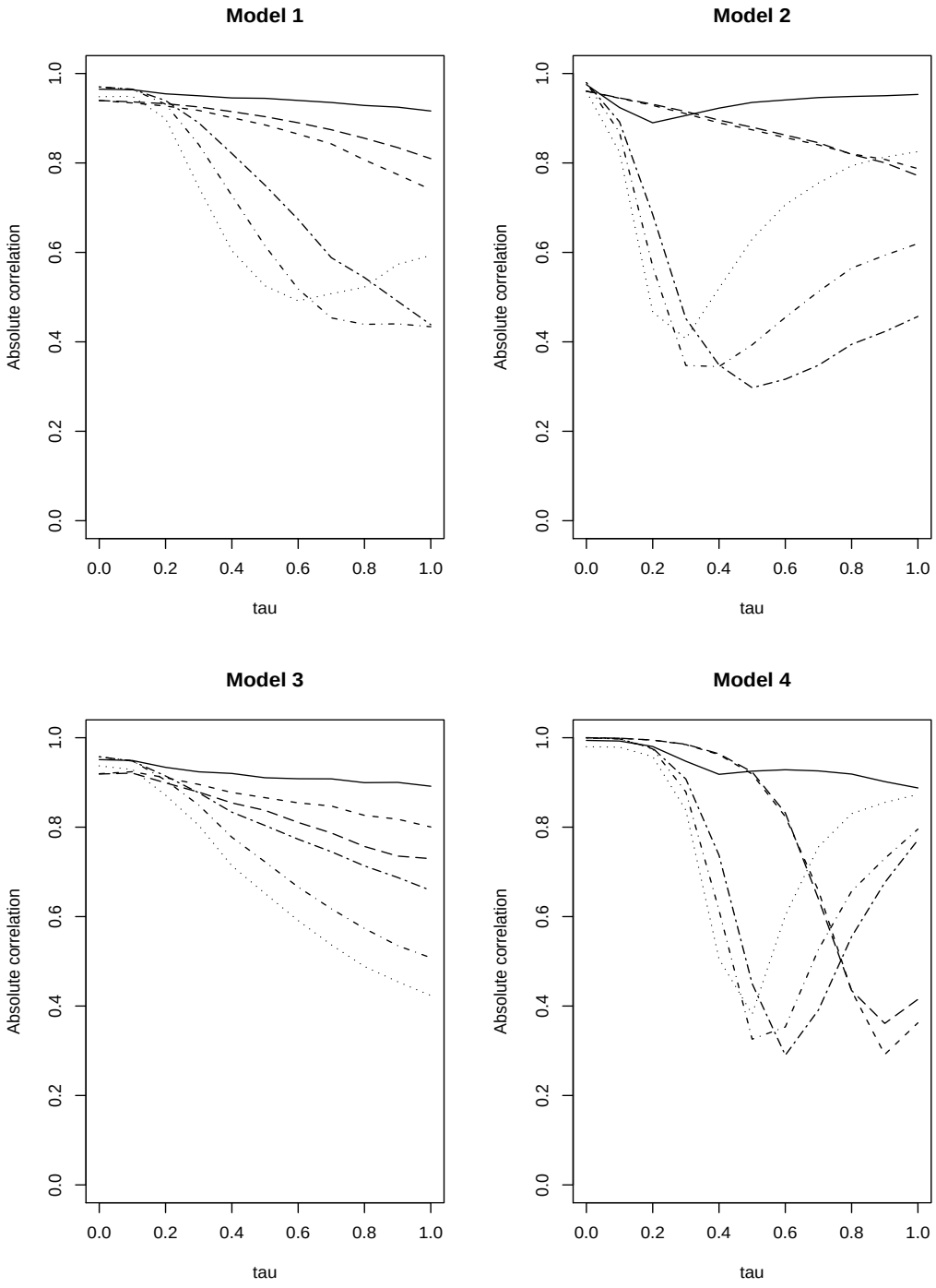

Figure 2: Plots of the standardized residuals from a linear regression with the observed logged total donations as the response and the fitted first SIR index as the predictor versus number of earners and age of head of household, respectively, for households with positive charitable donations.