The Annals of Statistics
1994, Vol. 22, No. 2, 747–762

ON USING STRATIFICATION IN THE ANALYSIS OF LINEAR REGRESSION MODELS WITH RIGHT CENSORING

BY MENDEL FYGENSON AND MAI ZHOU

Rutgers University and University of Kentucky

We study two modified synthetic data least-squares estimation methods for linear regression models with right censored response variables, unspecified residual distributions and random censoring variables which may not be i.i.d. These methods are the result of an investigation into the use of stratification. We conclude that stratification should be used whether or not the censoring variables are dependent on the covariates. We give the asymptotic results of the estimators and numerical results.

1. Introduction. Consider the linear regression model

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i, \tag{1.1}$$

where ε_i are i.i.d. random variables with mean zero and finite variance, and $\mathbf{X}_i$ are covariates which are either nonrandom variables or random variables that are independent of the ε_i. In this paper we consider the case in which the dependent variables are generated by a random censorship; that is, one observes $(T_i, \delta_i, \mathbf{X}_i)$, with

$$T_i = \min(Y_i, C_i), \qquad \delta_i = I_{[Y_i \le C_i]}, \qquad i = 1, 2, \dots, n,$$

where C_i are independent random variables that are independent of the ε_i, and $I_{[A]}$ is the indicator function of the event A.

This model has been used extensively in medical applications and in many other fields [cf. Miller (1981), Amemiya (1985) and Maddala (1983)]. The majority of the proposed estimators for the vector parameters β have been obtained by modifying the least-squares method to accommodate right censoring. These include the ones proposed by Miller (1976), Buckley and James (1979) and the "synthetic data" method of Koul, Susarla and Van Ryzin (1981) (hereafter abbreviated as KSV). These methods differ with respect to their assumptions about C_i [from the highly restrictive (Miller) to the minimally restrictive (Buckley–James)] and their reliance on the i.i.d. assumption of the ε_i (the Buckley–James method relies on this assumption, the KSV method does not). Unlike the Miller and Buckley–James methods, which require special programming and extensive computer time and have convergence problems, the KSV method is accessible since the estimators can be obtained quickly and easily using a standard regression package. In addition, the KSV method is applicable to survival data

Received May 1990; revised June 1993.
AMS 1991 *subject classifications.* Primary 62G05; secondary 62J05, 62E20, 62E25.
Key words and phrases. Censored data, linear models, least-squares estimators, stratification.

even when the error distribution differs from one patient to another, as often occurs in practice.

Despite the above advantages, the KSV method is not used because it has been found to provide unreasonable estimators in cases where the censoring times are not i.i.d. and/or are dependent on the covariates [cf. Miller and Halpern (1982) and Leurgans (1987)]. We conducted an extensive simulation study and found that the KSV method is extremely biased even when the censoring times *are* i.i.d. and do *not* vary with the covariates. Leurgans (1987), assuming i.i.d. censoring times, proposed an improved "synthetic data" method that seems to provide better estimators than the KSV procedure while retaining its advantages.

In this paper, we show how stratification can be used with these synthetic data methods to increase their applicability to many data sets and to increase the efficiency of their estimators when the censoring times are i.i.d. random variables.

First, we note that in practice the censoring times often are not i.i.d. random variables. We, therefore, consider the more realistic assumption that the censoring times are independent random variables from k, $1 < k < n$, strata. This allows group dependence of the censoring distribution on the covariates. We assume, without loss of generality, that the following hold:

$$\begin{aligned}
&C_1, C_2, \dots, C_{n_1} \text{ are i.i.d. distributed as } G_1(\cdot);\\
&C_{n_1+1}, C_{n_1+2}, \dots, C_{n_1+n_2} \text{ are i.i.d. distributed as } G_2(\cdot);\\
&\qquad\qquad\vdots \\
&C_{n_1+\cdots+n_{k-1}+1}, C_{n_1+\cdots+n_{k-1}+2}, \dots, C_{n_1+\cdots+n_{k-1}+n_k}\\
&\text{are i.i.d. distributed as } G_k(\cdot).
\end{aligned} \tag{1.2}$$

Under this assumption, we propose a corrected KSV (CKSV) method which corrects the bias without resorting to Leurgans' alternative synthetic data. The CKSV method is based on using stratification and redefining the largest observation in each stratum as uncensored. We show that, for the two-or-more-sample case, the CKSV estimators are identical to Leurgans estimators; see Theorem 2.2. Further support for our CKSV method is provided by numerical simulations.

We also investigate the use of a stratification-based method in cases where the censoring times are i.i.d. random variables. This method, termed *artificial stratification*, calls for treating the i.i.d. censoring times as if they were not i.i.d. random variables but rather independent random variables from $k > 1$ strata, as in assumption (1.2). The use of this counterintuitive method with the KSV and Leurgans procedures results in more efficient estimators of β in large samples; see Theorems 3.1 and 3.2.

The article is structured as follows. In Section 2 we propose our corrected KSV method and a stratified Leurgans procedure and establish the asymptotic distributions of their estimators. We present artificial stratification and its use in improving the KSV and Leurgans procedures in Section 3. We conclude in Section 4 with simulations and closing remarks.

We close this section with some notation and definitions: for any real number t, $P(Y_i \le t) = F_i(t)$ with density $f_i(t)$; $P(C_i \le t) = G_{g(i)}(t)$; and $1 - H_i(t) = P(T_i > t) = (1 - F_i(t))(1 - G_{g(i)}(t))$, where $g(\cdot)$ is the stratum index, that is, $g(i) = j$ if and only if $n_1 + \cdots + n_{j-1} < i \le n_1 + \cdots + n_{j-1} + n_j$.

We shall use the following well-known martingales associated with our model:

$$M_i^D(t) = I_{[T_i \le t,\, \delta_i = 1]} - \int_{-\infty}^{t} I_{[T_i \ge s]}\, d\Lambda_i^D(t),$$

$$M_i^C(t) = I_{[T_i \le t,\, \delta_i = 0]} - \int_{-\infty}^{t} I_{[T_i \ge s]}\, d\Lambda_i^C(t),$$

with respect to the filtration $\mathcal{F}_t = \sigma\{T_i I_{[T_i \le t]};\ \delta_i I_{[T_i \le t]};\ i = 1, 2, \dots, n\}$, where

$$\Lambda_i^D(t) = \int_{[-\infty, t]} \frac{dF_i(s)}{1 - F_i(s-)}, \qquad \Lambda_i^C(t) = \int_{[-\infty, t]} \frac{dG_{g(i)}(s)}{1 - G_{g(i)}(s-)}.$$

[See, e.g., Gill (1980).] We also define the two related counting processes by

$$R_j(t) = \sum_{g(i) = j} I_{[T_i \ge t]} \quad \text{and} \quad N_j^C(t) = \sum_{g(i) = j} I_{[T_i \le t,\, \delta_i = 0]}.$$

Finally, let

$$\overline{F}(t) = \lim \frac{1}{n} \sum F_i(t), \qquad 1 - \widehat{H}_i(t) = I_{[T_i > t]},$$

$$h_i(t) = \int_t^{\infty} s\, dF_i(s)_x \quad \text{and} \quad \tilde{h}_i(t) = \int_t^{\infty} \bigl(1 - F_i(s)\bigr)\, ds.$$

2. The corrected KSV and the stratified Leurgans estimators.

2.1. *The corrected KSV estimators.* Koul, Susarla and Van Ryzin (1981) considered the linear equation

$$E\left[\frac{\delta_i T_i}{1 - G(T_i)} \middle| \mathbf{X}_i\right] = \mathbf{X}_i \beta, \tag{2.1}$$

where $G(t)$ is the distribution of the censoring variables C_i, which they assume to be i.i.d. random variables. They then substituted an estimator $\widehat{G}(t)$ for $G(t)$ and solved the usual least-squares normal equations based on (2.1). While Koul, Susarla and Van Ryzin used a variation of the Kaplan–Meier (product-limit) estimator for $G(t)$ in their original proof to avoid some technical difficulties, we will use the standard, left-continuous version of the Kaplan–Meier estimator and a different technique in deriving the asymptotic results.

The great advantage of the KSV method is its accessibility and simplicity. Unlike the Buckley–James and Miller procedures, this method is not iterative. Therefore it does not require the selection of an initial value, and there are no convergence problems. Once the synthetic data $\widehat{Y}_i^* = \delta_i T_i(1 - \widehat{G}(T_i-))^{-1}$ are computed, one can use a standard least-squares computing package to obtain the estimators

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{Y}}^* \quad \text{where } \widehat{Y}_i^* = \frac{\delta_i T_i}{1 - \widehat{G}(T_i-)}. \tag{2.2}$$

This advantage is greatly magnified when one considers diagnostic procedures or alternative models to (1.1), as is usually done in regression analysis.

The main known disadvantage of this method is its unrealistic assumption that the C_i are i.i.d. random variables with a distribution which does not depend on the covariates [see Miller and Halpern (1982) and Leurgans (1987)]. To improve the KSV method, we replace this assumption with assumption (1.2) and make the following adjustments. First, we derive the Kaplan-Meier estimators $\widehat{G}_j(t)$ for each stratum $j = 1, 2\dots, k$. Then, we compute the corrected synthetic data

$$\widehat{Y}_i^{(c)} = \frac{\delta_i T_i}{1 - \widehat{G}_{g(i)}(T_i-)}, \tag{2.3}$$

where

$$\delta_i = \begin{cases} 1, & \text{if } Y_i \le C_i \text{ or } Y_i \text{ is the maximum in its stratum,} \\ 0, & \text{otherwise.} \end{cases}$$

Based on the corrected synthetic data $(\widehat{Y}_i^{(c)}, \mathbf{X}_i)$, the least-squares estimator is

$$\widehat{\beta}^{(c)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{Y}}^{(c)}. \tag{2.4}$$

Now we consider the asymptotic distribution of the corrected estimators, hereafter referred to as the CKSV estimators. For simplicity, we present the results for the simple linear regression. The extension of these results to the multiple regression is straightforward. We also assume the X_i are i.i.d. from some nondegenerate distribution with a finite fourth moment, that is, a random design.

The following notations and assumptions are needed to state our theorem:

$$b_{ni} = \frac{X_i - \overline{X}}{\sum_{j=1}^n (X_j - \overline{X})^2}; \qquad a_{ni} = \frac{1}{n} - \overline{X} b_{ni}; \qquad \overline{X} = \frac{1}{n}\sum X_i.$$

(A1) The density $f_i(t)$ has bounded variation on any finite interval.

(A2) There exist $K > 0$ and $0 < b < 1$ such that, for all $i, 1 - G_{g(i)}(t) > K(1 - F_i(t))^b$.

(A3)
$$\frac{f_i(t)}{[1 - F_i(t)]^a} \le g(t) \quad \text{with} \quad a = \frac{1+b}{2} \quad \text{and}$$
$$\int_\tau^\infty t g(t)\, dt < \infty \quad \text{for some } \tau < \infty.$$

(A4)
$$\sup_{0<t<\infty} \frac{t^2 (1/n) \sum f_i^2(t)}{(1 - \overline{F})^\gamma} < \infty \quad \text{for some } \gamma < 2.$$

(A5)
$$\sup_i \int_{-\infty}^\tau t^2\, dF_i(t) \to 0 \quad \text{as } \tau \to -\infty.$$

(A6)
$$\sup_{t>0} E\left[\varepsilon_i - t \mid \varepsilon_i > t\right] < \infty.$$

(B1)
$$\sqrt{n} \max_i \frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2} \frac{T_j^n}{\left[1 - G_j(T_j^n)\right]} \to_P 0 \ , \quad \text{as } n \to \infty.$$

(B2)
$$\frac{d\left(F_i(t) \big/ [1 - F_i(t)]\right) \big/ dt}{d\left(G_j(t) \big/ [1 - G_j(t)]\right) \big/ dt} \to 0 \quad \text{as } t \to \infty.$$

(S1) For $j = 1, \ldots, k$, $T_j^n \to \infty$ as $n \to \infty$, where $T_j^n = \max_i \{T_{g(i)}; g(i) = j\}$.

(S2) $(n_j/n) \to \lambda_j$ as $n \to \infty$ and $\lambda_j \in (0, 1)$.

REMARK 2.1. Conditions (A2)–(A4) are needed to insure the estimators' behavior as $t \to \infty$ in cases where the censoring distribution $G_{g(i)}$ and the survival time distribution F_i share a common support on the positive real line. They require that F_i have a smaller tail than $G_{g(i)}$. For example, when $1 - F_i(t) \sim \exp(-\lambda_i t)$ and $1 - G(t) \sim \exp(-\mu_{g(i)} t)$ with $\lambda_i - \mu_{g(i)} \le \varepsilon > 0$, (A2)–(A4) all hold. These conditions are not needed, however, in cases where the support of the censoring distribution is larger than that of the survival distribution.

In redefining the largest observation in each stratum [see (2.3)] the martingale structure of the estimator is destroyed. To overcome this difficulty when deriving the asymptotic results, we need to assume that either the size of the last jump is asymptotically negligible [assumption (B2)] or that jump occurs with a negligible probability [assumption (B1)].

THEOREM 2.1. *Suppose that the variance–covariance matrix elements $\sigma_{ij}(\tau)$ defined below are finite for all τ and that $\sigma_{ij}(\tau) \to \sigma_{ij}(\infty)$ as $\tau \to \infty$. Under assumptions* (A1)–(A6), (S1), (S2) *and either* (B1) *or* (B2), *the CKSV estimators have limiting normal distributions with means (α^*, β^*) and covariance matrix $\Sigma = (\sigma_{ij})$, where*

$$\alpha^* = \sum_i a_{ni} \int_{-\infty}^{T^n_{g(i)}} t\, dF_i(t), \qquad \beta^* = \sum_i b_{ni} \int_{-\infty}^{T^n_{g(i)}} t\, dF_i(t)$$

and

$$\begin{aligned}\sigma_{22}(\tau) = {}& \lim n \sum_{i=1}^{n} b_{ni}^2 \int_{-\infty}^{\tau} \left(\frac{t}{1 - G_{g(i)}(t)} - \frac{\int_t^\tau s\, dF_i(s)}{1 - H_i(t)} \right)^2 [1 - H_i(t)]\, d\Lambda_i^D(t) \\ & + \lim n \sum_{i=1}^{n} b_{ni}^2 \int_{-\infty}^{\tau} \left(\frac{\sum_{l:g(l)=g(i)} b_{nl} \int_t^\tau s\, dF_l(s)}{\sum_{l:g(l)=g(i)} (1 - H_l(t))} - \frac{b_{ni} \int_t^\tau s\, dF_i(s)}{1 - H_i(t)} \right)^2 \\ & \times [1 - H_i(t)]\, d\Lambda_i^C(t),\end{aligned}$$

$\sigma_{11}(\tau) =$ *the same as* σ_{22} *with* b_{ni} *replaced by* a_{ni},

$$\begin{aligned}\sigma_{12}(\tau) = {}& \lim n \sum_{i=1}^{n} a_{ni} b_{ni} \int_{-\infty}^{\tau} \left(\frac{t}{1 - G_{g(i)}(t)} - \frac{\int_t^\tau s\, dF_i(s)}{1 - H_i(t)} \right)^2 [1 - H_i(t)]\, d\Lambda_i^D(t) \\ & + \lim n \sum_{i=1}^{n} a_{ni} b_{ni} \int_{-\infty}^{\tau} \prod_{e_{nl} = b_{nl}, a_{nl}} \left(\frac{\sum_{l:g(l)=g(i)} e_{nl} \int_t \tau s\, dF_l(s)}{\sum_{l:g(l)=g(i)} (1 - H_l(t))} \right. \\ & \left. - \frac{\int_t^\tau s\, dF_i(s)}{1 - H_i(t)} \right) [1 - H_i(t)]\, d\Lambda_i^C(t).\end{aligned}$$

PROOF From the definition of $\widehat{\beta}^{(c)}$ and β^*, we have

$$\widehat{\beta}^{(c)} - \beta^* = \sum b_{ni} \left[\int_{-\infty}^{T^n_{g(i)}} \frac{t}{1 - \widehat{G}_{g(i)}(t)}\, dI_{[T_i \le t,\, \delta_i = 1]} - \int_{-\infty}^{T^n_{g(i)}} t\, dF_i(t) \right].$$

This can be written as

$$\begin{aligned}\widehat{\beta}^{(c)} - \beta^* = {}& \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \frac{t}{1 - \widehat{G}_{g(i)}(t-)}\, dM_i^D(t) \\ & + \sum b_{ni} \int_{-\infty}^{T^n_g(i)} \left(\frac{1 - G_{g(i)}}{1 - \widehat{G}_{g(i)}(t-)} \frac{1 - \widehat{H}_i}{1 - H_i} - 1 \right) t\, dF_i(t) \\ = {}& \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \frac{t}{1 - \widehat{G}_{g(i)}(t-)}\, dM_i^D(t) \\ & + \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \left(\frac{H_i - \widehat{H}_i}{1 - H_i} + \frac{\widehat{G}_{g(i)}(t-) - G_{g(i)}}{1 - G_{g(i)}} \right) t\, dF_i(t) \\ & + \text{two higher-order terms.}\end{aligned}$$

Ignoring the higher-order terms and integrating the second term by parts, we get

$$
\begin{aligned}
\widehat{\beta}^{(c)} - \beta^* \approx{} & \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \frac{t}{1-\widehat{G}_{g(i)}(t-)}\, dM_i^D(t) \\
& + \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \left[\int_t^{\infty} s\, dF_i(s)\right] d\left(\frac{H_i(t)-\widehat{H}_i(t)}{1-H_i(t)}\right) \\
& + \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \left[\int_t^{\infty} s\, dF_i(s)\right] d\left(\frac{\widehat{G}_{g(i)}(t)-G_{g(i)}(t)}{1-G_{g(i)}}\right).
\end{aligned}
$$

Rewriting the above in terms of integration with respect to the individual martingales M_i^D and M_i^C [recall $h_i(t) = \int_t^{\infty} s\, dF_i(s)$] and collecting terms, we get

$$
\begin{aligned}
\widehat{\beta}^{(c)} - \beta^* ={} & \sum b_{ni} \int_{-\infty}^{T^n_{g(i)}} \left(\frac{t}{1-\widehat{G}_{g(i)}(t-)} - \frac{h_i(t)}{1-H_i(t)}\right) dM_i^D(t) \\
& + \sum \int_{-\infty}^{T^n_{g(i)}} \left(\sum_{l:g(l)=g(i)} b_{nl} h_l(t) \frac{1-\widehat{G}_{g(i)}(t-)}{1-G_{g(i)}} \frac{1}{R_{g(i)}(t)} \right. \\
& \qquad \left. - \frac{b_{ni}h_i(t)}{1-H_i(t)} \right) dM_i^C(t).
\end{aligned} \tag{2.5}
$$

The integrands in the last sums are predictable processes. It follows that the above is a martingale. To show that the central limit theorem for martingales is applicable here, we checked the Lindeberg condition and demonstrated that the higher-order terms are $o_p(1)$. [A similar verification can be found in Srinivasan and Zhou (1991).]

Next, consider the predictable variation process of the martingale (2.5):

$$
\begin{aligned}
\langle(2.5)\rangle ={} & \sum b_{ni}^2 \int_{-\infty}^{T^n_{g(i)}} \left(\frac{t}{1-\widehat{G}_{g(i)}(t-)} - \frac{h_i(t)}{1-H_i(t)}\right)^2 I_{[T_i \ge t]}\, d\Lambda_i^D(t) \\
& + \sum \int_{-\infty}^{T^n_{g(i)}} \left(\sum_{l:g(l)=g(i)} b_{nl} h_l(t) \frac{1-\widehat{G}_{g(i)}(t-)}{1-G_{g(i)}(t)} \frac{1}{R_{g(i)}(t)} - \frac{b_{ni}h_i(t)}{1-H_i(t)} \right)^2 \\
& \qquad \times I_{[T_i \ge t]}\, d\Lambda_i^D(t).
\end{aligned}
$$

When normalized by multiplying by n, it is not hard to show that the above has a p-limit [$= \sigma_{22}(\infty)$],

$$
\begin{aligned}
& \lim n\langle \widehat{\beta}^{(c)} - \beta^* \rangle \\
& \quad = \lim n \sum b_{ni}^2 \int_{-\infty}^{\infty} \left(\frac{t}{1-G_{g(i)}(t)} - \frac{h_i(t)}{1-H_i(t)}\right) [1-H_i(t)]\, d\Lambda_i^D(t) \\
& \qquad + \lim n \sum \int_{-\infty}^{\infty} \left(\frac{\sum_{l:g(l)=g(i)} b_{nl} h_l(t)}{\sum_{l:g(l)=g(i)} (1-H_l)} - \frac{b_{ni}h_i(t)}{1-H_i(t)} \right)^2 (1-F_i)\, dG_{g(i)}(t). \qquad \square
\end{aligned}
$$

The following alternative expression for the asymptotic variance can be obtained by expanding the square and collecting terms in σ_{22}:

$$\begin{aligned}\sigma_{22}(\infty) &= \lim n\langle\widehat{\beta}^{(c)} - \beta^*\rangle \\ &= \lim \operatorname{Var}\left(\sqrt{n}\sum b_{ni}\frac{T_i\delta_i}{1-G_{g(i)}(T_i)}\right) - \lim n\sum_j \frac{\left(\sum b_{nl}h_l(t)\right)^2}{\sum(1-H_l)}\frac{dG_j}{1-G_j}.\end{aligned} \tag{2.6}$$

This expression suggests an estimator of σ_{22}:

$$\begin{aligned}\widehat{\sigma}_{22} &= n\sum b_{ni}^2\left(\frac{T_i\delta_i}{1-\widehat{G}_{g(i)}(T_i)} - \widehat{\alpha}^{(s)} - \widehat{\beta}^{(s)}X_i\right)^2 \\ &\quad - n\sum_j\int\frac{\left[\sum_{l:g(l)=j} b_{nl}\int_t^\infty s\,d\left(I_{[T>s]}/[1-\widehat{G}_l(s)]\right)\right]^2}{R_j(t)-1}\frac{dN_j^C(t)}{R_j(t)}.\end{aligned}$$

2.2. *The stratified Leurgans estimators.* Assuming that the C_i are i.i.d. random variables and independent of the $\mathbf{X}_i$, Leurgans (1987) suggested transforming the censored observations (T_i, δ_i) into the synthetic data

$$\widehat{Y}_i^{(L)} = \int_{-\infty}^{T^n \vee 0}\left(\frac{I_{[T_i\ge t]}}{1-\widehat{G}(t)} - I_{[t<0]}\right)dt, \qquad i = 1,\dots,n, \text{ where } T^n = \max T_j^n, \tag{2.7}$$

and $\widehat{G}(t)$ is the Kaplan–Meier estimator of the censoring distribution. [We introduced a minor modification, $\vee 0$, in the integral limit, which does not matter asymptotically if $T^n \to \infty$ as we assume in (S1).] She then applied the usual least-squares procedure to $(\widehat{Y}_i^{(L)}, \mathbf{X}_i)$. Zhou (1992) showed that these estimators are asymptotically normally distributed.

We use stratification to adjust Leurgans' method to meet assumption (1.2). Leurgans (1987) used this same adjustment to her method in the analysis of the Stanford heart transplant data without formally considering assumption (1.2) and without deriving theoretical results for the estimators.

For the simple linear model

$$E(Y_i \mid X_i) = \alpha + \beta X_i, \qquad i = 1,\dots,n,$$

assuming (1.2), we define the stratified synthetic data as

$$\widehat{Y}_i^{(s)} = \int_{-\infty}^{T_{g(i)}^n \vee 0}\left(\frac{I_{[T_i\ge t]}}{1-\widehat{G}_{g(i)}} - I_{[t<0]}\right)dt, \tag{2.8}$$

where $\widehat{G}_{g(i)}(t)$ is derived in subsection 2.1. Applying the least-squares procedure to $(\widehat{Y}_i^{(s)}, X_i)$, the stratified Leurgans estimators are

$$\widehat{\beta} = \sum b_{ni}\widehat{Y}_i^{(s)}, \qquad \widehat{\alpha} = \sum a_{ni}\widehat{Y}_i^{(s)}.$$

In a manner similar to the proof of Theorem 2.1, we can prove the asymptotic normality of the stratified Leurgans estimators. (Details are available from the authors upon request.) See Section 3.2 for an expression for the asymptotic variance.

Although the CKSV procedure differs from the stratified Leurgans method in motivation and in its synthetic data transformation, we prove next that it is equivalent to the Leurgans method in some special cases. This provides further support for the CKSV procedure.

THEOREM 2.2. *For the simplest nontrivial linear model, the $k(\geq 2)$-sample case, the CKSV and stratified Leurgans methods produce the same least-squares estimators.*

PROOF. We have to show that $(1/n_j)\Sigma_{g(i)=j}\widehat{Y}_i^{(s)} = (1/n_j)\Sigma_{g(i)=j}\widehat{Y}_i^{(c)}$ for each $j = 1, 2, \dots, k$, where $\widehat{Y}_i^{(s)}$ is the Leurgans synthetic data defined in (2.9) and $\widehat{Y}_i^{(c)}$ is defined in (2.3).

We start by rewriting the Leurgans estimator:

$$(2.9) \qquad \frac{1}{n_j}\sum_{g(i)=j}\widehat{Y}_i^{(s)} = \frac{1}{n_j}\sum_{g(i)=j}\int_{-\infty}^{T_j^n \vee 0}\left\{\frac{I_{[T_i > t]}}{1-\widehat{G}_j(t)} - I_{[t<0]}\right\}dt;$$

bringing the summation inside and omitting the range of the summation,

$$(2.10) \qquad \begin{aligned}(2.9) &= \int_{-\infty}^{T_j^n \vee 0}\left\{\frac{(1/n_j)\sum I_{[T_i \geq t]}}{1-\widehat{G}_j(t)} - I_{[t<0]}\right\}dt \\ &= \int_{-\infty}^{T_j^n \vee 0}\left\{\frac{1-\widehat{H}_j(t)}{1-\widehat{G}_j(t)} - I_{[t<0]}\right\}dt,\end{aligned}$$

where

$$1-\widehat{H}_j(t) = \frac{1}{n_j}\sum_{g(i)=j} I_{[T_i \geq t]}.$$

Using the facts that $(1-\widehat{F}_j)(1-\widehat{G}_j) = 1-\widehat{H}_j$ and that, for $t > T_j^n$, either $1-\widehat{F}_j = 0$ or $1-\widehat{G}_j = 0$, the right-hand side of (2.10) becomes

$$(2.11) \qquad \int_{-\infty}^{T_j^n \vee 0}\{1-\widehat{F}_j - I_{[t<0]}\}\,dt.$$

Notice that $1-\widehat{F}_j$ is always a proper distribution under the modification given above [i.e., $\widehat{F}(T_j^n) = 1$]. Integrating by parts, we find

$$(2.12) \qquad (2.11) = \int_{-\infty}^{T_j^n} t\, d\widehat{F}_j(t) = \sum_{t_i} t_i\, \Delta\widehat{F}_j(t_i),$$

where $\Delta\widehat{F}_j(t_i) = \widehat{F}_j(t_i+) - \widehat{F}_j(t_i-)$ and the t_i are the jump points of $\widehat{F}_j$. Finally, by noticing that $\widehat{F}_j$ jumps at T_i with jump size $1/n_j \times \delta_i/[1-\widehat{G}_j(T_i)]$, the right-hand side of (2.12) reduces to the CKSV estimator $(1/n_j)\Sigma_{g(i)=j}\widehat{Y}_i^{(c)}$. □

REMARK 2.2. If the C_i are i.i.d. random variables with distribution G_1 [$k = 1$ in assumption (1.2)], the CKSV differs from the KSV method only in its treatment of the largest T_i. This is analogous to the derivation of the Kaplan–Meier estimator for a survival function using the "redistribution-to-the-right" algorithm [Efron (1967)]. The KSV procedure puts zero "weight" on a censored observation and makes a larger uncensored observation carry a "weight" of $(1/n)1/[1-\widehat{G}(\cdot)]$ (where $1/[1-\widehat{G}(\cdot)]$ is the inflation factor). This suggests, to complete the analogy, that we define the largest Y_i as uncensored (even if it is censored), since there are no larger Y_i to which we can redistribute its "weight."

3. Artificial stratification.

It is a key assumption of the KSV procedure that the censoring variables do not depend on the covariates and do share a common distribution $G(t)$. Without this assumption, the estimator can be inconsistent, as was first confirmed by Miller and Halpern (1982).

Our stratification adjustment in Section 2 relaxes this stringent assumption, requiring the C_i to be i.i.d. only "locally" (within each stratum). What happens if the stringent i.i.d. censoring assumption is valid and we use stratification artificially? Will efficiency be sacrificed? When stratification is unnecessary, it should produce a poor estimator of G compared to the original KSV procedure. Our goal, however, is to derive the best estimator for β; G is a nuisance parameter and thus the quality of its estimator is less important.

Given that the censoring times are i.i.d. random variables and do not depend on the covariates, we will show that using artificial stratification with the KSV and the Leurgans methods results in more efficient estimators of β; that is, the resulting estimators have smaller asymptotic variances and MSE than the estimators derived without artificial stratification. For clarity, we present the technique in a simple linear model and use two artificial strata.

3.1. *Artificial stratification in the KSV method.*

In this section we consider the simple linear regression model with censoring times that are i.i.d. random variables. Throughout, the *asymptotic variance* refers to the variance of the asymptotic distribution and it is denoted by AsVar. We begin with two lemmas that provide expressions for the asymptotic variances in cases where we have one or two strata. They follow directly from (2.6) upon taking $k = 1$ or $k = 2$.

LEMMA 3.1. *Suppose that $k = 1$ in assumption (1.2) and that the conditions of Theorem 2.1 hold. Then the CKSV estimator is asymptotically normal with variance*

$$(3.1)\quad \operatorname{AsVar}(\widehat{\beta}^{(c)}) = \lim n \sum b_{ni}^2 \operatorname{Var}\left(\frac{\delta_i T_i}{1-G(T_i)}\right) - \lim n \int \frac{\left[\sum b_{nj} h_j(t)\right]^2}{\sum 1 - H_j(t)} \frac{dG}{1-G}.$$

REMARK 3.1. Koul, Susarla and Van Ryzin (1981) derived an expression for the asymptotic variance. However, their variance formulas (3.6), (3.7) and (3.8) are missing the factor n in the second (negative) term.

LEMMA 3.2. *Suppose that $k = 2$ in assumption* (1.2) [*i.e.*, $G_1(t) = G_2(t) = G(t)$] *and that the conditions of Theorem* 2.1 *hold. Then the CKSV estimator is asymptotically normal with variance*

$$\begin{aligned}\operatorname{AsVar}(\widehat{\beta}^{(c)}) &= \lim n \sum b_{ni}^2 \operatorname{Var}\Big(\frac{\delta_i T_i}{1-G(T_i)}\Big)\\ &\quad - \lim n \int \frac{\big[\sum_{g(j)=1} b_{nj} h_j(t)\big]^2}{\sum_{g(j)=1}(1-H_j(t))}\frac{dG}{1-G}\\ &\quad - \lim n \int \frac{\big[\sum_{g(j)=2} b_{nj} h_j(t)\big]^2}{\sum_{g(j)=2}(1-H_j(t))}\frac{dG}{1-G}.\end{aligned} \tag{3.2}$$

The next theorem presents one of two main results. It shows that, by choosing particular strata, the variance in (3.2) can be made smaller than the variance in (3.1).

THEOREM 3.1. *Suppose that the strata are such that $g(i) = 1$ if $b_{ni} \le 0$ and $g(i) = 2$ if $b_{ni} > 0$. Then the asymptotic variance in* (3.2) *is no greater than the asymptotic variance in* (3.1). *The inequality is strict when*

$$\lim \int_0^\infty \Big[\sum_{g(j)=1} b_{nj} h_j(t)\Big]^2 \Big[\sum_{g(j)=2} b_{nj} h_j(t)\Big]^2 \frac{dG}{1-G} > 0. \tag{3.3}$$

PROOF. In comparing (3.2) with (3.1), it is clear that we only need to show that the integrand of the negative term in (3.1) is smaller than its counterpart in (3.2), that is,

$$\begin{aligned}\lim n \frac{\big[\sum b_{nj} h_j(t)\big]^2}{\sum(1-H_j)} &\le \lim n \frac{\big[\sum_{g(i)=1} b_{ni} h_i(t)\big]^2}{\sum_{g(i)=1}(1-H_i)}\\ &\quad + \lim n \frac{\big[\sum_{g(i)=2} b_{ni} h_i(t)\big]^2}{\sum_{g(i)=2}(1-H_i)}.\end{aligned} \tag{3.4}$$

Let

$$A_1 = \lim \sum_{g(i)=1} b_{ni} h_i(t), \qquad A_2 = \lim \sum_{g(i)=2} b_{ni} h_i(t) \quad \text{and}$$

$$\lambda = \lim \frac{\sum_{g(i)=1}(1-H_i)}{\sum(1-H_j)}.$$

We can now apply the following lemma to complete the proof. Note that the strict inequality holds in (3.4) if A_1 and A_2 are both nonzero and have different signs, which is the case when (3.3) holds. □

LEMMA 3.3. *For any two constants* A_1 *and* A_2,

$$A^2 \le \frac{A_1^2}{\lambda} + \frac{A_2^2}{1-\lambda}, \quad \textit{where } A = A_1 + A_2,\ \lambda \in (0,1). \tag{3.5}$$

Generalization to more than two constants is immediate: for any constants A_i,

$$A^2 \le \frac{A_1^2}{\lambda_1} + \frac{A_2^2}{\lambda_2} + \cdots + \frac{A_k^2}{\lambda_k}, \quad \textit{where } A = A_1 + A_1 + \cdots + A_k,$$
$$\lambda_i \in (0,1)\ \textit{and}\ \sum \lambda_i = 1.$$

PROOF. We shall only prove (3.5). The inequality is trivially true if $A_1^2 = 0$ or $A_2^2 = 0$. Therefore, we consider the case in which $A_1^2 > 0$ and $A_2^2 > 0$.

Minimizing the right-hand side of (3.5) with respect to λ we find a unique minimum with the value $A_1^2 + 2\sqrt{A_1^2}\sqrt{A_2^2} + A_2^2$. This is certainly greater than or equal to $A_1^2 + 2A_1A_2 + A_2^2$, which is the left-hand side of (3.5). It is not hard to see that the strict inequality holds unless A_1 and A_2 are of the same sign and λ is equal to one and only one specific value. □

REMARK 3.2. Lemma 3.3 in fact implies that *any* stratification can reduce the asymptotic variance of the estimator. By repeatedly applying Lemma 3.3, we get the following upper bound for the right-hand side of (3.4) :

$$\sum_{i=1}^{n} \frac{[b_{ni}h_i(t)]^2}{1 - H_i(t)}, \tag{3.6}$$

that is, for any stratification, the integrand of the negative terms in (3.1) and (3.2) does not get larger than (3.6). Thus, we have the following lower bound for the asymptotic variance of any artificially stratified CKSV estimator of β:

$$\operatorname{AsVar}(\widehat{\beta}^{(c)}) \ge \lim n \sum b_{ni}^2 \operatorname{Var}\left(\frac{\delta_i T_i}{1 - G(T_i)}\right)$$
$$- \lim n \int \sum_{j=1}^{n} \frac{[b_{nj}h_j(t)]^2}{1 - H_j(t)} \frac{dG}{1-G}.$$

Using the artificial stratification of Theorem 3.1, the asymptotic variance of $\widehat{\alpha}^{(c)}$ can also be reduced; see Section 4 for numerical examples.

3.2. *Artificial stratification in the Leurgans method.* Results similar to the previous section also hold for Leurgans estimators. Assuming the conditions of Zhou [(1992), Theorem 3.1] plus the strata conditions (S1) and (S2), we can show that, with or without artificial stratification, both estimators are asymptotically normal with variances given in the next lemma.

LEMMA 3.4. *Suppose that* $k = 1$. *Then the asymptotic variance of the Leurgans estimator is*

$$\begin{aligned}\operatorname{AsVar}(\widehat{\beta}) &= \lim n \sum b_{ni}^2 \int \left[\frac{\tilde{h}_i(t)}{1-H_i}\right]^2 \{(1-G)\,dF_i + (1-F_i)\,dG\}\\ &\quad - \lim n \int \frac{\left(\sum b_{nj}\tilde{h}_j(t)\right)^2}{\sum_j (1-F_j)} \frac{dG}{(1-G)^2}.\end{aligned} \tag{3.7}$$

When $k = 2$ *(artificial strata),*

$$\begin{aligned}\operatorname{AsVar}(\widehat{\beta}^s) &= \lim n \sum b_{ni}^2 \int \left[\frac{\tilde{h}_i(t)}{1-H_i}\right]^2 \{(1-G)\,dF_i + (1-F_i)\,dG\}\\ &\quad - \lim n \int \left[\frac{\left[\sum_{g(j)=1} b_{nj}\tilde{h}_j(t)\right]^2}{\sum_{g(j)=1}(1-F_j)} + \frac{\left[\sum_{g(j)=2} b_{nj}\tilde{h}_j(t)\right]^2}{\sum_{g(j)=2}(1-F_j)}\right] \frac{dG}{1-G}.\end{aligned} \tag{3.8}$$

THEOREM 3.2. *Under the artificial stratification given in Theorem* 3.1, *the asymptotic variance in* (3.8) *is no greater than the asymptotic variance in* (3.7).

The proof of this theorem follows closely that of Theorem 3.1 and is therefore omitted.

Similar to Remark 3.2, by repeatedly using Lemma 3.3 on the integrand of the negative term in (3.7), we get the following lower bound for the asymptotic variance of any stratified Leurgans estimator:

$$\begin{aligned}\operatorname{AsVar}(\widehat{\beta}^{(s)}) &\geq \lim n \sum b_{ni}^2 \int \left[\frac{\tilde{h}_i(t)}{1-H_i}\right]^2 \{(1-G)\,dF_i + (1-F_i)\,dG\}\\ &\quad - \lim n \int \sum_{j=1}^{n} \frac{[b_{nj}\tilde{h}_j(t)]^2}{1-F_j} \frac{dG}{1-G}.\end{aligned}$$

The conclusion that artificial stratification can provide a smaller asymptotic variance of the estimator is intriguing and somewhat counterintuitive. The results for the two strata suggest that one should consider the use of artificial stratification at least for large samples when the censoring times are i.i.d. random variables. For multiple regression, however, the implementation of artificial stratification is more complicated and demands further research. One

TABLE 1

(a) $\varepsilon_i \sim$ N(0, 0.5^2), C$_i \sim$ U(0, 4); sample size 100; censoring percentage 50.68%

Method	E$\widehat{\alpha}$	Var$\widehat{\alpha}$	E$\widehat{\beta}$	Var($\widehat{\beta}$)	MSE$_{\text{ksv}}$($\widehat{\beta}$)/MSE($\widehat{\beta}$)
KSV	1.6364	0.0661	0.5483	0.1076	1
CKSV	1.9688	0.0093	0.9637	0.0669	4.56
ACKSV	1.9696	0.0069	0.9639	0.0498	6.098
Leurgans	1.9688	0.0095	0.9603	0.0217	13.37
ALeurgans	1.9696	0.0070	0.9638	0.0141	20.19

(b) $\varepsilon_i \sim$ N(0, 2.1^2), C$_i \sim$ U(0, 4); sample size 100; censoring percentage 50.03%

Method	E$\widehat{\alpha}$	Var($\widehat{\alpha}$)	E$\widehat{\beta}$	Var($\widehat{\beta}$)	MSE$_{\text{ksv}}$($\widehat{\beta}$)/MSE($\widehat{\beta}$)
KSV	0.8729	0.1297	0.3106	0.1058	1
CKSV	1.6930	0.0403	0.8001	0.5038	1.068
ACKSV	1.6712	0.0389	0.7914	0.1476	3.040
Leurgans	1.6930	0.0402	0.7976	0.0492	6.44
ALeurgans	1.6712	0.0390	0.7916	0.0345	7.455

(c) $\varepsilon_i \sim$ N(0, 0.5^2), C$_i \sim$ U(−4, 8); sample size 100; censoring percentage 49.87%

Method	E$\widehat{\alpha}$	Var($\widehat{\alpha}$)	E$\widehat{\beta}$	Var($\widehat{\beta}$)	MSE$_{\text{ksv}}$($\widehat{\beta}$)/MSE($\widehat{\beta}$)
KSV	1.9902	0.0154	0.9780	0.0558	1
CKSV	1.9995	0.0140	0.9908	0.0530	1.060
ACKSV	1.9937	0.0067	0.98129	0.0290	1.917
Leurgans	1.9993	0.0137	1.0025	0.0841	0.669
ALeurgans	1.9937	0.0067	0.9824	0.0288	1.933

(d) $\varepsilon_i \sim$ N(0, 0.5^2), C$_i \sim$ U(−4, 8); sample size 100; censoring percentage 50.63%

Method	E$\widehat{\alpha}$	Var($\widehat{\alpha}$)	E$\widehat{\beta}$	Var($\widehat{\beta}$)	MSE$_{\text{ksv}}$($\widehat{\beta}$)/MSE($\widehat{\beta}$)
KSV	1.8306	0.1230	0.8627	0.2408	1
CKSV	1.9799	0.0741	0.9777	0.2453	1.0564
ACKSV	1.9654	0.0670	1.0033	0.1158	2.242
Leurgans	1.9795	0.0744	0.9955	0.1664	1.560
ALeurgans	1.9655	0.0671	0.9966	0.0834	3.112

possible approach would be to use a grid partitioning of the vector of covariates. A reduction in the asymptotic variance seems likely.

4. Monte Carlo simulations. To demonstrate the reduction in the variance and MSE by the artificial stratification of Theorems 3.1 and 3.2, we present simulations that incorporate 200 samples of size 100 each. The data were generated according to the model

$$Y_i = 2 + X_i + \varepsilon_i, \quad \text{with} \quad X_i = -2 + 0.04i \quad \text{and} \quad \varepsilon_i \sim N(0, \sigma^2).$$

For the model $T_i = \min(Y_i, C_i)$, the C_i are drawn from a Uniform(a, b) distribution.

In Table 1, ACKSV denotes the CKSV method under artificial stratification, and ALeurgans denotes the method of Leurgans under artificial stratification (of Theorems 3.1 and 3.2).

The simulations presented here are a representative selection of the simulations we conducted. We summarize the results as follows:

1. The KSV estimator is consistently worse than the others. It is extremely biased and its mean squared error is usually largest.
2. Under i.i.d. censoring times, the CKSV and the Leurgans methods are comparible. Both methods estimate the slope with relatively small bias and low variance. The estimators of the intercept were more biased than those of the slope, which is consistent with other methods of estimation [see Buckley and James (1979)].
3. In most cases, artificial stratification reduces the variance of the CKSV and the Leurgans estimators by up to 50% without significant change in bias. This improvement is obtained when estimating either the slope or the constant term of the regression model.
4. The behavior of the estimators was relatively unaffected when nonnormal errors were simulated.

4.1. *Closing remarks.* It is our experience, and it is well documented [e.g., Miller and Halpern (1982)], that the unstratified synthetic methods tend to produce extremely biased estimators in the following cases:

1. when the censoring distribution $G(\cdot)$ varies with the covariates;
2. when the censored observations are spread unevenly over the range of the covariates.

Condition (1.2) relaxes the often unrealistic assumption of i.i.d. censoring times, making the methods we present less sensitive to the two departures listed above. Even when the true strata are not known, we recommend stratifying the censoring times with respect to the covariates so that within each stratum cases 1 and 2 are less likely to occur.

The need for stratification can usually be detected by plotting the log-survival time versus the various covariates. Nonuniformity in the proportion of the censored observations over the range of a covariate calls for stratification and also indicates a possible set of strata. As is the case with other methods of data analysis, one cannot provide exact rules for how to stratify. However, we have found that when working with real data that needed stratification (e.g., Stanford heart transplant data) any "sensible" stratification used with the CKSV method provided better estimators than no stratification.

A "sensible" stratification can be obtained in several ways. Cluster analysis techniques [Kaufman and Rousseeauw (1990)] applied to the log-censoring times and the covariates may indicate a set of strata. Also there may exist scientific evidence, interest and/or information about how the data was collected that favors a particular stratification. When no such preference exists, another possibility is to stratify the censoring times in as many groups as possible. Upon stratification, one can derive the Kaplan–Meier estimators $\widehat{G}_i$ and calculate the proportion of censored observations p_i for each stratum. A comparison of the $\widehat{G}_i$ and p_i of different strata may indicate which strata can be combined.

The possibility of using different stratifications for the same data set raises questions about the sensitivity of the proposed methods to the particular stratification. Would an incorrect stratification be worse than no stratification at all when stratification is present but not identified? This and other questions of robustness call for further research.

Acknowledgments. We thank Jeff Albert and Shula Gross for assisting us with the computer runs. We are also very grateful to the Editor and an Associate Editor for many constructive suggestions.

REFERENCES

AMEMIYA, A. (1985). *Advanced Econometrics.* Harvard Univ. Press.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

EFRON, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853. Univ. California Press, Berkeley.

GILL, R. (1980). *Censoring and Stochastic Integrals. Math. Centre Tract* **124**. Math. Centrum, Amsterdam.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

KOUL, H., SUSARLA, V. and VAN RYZIN, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9** 1276–1288.

LEURGANS, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74** 301–309.

MADDALA, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge Univ. Press.

MILLER, R. G. (1976). Least squares regression with censored data. *Biometrika* **63** 449–464.

MILLER, R. G. (1981). *Survival Analysis.* Wiley, New York.

MILLER, R. G. and HALPERN, J. (1982). Regression with censored data. *Biometrika* **69** 521–531.

SRINIVASAN, C. and ZHOU, M. (1991). Linear regression with censoring. Technical Report 305, Dept. Statistics, Univ. Kentucky.

ZHOU, M. (1992). Asymptotic normality of the "synthetic data" regression estimator for censored survival data. *Ann. Statist.* **20** 1002–1021.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY 08903

DEPARTMENT OF STATISTICS
UNIVERSITY OF KENTUCKY
LEXINGTON, KENTUCKY 40506-0027