ESTIMATING THE DISTRIBUTION FUNCTION OF A FINITE POPULATION: ESTIMATORS OF CHAMBERS-DUNSTAN TYPE BASED ON LINEAR SUPERPOPULATION MODELS WITH UNKNOWN HETEROSCEDASTIC ERRORS

M. J. Lombardía, W. González-Manteiga
and J. M. Prada-Sánchez

Report 03-11

Reports in Statistics and Operations Research
Estimating the distribution function of a finite population:

Estimators of Chambers-Dunstan type based on linear superpopulation models with unknown heteroscedastic errors

Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M.
Department of Statistics and Operative Research.
Faculty of Mathematics. Santiago de Compostela University.
15782 Santiago de Compostela, Spain.
mjoselc@usc.es.

Abstract

We consider a finite population $P = \{(Y_k, x_k)\}$ conforming to a linear superpopulation model with unknown heteroscedastic errors, the variance of which is a smooth enough function of the auxiliary variable $x$ for their nonparametric estimation. In this article we describe a method of Chambers-Dunstan type for estimation of the distribution of $Y$ from a sample drawn from $P$ without replacement, and determine the asymptotic distribution of its estimation error. For estimation of its error variance and other characteristics in any particular case we define a bootstrap method that combines the wild bootstrap with nonparametric estimation of the superpopulation regression error variance and conforms to Presnell and Booth’s (1994) paradigm for bootstrapping finite populations, and we prove the model-consistency of its resampling scheme. In simulation studies and in an application to real data the new distribution function estimators proved to be significantly superior to both the empirical and homoscedastic Chambers-Dunstan estimators, and the bootstrap estimation was also satisfactory.
1 Introduction

Given a finite population \( P = \{(Y_k, x_k)\} (k = \{1, \ldots, N\}) \), where the \( x_k \) are known for all members of the population and the \( Y_k \) only for a sample \( S \subset P \), Chambers and Dunstan (1986) estimated the distribution function \( F_N \) of \( Y \) by treating \( P \) as a realization of a superpopulation conforming to a linear regression model of the form \( Y = \beta x + \varepsilon \), where the mean of the random error \( \varepsilon \) is zero and its variance \( \sigma^2(x) \) is some known function of \( x \). The distribution of the error of this estimator was shown to be asymptotically normal. For homoscedastic \( \varepsilon \), the asymptotic variance of the corresponding estimator for the model \( Y = \alpha + \beta x + \varepsilon \) was calculated by Chambers et al. (1992), but no analogous results have been forthcoming for the more realistic heteroscedastic case.

In Section 2 below we define, and prove asymptotic properties of, two estimators of Chambers-Dunstan type that do not require \( \sigma^2(x) \) to be known. One of them, \( \hat{F} \), consistently treats \( \varepsilon \) as heteroscedastic. The other, \( \hat{F}_L \), which is included for the purposes of comparison, may be considered as intermediate between \( \hat{F} \) and the estimator considered by Chambers et al. (1992): it estimates \( \alpha \) and \( \beta \) by unweighted least squares, but normalizes the corresponding estimated errors by an estimate of the function \( \sigma(x) \) when using these errors to estimate the distribution function itself. In both cases, \( \sigma^2(x) \) is estimated nonparametrically (Carroll 1982).

The estimation errors of both the new estimators prove to have asymptotically normal distributions, but in both cases the analytic expression for the variance of the asymptotic distribution - a generalization of that obtained for known constant \( \sigma^2(x) \) by Chambers et al. (1992) - is cumbersome and hard to apply in practice. We have accordingly developed a bootstrap method that in any particular application allows
estimation of the estimation error variance. This method, which combines the wild bootstrap (Shao and Tu, 1995, p.292) with nonparametric estimation of the superpopulation regression model error variance (Carroll 1982) and Presnell and Booth’s (1994) approach to bootstrapping finite populations, involves a two-stages resampling scheme: in the first stage, a set of bootstrap populations \( P^* = \{ (Y_k^*, x_k) \} \) are generated using the superpopulation regression model constructed from the given sample \( S \); and in the second, a bootstrap sample \( S^* \) is taken without replacement from each \( P^* \). The method is described in greater detail in Section 3, where its consistency within the superpopulation model is also proved. In Section 4 we report the behaviour of the new estimators, and of the bootstrap of Section 3, in a simulation study in which data were generated from three superpopulation regression models with varying degrees of heteroscedasticity, and in a study using real data previously used by Chambers and Dunstan (1986). Finally, in Section 5, we summarize our conclusions, while proofs of the results of Sections 2 and 3 are given in an Appendix.

2 Chambers-Dunstan estimators for superpopulation regression models with unknown error variance

Let \( P \) be the set of integers \( \{1, ..., N\} \), \( S \) a randomly selected \( n \) element subset of \( P \), and \( P - S \) the complement of \( S \) in \( P \). We consider a finite population \( P = \{ (Y_k, x_k) \}_{k \in P} \), where the values \( x_k \) of the auxiliary variable \( X \) are known for all the members of \( P \) and the variable \( Y \) is related to \( X \) by the model

\[
\xi: Y(x_k) = \alpha + \beta x_k + \sigma(x_k) \varepsilon_k,
\]

in which the \( \varepsilon_k \) are independent identically distributed random variables of zero mean, unit variance but unknown distribution function, and the model parameters (\( \alpha \) and \( \beta \)) and error variances \( \sigma^2(x_k) \) are unknown quantities to be estimated from the data of the random sample \( S = \{ (Y_i, x_i) \}_{i \in S} \) without replacement form \( P \). Our objective is to estimate the distribution function \( F_N(t) = N^{-1} \sum_{k \in P} I(Y_k \leq t) \) of \( Y \), which depends on \( P \), then it is random under \( \xi \).
Since $F_N(t)$ can be partitioned in the form

$$F_N(t) = (n/N)F_n(t) + (1 - n/N)F_r(t) \quad (2)$$

where $F_n(t) = n^{-1} \sum_{i \in S} I(Y_i \leq t)$ is known and $F_r(t) = (N - n)^{-1} \sum_{j \in P - S} I(Y_j \leq t)$, our task reduces to the construction of an estimate $\hat{F}_r(t)$ of $F_r(t)$. To perform it we note that

$$E \left\{ \sum_{j \in P - S} I(Y_j \leq t) \right\} = E \left\{ \sum_{j \in P - S} I(\eta_j \leq t - \alpha - \beta x_j) \right\}$$

$$= \sum_{j \in P - S} P \left\{ \frac{\eta_j}{\sigma(x_j)} \leq \frac{t - \alpha - \beta x_j}{\sigma(x_j)} \right\}$$

$$= \sum_{j \in P - S} G \left( \frac{t - \alpha - \beta x_j}{\sigma(x_j)} \right),$$

where $G$ is the distribution function of the normalized errors $\varepsilon_j$, and that to estimate $F_r(t)$ it is therefore natural to approximate $I(Y_j \leq t)$ by an estimate of $G[(t - \alpha - \beta x_j)/\sigma(x_j)]$:

$$\hat{F}_r(t) = (N - n)^{-1} \sum_{j \in P - S} \hat{G} \left( \frac{t - \alpha - \beta x_j}{\sigma_h(x_j)} \right) \quad (3)$$

where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}_h^2(x)$ are estimates of the regression parameters and error variances of $\xi$ and $\hat{G}$ is an estimate of $G$.

To estimate the error variances of $\xi$ from the sample data we follow Carroll (1982):

$$\hat{\sigma}_h^2(x) = \frac{\sum_{i \in S} k \{(x - x_i)/h\} \left( Y_i - \hat{\alpha}_L - \hat{\beta}_L x_i \right)^2}{\sum_{i \in S} k \{(x - x_i)/h\}}, \quad (4)$$

where $k$ is a symmetric density function, $h$ is a bandwidth parameter, and $\hat{\alpha}_L$ and $\hat{\beta}_L$ are the unweighted least-squares estimates of $\alpha$ and $\beta$ obtained from the sample, i.e. the values of $a$ and $b$ that minimize $\sum_{i \in S}(Y_i - a - bx_i)^2$.

In the construction of $\hat{F}$ (see Introduction), the estimates $\hat{\alpha}$ and $\hat{\beta}$ in eq.3 are those obtained by the weighted least-squares method with weights $w_i = 1/\hat{\sigma}_h^2(x_i)$, $w_i = 1/\hat{\sigma}_h^2(x_i),

$$\hat{\alpha}_E = \bar{y}_w - \hat{\beta}_E \bar{x}_w, \quad (5)$$

$$\hat{\beta}_E = \frac{\sum_{i \in S} w_i Y_i x_i - \bar{y}_w \bar{x}_w}{\sum_{i \in S} w_i (x_i - \bar{x}_w)^2}$$
(where \( \bar{y}_w = (\sum_{i \in S} w_i)^{-1} \sum_{i \in S} w_i y_i \) and \( \bar{x}_w = (\sum_{i \in S} w_i)^{-1} \sum_{i \in S} w_i x_i \)), and the estimate of \( G \) is \( \hat{G}(u) = n^{-1} \sum_{i \in S} I \left( (r_i/\hat{\sigma}_h(x_i)) \leq u \right) \) (where \( r_i = Y_i - \hat{\alpha}_E - \hat{\beta}_E x_i \)). Thus

\[
\hat{F}(t) = N^{-1} \left\{ nF_n(t) + (N-n)\hat{F}_n(t) \right\} \\
= N^{-1} \left\{ \sum_{i \in S} I(Y_i \leq t) + \sum_{j \in P-S} \hat{G} \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E x_j}{\hat{\sigma}_h(x_j)} \right) \right\}.
\]

For \( \hat{F}_L \) (see Introduction), we use the unweighted least-squares estimates \( \hat{\alpha}_L \) and \( \hat{\beta}_L \) in eq.3 together with the corresponding estimate \( \hat{G}_L(u) = n^{-1} \sum_{i \in S} I[(\hat{r}_i/\hat{\sigma}_h(x_i)) \leq u] \) (where \( \hat{r}_i = Y_i - \hat{\alpha}_L - \hat{\beta}_L x_i \)), so that

\[
\hat{F}_L(t) = N^{-1} \left\{ \sum_{i \in S} I(Y_i \leq t) + \sum_{j \in P-S} \hat{G}_L \left( \frac{t - \hat{\alpha}_L - \hat{\beta}_L x_j}{\hat{\sigma}_h(x_j)} \right) \right\}.
\]

2.1 Asymptotic behaviour of \( \hat{F} \) and \( \hat{F}_L \)

2.1.1 Notation

We denote by \( a \wedge b \) the lesser of numbers \( a \) and \( b \); by \( P, E, Var \) and \( MSE \) the probability, expectation, variance and mean square error operators under \( \xi \); by \( g(= G') \) the density of the normalized errors \( \varepsilon_k \) of eq.1; and by \( d \) the density of the design set \( \{x_k\} \), which is assumed to have compact support \( \Gamma \), asymptotic mean \( \mu_x = \int x d(x) dx \) and asymptotic variance \( \tau_x^2 = \int (x - \mu_x)^2 d(x) dx \). The asymptotic weighted mean and variance of design points are defined by \( \mu_{wx} = W^{-1} \int x \sigma^{-2}(x) d(x) dx \) and \( \tau_{wx}^2 = W^{-1} \int (x - \mu_{wx})^2 \sigma^{-2}(x) d(x) dx \), respectively, where \( W = \int [\sigma^2(x)]^{-1} d(x) dx \). The following abbreviations are used for certain integrals:

\[
I_1 = \int \int \left( G \left( \frac{t - \alpha - \beta u}{\sigma(u)} \right) - G \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \right) G \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) d(u) d(v) dudv,
\]
\[
I_2 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{1}{\sigma(u)} - \frac{1}{\sigma(v)} \right) d(u) d(v) dudv,
\]
\[
I_3 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{u}{\sigma(u)} - \frac{v}{\sigma(v)} \right) d(u) d(v) dudv,
\]
\[
I_4 = \int_{-\infty}^{\infty} zg(z) d(u) dudz,
\]
\[
I_5 = \int \left[ G \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) - G \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right)^2 \right] d(v) dv.
\]
2.1.2 Assumptions

H1. i) The first two derivatives of \( d \) are continuous.

\[ n^{-1} \sum_{i \in S} I(x_i \leq x) \rightarrow \int_{-\infty}^{x} d(y)dy. \]

ii) \((N-n)^{-1} \sum_{j \in P-S} I(x_j \leq x) \rightarrow \int_{-\infty}^{x} d(y)dy.\]

iii) \((N-n) \sum_{j \in P} I(x_j \leq x) \rightarrow \int_{-\infty}^{x} d(y)dy.\)

H2. \( \sum_{k \in P} x_k^2 \) and \( \sum_{k \in P} x_k^2 / \sigma^2(x_k) \) are strictly positive and bounded for all \( N \).

H3. The first two derivatives of the variance function \( \sigma^2(x_k) \) are continuous in \( \Gamma \), and \( \sigma^2(x_k) \) has a positive infimum: \( 0 < \inf_{\{\Gamma\}} \sigma^2(x_k) < +\infty. \)

H4. Both \( \max_{\{i \in P, j \in P-S\}} \left| \frac{1}{\sigma(x_i)} - \frac{1}{\sigma(x_j)} \right| = O(n^\delta) \)

and \( \max_{\{i \in P, j \in P-S\}} \left| \frac{x_i}{\sigma(x_i)} - \frac{x_j}{\sigma(x_j)} \right| = O(n^\delta) \)

where \( \delta \in (0,1/2). \)

H5. \( \exists \delta > 0 \) such that \( \mathbb{E} \{|\epsilon_i|^{\delta+2}\} < \infty. \)

H6. As \( N \) tends to infinity, \( n \) tends to infinity in such a way that the sampling fraction \( f = n/N \) tends to a limit that, without risk of confusion, will also be denoted by \( f \), and \( 0 < f < 1. \)

K1. In eq.4, the kernel \( k \) has compact support and three continuous derivatives.

K2. As \( n \to \infty \), the bandwidth parameter \( h \) is taken such that \( nh^8 \to 0 \) but \( nh^2 \to \infty. \)

Remark 2.1. Since the limit of \( f \) as \( n \to \infty \) is nonzero (H6), the use of the without-replacement sample \( S \) is not equivalent in the limit to sampling with replacement.

Remark 2.2. K2 is weaker than the corresponding assumption made by Carroll (1982). Carroll required that \( nh^4 \to 0 \) and \( n^{5/4} h^4 \to \infty \) in order to be able to estimate the regression parameters of the model \( Y_i = \alpha + \beta X_i + \sigma(X_i) \epsilon_i \), where \( \{X_i\} \) and \( \{\epsilon_i\} \) are mutually independent sets of independent identically distributed random variables, a situation in which both the numerator and the denominator of the right-hand side of eq.4 are random variables. In our case the \( x_i \) are fixed and in eq.4 only the numerator is random, which facilitates proof of the results obtained by Carroll (1982).
2.1.3 Theorems

Theorem 1. For $\tilde{F}(t)$ as defined in eq.6, and assuming H1-H6, K1 and K2,

$$\sqrt{n} \left\{ \frac{\tilde{F}(t) - F_N(t)}{n} \right\} \to_d N(0, V),$$

where

$$V = \left\{ (1 - f)^2 \left[ I_1 + I_2^2 W^{-1} (1 + \tau_{wx}^2 \mu_{wx}) + \tau_{wx}^2 W^{-1} \tau_{wx}^2 - I_2 I_3 W^{-1} \tau_{wx}^2 \mu_{wx} \right. \right.$$  
$$+ I_2 I_4 \int (1 - \tau_{wx}^2 \mu_{wx} (x - \mu_{wx})) \sigma^{-1} (x) d(x) dx + I_3 I_4 \tau_{wx}^2 \int (x - \mu_{wx}) \sigma^{-1} (x) d(x) dx \left. \right\}$$

Proof. See the Appendix.

Theorem 2. For $\hat{F}_L(t)$ as defined in eq.7, and assuming H1-H6, K1 and K2,

$$\sqrt{n} \left\{ \frac{\hat{F}_L(t) - F_N(t)}{n} \right\} \to_d N(0, V_L),$$

where

$$V_L = \left\{ (1 - f)^2 \left[ I_1 + I_2^2 \int (1 - \tau_{wx}^2 \mu_{wx} (x - \mu_{wx}))^2 \sigma^2 (x) d(x) dx + I_2^2 \int (\tau_{wx}^2)^2 (x - \mu_{wx})^2 \sigma^2 (x) d(x) dx \right. \right.$$  
$$+ I_2 I_3 \int (1 - \tau_{wx}^2 \mu_{wx} (x - \mu_{wx})) \tau_{wx}^2 (x - \mu_{wx}) \sigma^2 (x) d(x) dx \left. \right\}$$

Proof. Analogous to that of Theorem 1.

Remark 2.3. $\text{Var} \{ \tilde{F}(t) - F_N(t) \}$ and $\text{Var} \{ \hat{F}_L(t) - F_N(t) \}$ are both $O(n^{-1})$. They differ asymptotically in the terms involving the variances and covariances of their estimates of $\alpha$ and $\beta$.

Remark 2.4. If $\sigma(x_L)$ is constant in eq.1, then both $V$ (eq.9) and $V_L$ (eq.11) coincide with the expression obtained by Chambers et al. (1992), because under this assumption $v_L = v_E, \hat{v}_L = \hat{v}_E, I_2 = 0$, and $\int (x - \mu_{wx}) \sigma^{-1} (x) d(x) dx = \int (x - \mu_{wx}) \sigma (x) d(x) dx = 0.$
3 The bootstrap method

In general, given a statistic \( \theta \) that is a functional of an unknown distribution function \( F \), a bootstrap estimate \( \hat{\theta}^* \) of \( \theta \) is obtained by applying the same functional to the empirical distribution \( F_n \) constructed from a sample \( S \), and properties of \( \hat{\theta}^* \) such as its bias and variance are estimated from a large set of values \( \hat{\theta}^{*b} \) that are similarly obtained from a set of bootstrap distributions \( F^{*b}_n \) generated by repeated resampling of \( F_n \). However, when the population \( P \) is finite and the sample \( S \) is obtained without replacement, the most natural approach is to define a bootstrap estimate of \( \theta \) as the average of a set of “raw” bootstrap values \( \hat{\theta}^{*b} \), each of which is derived from a bootstrap population \( P^{*b} = \{(Y^*_k, x_k)\}_{k \in \mathcal{P}} \) that is generated from \( S \).

\[
\begin{array}{ccc}
\text{Real World} & & \text{Bootstrap World} \\
\xi & & \xi^* \\
\downarrow \text{id} & & \downarrow \text{id} \\
\downarrow \text{SR} & & \downarrow \text{SR} \\
\downarrow & & \\
\hat{\theta} & & \hat{\theta}^*
\end{array}
\]

In the present context, moreover, it is natural for each \( \hat{\theta}^{*b} \) to be constructed from a sample \( S^{*b} \) of size \( n \) obtained without replacement from \( P^{*b} \), and for the \( P^{*b} \) to be generated using a superpopulation model in which the regression parameters are the estimates obtained in Section 2 and the heteroscedasticity is dealt with by a wild bootstrap incorporating a nonparametric estimate of the variance of the errors:

\[
\xi^*: Y^*_k = \alpha^* + \beta^* x_k + \hat{\sigma}_{\theta}(x_k) Z_k, \quad (12)
\]

where \((\alpha^*, \beta^*) = (\hat{\alpha}_L, \hat{\beta}_L) \) or \((\hat{\alpha}_E, \hat{\beta}_E)\), \( \hat{\sigma}_{\theta}(x_k) \) is obtained via eq.4 using a “pilot” bandwidth parameter \( h_0 \), and the \( Z_k \) are independent zero-mean, unit-variance random variables with a distribution \( G^* \) that estimates \( G \). In order for \( G^* \) to be sufficiently smooth and accurate, in this work we use a smoothed
version of the empirical maximum likelihood estimator (Owen 2001):

\[ G^*(u) = \sum_{i \in S} \hat{p}_i L \left\{ \left( u - (\hat{\epsilon}_i - \bar{\epsilon})/h_1 \right) \right\}, \tag{13} \]

where \( \hat{\epsilon}_i = r_i/\hat{\sigma}_h(x_i) \) or \( \tilde{r}_i/\hat{\sigma}_h(x_i) \), \( \bar{\epsilon} \) is the average of the \( \hat{\epsilon}_i \) for \( i \in S \), \( L \) is the distribution function of a bounded symmetric kernel density, \( h_1 > 0 \) is a bandwidth parameter, and the \( \hat{p}_i \) \( i \in S \) \) are the empirical maximum likelihood estimates of the probability measure corresponding to \( G \). These are obtained subject to the constraints \( \int dG(v) = 1, \int v dG(v) = 0 \) and \( \int v^2 dG(v) = 1 \) using Lagrange multipliers, with the result that

\[ \hat{p}_i = \frac{1}{n \left[ 1 + \delta_n(\hat{\epsilon}_i - \bar{\epsilon}) + \lambda_n ((\hat{\epsilon}_i - \bar{\epsilon})^2 - 1) \right]}, \tag{14} \]

where the Lagrange multipliers \( \delta_n \) and \( \lambda_n \) satisfy

\[
\begin{align*}
\sum_{i \in S} \frac{(\hat{\epsilon}_i - \bar{\epsilon})}{n \left[ 1 + \delta_n(\hat{\epsilon}_i - \bar{\epsilon}) + \lambda_n ((\hat{\epsilon}_i - \bar{\epsilon})^2 - 1) \right]} &= 0, \\
\sum_{i \in S} \frac{(\hat{\epsilon}_i - \bar{\epsilon})^2 - 1}{n \left[ 1 + \delta_n(\hat{\epsilon}_i - \bar{\epsilon}) + \lambda_n ((\hat{\epsilon}_i - \bar{\epsilon})^2 - 1) \right]} &= 0. \tag{15}
\end{align*}
\]

In the remainder of this Section it is assumed that \( (\alpha^*, \beta^*) = (\hat{\alpha}_E, \hat{\beta}_E) \) in eq.12, that \( \hat{\epsilon}_i = r_i/\hat{\sigma}_h(x_i) \) in eq.13, and that all bootstrap estimates are calculated using procedures paralleling those leading to \( \hat{F} \) in Section 2 (procedures and results for the bootstrap of \( \hat{F}_L \) are analogous). Thus, given a bootstrap population \( P^* \) generated as above, and an \( n \)-member sample \( S^* \subset P^* \) defined by a random subset \( S^* \) of \( P \), the “raw” bootstrap Chambers-Dunstan estimator corresponding to \( F_N \), i.e. the estimator, in the bootstrap world, of the distribution function of \( Y^* \)

\[
F_N^*(t) = N^{-1} \sum_{k \in P} I(Y_k^* \leq t) = N^{-1} \left\{ \sum_{i \in S^*} I(Y_i^* \leq t) + \sum_{j \in P - S^*} I(Y_j^* \leq t) \right\}, \tag{16}
\]

is

\[
\hat{F}^*(t) = N^{-1} \left\{ \sum_{i \in S^*} I(Y_i^* \leq t) + \sum_{j \in P - S^*} \hat{G}^* \left( \frac{t - \hat{\alpha}_E^* - \hat{\beta}_E^* x_j}{\hat{\sigma}_h^*(x_j)} \right) \right\}, \tag{17}
\]

where, paralleling the constructions of Section 2,

\[
\hat{\sigma}_h^2(x) = \frac{\sum_{i \in S^*} k \{(x - x_i)/h\} \{Y_i^* - \hat{\alpha}_L^* - \hat{\beta}_L^* x_i\}^2}{\sum_{i \in S^*} k \{(x - x_i)/h\}}, \tag{18}
\]
(\hat{\alpha}_E^* \text{ and } \hat{\beta}_E^* \text{ being the } S^*\text{-based ordinary (unweighted) least-squares estimates of } (\alpha^*, \beta^*) \text{ in eq.12 when } (\alpha^* = \hat{\alpha}_E \text{ and } \beta^* = \hat{\beta}_E);)

ii)

\[
\hat{\alpha}_E^* = \hat{y}_{w^*} - \hat{\beta}_E^* \hat{x}_{w^*},
\]

\[
\hat{\beta}_E^* = \frac{\sum_{i \in S^*} w_i^* Y_i^* x_i - \hat{y}_{w^*} \hat{x}_{w^*}}{\sum_{i \in S^*} w_i^* (x_i - \hat{x}_{w^*})^2}
\]

(with \( w_i^* = 1/\hat{\sigma}_{h_i}^2(x_i) \) and \( \hat{y}_{w^*} = \left[ \sum_{i \in S^*} w_i^* \right]^{-1} \sum_{i \in S^*} w_i^* Y_i^* \), \( \hat{x}_{w^*} = \left[ \sum_{i \in S^*} w_i^* \right]^{-1} \sum_{i \in S^*} w_i^* x_i \); and

iii) \( \hat{G}^*(u) = n^{-1} \sum_{i \in S^*} I\left( \frac{x_i}{\hat{\sigma}_{h_i}(x_i)} \leq u \right) \) (with \( r_i^* = Y_i^* - \hat{\alpha}_E^* - \hat{\beta}_E^* x_i \)).

### 3.1 Model-consistency of the resampling scheme

#### 3.1.1 Notation

We denote by \( P \{ \cdot \}, E \{ \cdot \}, Var \{ \cdot \} \text{ and } MSE \{ \cdot \} \) the probability, expectation, variance and mean square error operators under \( \xi^* \); by \( g^* (= G^*) \) the density of the normalized errors \( Z_k \) of eq.12; and by \( f^{(\nu)} \) the \( \nu \)th derivative of any function \( f \). The asymptotic weighted mean and variance of the design set are defined by

\[
\hat{\mu}_{ho} = \hat{W}_{ho}^{-1} \int x \hat{\sigma}_{ho}^{-2}(x)d(x)dx
\]

and

\[
\hat{\tau}_{ho}^2 = \hat{W}_{ho}^{-1} \int (x - \hat{\mu}_{ho})^2 \hat{\sigma}_{ho}^{-2}(x)d(x)dx,
\]

respectively, where \( \hat{W}_{ho} = \int [\hat{\sigma}_{ho}^2(x)]^{-1}d(x)dx \). The following abbreviations are used for certain integrals:

\[
I_1^* = \int \int \left[ G^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E u}{\hat{\sigma}_{ho}(u)} \right) - G^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E u}{\hat{\sigma}_{ho}(v)} \right) \right] d(u)d(v)dv,
\]

\[
I_2^* = \int \int g^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E u}{\hat{\sigma}_{ho}(v)} \right) \frac{1}{\hat{\sigma}_{ho}(u)} d(u)d(v)dv,
\]

\[
I_3^* = \int \int g^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E u}{\hat{\sigma}_{ho}(v)} \right) \frac{u}{\hat{\sigma}_{ho}(u)} - \frac{v}{\hat{\sigma}_{ho}(v)} d(u)d(v)dv,
\]

\[
I_4^* = \int_{-\infty}^{-\hat{\alpha}_E - \hat{\beta}_E u} g^* (z) d(u)duz,
\]

\[
I_5^* = \int \left[ G^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E u}{\hat{\sigma}_{ho}(v)} \right) - G^* \left( \frac{t - \hat{\alpha}_E - \hat{\beta}_E v}{\hat{\sigma}_{ho}(v)} \right) \right] d(v)dv.
\]
3.1.2 Hipótesis

B1. The density function $g^*$ of the independent, identically distributed random variables $Z_k \ (k \in P)$ has a bounded derivative in $\Gamma$ and is such that $E \left[ |Z_i|^{\delta+2} \right] < \infty$ for some $\delta > 0$.

B2. As $n \to \infty$, the pilot window parameter $h_0$ is taken such that $h_0 \to 0$ but $nh_0^5$ tends to infinity.

Remark 3.1. B2 is required for $\hat{\sigma}_{h_0}^{(\nu)}(x)$ to converge in probability to $\sigma^{(\nu)}(x) \ (\nu = 0, 1, 2)$; see Härdle (1990, p.33).

3.1.3 Results

Carroll (1982) showed that $\hat{\beta}_E$ tends in probability to $\hat{\beta}_T$, the weighted least-squares estimate when the variances $\sigma^2(x_i)$ are known ($\hat{\beta}_T = \left[ \sum_{i \in S} w_i Y_i x_i - \hat{y}_w \hat{x}_w \right] / \left[ \sum_{i \in S} w_i (x_i - \hat{x}_w)^2 \right]$), where $w_i = 1/\sigma^2(x_i)$ and the weighted means $\hat{y}_w$ and $\hat{x}_w$ are now defined in terms of these weights. The following lemmas show that the resampling scheme described above is model-consistent in this respect in that, under $\xi^*$, $\hat{\beta}_E$ similarly tends asymptotically to

$$
\hat{\beta}^{\nu}_T = \frac{\sum_{i \in S^*} w_0 Y_i^* x_i - \hat{y}_w^* \hat{x}_w^*}{\sum_{i \in S^*} w_0 (x_i - \hat{x}_w^*)^2}
$$

where $w_0 = 1/\hat{\sigma}^2_{h_0}(x_i)$ and the weighted means $\hat{y}_w^*$ and $\hat{x}_w^*$ are defined in terms of these weights. For a general discussion of the validation of bootstrap sampling schemes, see Shao and Tu (1995, p.76).

**Lemma 1.** Given H1-H6, K1, K2, B1 and B2,

$$
n^{-1/4} \sup |\hat{\beta}_h^2(x) - \hat{\sigma}_{h_0}^2(x)| \to_p 0
$$

where the supremum is taken over $\Gamma$.

**Proof.** See the Appendix.

**Lemma 2.** Given H1-H6, K1, K2, B1 and B2,

$$
\sqrt{n}(\hat{\beta}_E - \hat{\beta}_T) \to_p 0.
$$

**Proof.** See the Appendix.
Lemma 3. Given $H_1$-$H_6$, $K_1$, $K_2$, $B_1$ and $B_2$, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_T - \hat{\beta}_E)$ is the same as that of $\sqrt{n}(\hat{\beta}_T - \beta)$:

$$\sqrt{n}(\hat{\beta}_T - \hat{\beta}_E) \rightarrow_d N \left( 0, \sqrt{V} \right),$$

in probability.

Proof. See the Appendix.

Similarly, the following theorem proves model-consistency in the sense that the asymptotic behaviour of $\hat{F}^*$ relative to $F_N^*$ is the same as the asymptotic behaviour of $\hat{F}$ relative to $F_N$ that was described by Theorem 1.

Theorem 3. Given $H_1$-$H_6$, $K_1$, $K_2$, $B_1$ and $B_2$,

$$\sqrt{n} \left\{ \hat{F}^*(t) - F_N^*(t) \right\} \rightarrow_d N(0, V^*),$$

in probability, where

$$V^* = \left\{ (1 - f)^2 \left[ I_1^* + I_2^* \hat{W}_{ho}^{-1} (1 + \tau_{ho}^{-2} \hat{\mu}_{ho}^2) + I_3^* \hat{W}_{ho}^{-1} \tau_{ho}^{-2} - I_2^* I_4^* \hat{W}_{ho}^{-1} \tau_{ho}^{-2} \mu_{ho} \right. \right.$$  \[21\]

$$+ \left. I_2^* I_5^* \int (1 - \tau_{ho}^{-2} \hat{\mu}_{ho}(x - \hat{\mu}_{ho})) \hat{\sigma}_{ho}^{-1}(x) dx + I_5^* I_1^* I_4^* \int (x - \hat{\mu}_{ho}) \hat{\sigma}_{ho}^{-1}(x) dx \right\}^{1/2}.$$  \[21\]

Proof. Analogous to that of Theorem 1, using Lemmas 1-3.

Remark 3.2. An alternative approach to bootstrapping $\hat{F}$ would be to use a bootstrap model

$$\xi: Y_i^* = \hat{\alpha}_E + \hat{\beta}_E x_i + r_i Z_i$$

(where the $Z_i$ are as above and the $r_i$ are the errors $Y_i - \hat{\alpha}_E - \hat{\beta}_E x_i$ of $\xi$) and to use samples $S^*$ of bootstrap populations $P^*$ generated using this model. However, this approach is not model-consistent in the sense of Theorem 3, because the resulting $[\hat{F}^*(t) - F_N^*(t)]$ are asymptotically functions of the $r_i$ rather than of the variance $\sigma^2(x)$. 

12
4 Simulation studies and application to real data

4.1 Simulations

The non-asymptotic behaviour of \( \hat{F} \) and \( \hat{F}_L \) was first investigated in simulation studies using artificial populations \( P \) of sizes \( N = 240 \) and \( N = 500 \). For each \( P \), the design set \( \{x_k\} \) was generated by simple random sampling from a uniform distribution on \([-0.5, 0.5]\), and the corresponding \( Y_k \) were obtained from eq.1 with \( \alpha = 50 \), \( \beta = 60 \), \( \varepsilon_k \) generated by simple random sampling from a standard normal distribution, and \( \sigma(x_k) \) calculated from one of three different variance models of increasing heteroscedasticity:

1. \( \sigma_1^2(x_k) = 100 + 0.25t(x_k)^2 \),
2. \( \sigma_2^2(x_k) = \{0.25 \exp[0.04|t(x_k)|]\}^2 \),
3. \( \sigma_3^2(x_k) = \{0.25 \exp[t(x_k)^2/3200]\}^2 \),

where \( t(x_k) = \alpha + \beta x_k \). To evaluate the influence of the relative sample size \( f \), for \( N = 240 \) the sample size \( n \) was fixed as 60 \( (f = 0.25) \), and for \( N = 500 \) \( n = 50 \) \( (f = 0.1) \).

For each of the six combinations of variance model and \( N \), \( I = 1000 \) populations \( P^i \) were generated as above; a sample \( S^i \) of size \( n \) was taken from each without replacement; and for the quartile points \( t_q \) such that \( F_N(t_q) = q = 0.25, 0.50 \) and 0.75, values \( \hat{F}_N^i(t_q) \) and \( \hat{F}_L^i(t_q) \) were calculated from each sample \( S^i \) as in Section 2, for which purpose the variance estimator \( \hat{\sigma}_h^2(x) \) (eq.4) was calculated using the Epanechnikov kernel and a window parameter \( h \) found by cross-validation and WARPing with approximation error \( \delta = 0.01 \) (see Härdle 1991, p.137). From each \( S^i \) we also calculated the estimate \( F_n^i(t_q) \) afforded by the empirical distribution function. The performance of \( \hat{F}, \hat{F}_L \) and \( F_n \) at the test points \( t_q \) was evaluated in terms of their mean square error \( MSE, MSE_L \) and \( MSE_n \), which are approximated by Monte Carlo:

\[
\begin{align*}
MSE(t) & \approx \frac{1}{I} \sum_{i=1}^{I} \left( \hat{F}_N^i(t) - F_N^i(t) \right)^2, \\
MSE_L(t) & \approx \frac{1}{I} \sum_{i=1}^{I} \left( \hat{F}_L^i(t) - F_N^i(t) \right)^2, \\
MSE_n(t) & \approx \frac{1}{I} \sum_{i=1}^{I} \left( F_n^i(t) - F_N^i(t) \right)^2,
\end{align*}
\]
Table I. Mean square error ($MSE \times 10^4$) of $F_n$, $\hat{F}_L$ and $\hat{F}$ in 1000-trial simulations, for the quartiles of $F_N(t_q)$, and the bootstrap estimation of $MSE$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$F_N(t)$</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>$MSE_n$</td>
<td>23.80</td>
<td>31.29</td>
<td>23.65</td>
</tr>
<tr>
<td>240</td>
<td>$MSE_L$</td>
<td>17.86</td>
<td>23.23</td>
</tr>
<tr>
<td>60</td>
<td>$MSE$</td>
<td>17.81</td>
<td>23.24</td>
</tr>
<tr>
<td>$MSE_n$</td>
<td>17.62</td>
<td>22.88</td>
<td>17.76</td>
</tr>
<tr>
<td>$MSE_n$</td>
<td>33.94</td>
<td>45.48</td>
<td>33.06</td>
</tr>
<tr>
<td>500</td>
<td>$MSE_L$</td>
<td>25.86</td>
<td>30.98</td>
</tr>
<tr>
<td>50</td>
<td>$MSE$</td>
<td>25.96</td>
<td>30.74</td>
</tr>
<tr>
<td>$MSE_n$</td>
<td>24.67</td>
<td>31.63</td>
<td>24.23</td>
</tr>
</tbody>
</table>

where $F_N^3$ is the distribution function of $P^3$. The results (Table I) showed, as expected, that $F_n$ was not only less accurate than $\hat{F}$ and $\hat{F}_L$, but was also hardly affected by which of the heteroscedasticity models was employed (since $F_n$ ignores the existence of auxiliary information). By contrast, the performance of $\hat{F}$ and $\hat{F}_L$ improved with increasing heteroscedasticity. Furthermore, the mean square error of $\hat{F}$ was in almost all cases less than that of $\hat{F}_L$, especially for the two more severe heteroscedasticity models.

For the least heteroscedastic model, reducing the value of $f$ from 0.25 to 0.1 increased the mean square error of all three estimators considerably. For $F_n$ the same was true for the more heteroscedastic models, but for $\hat{F}$ and $\hat{F}_L$ the mean square error with these models was increased little, if at all, by this change in $f$.

To obtain a bootstrap estimate $MSE_n^*$ of the mean square error in $\hat{F}^*$ we proceeded as follows. $B = 1000$ bootstrap populations $P^{*ib}$ of size $N$ were generated from $S^i$ in accordance with eq.12 using the pilot bandwidth $h_0 = 2h$ (with $h$ determined from $S^i$ as above) and $Z_k$ from the distribution of eq.13, in which the $\hat{p}_i$ were obtained by numerical solution of eqs.14 and 15 and the bandwidth parameter $h_1$.
as per Bowman et al. (1998), i.e. by minimization of

\[ n^{-1} \sum_{i \in S} \left\{ I \left( (\hat{\xi}_i - \bar{\xi}) \leq u \right) - G^*_i(u) \right\}^2 du, \]

where

\[ G^*_i(u) = \sum_{i_1 \in S \setminus i \neq i} \hat{p}_{i_1} L \left\{ (u - (\hat{\xi}_{i_1} - \bar{\xi})) / h_1 \right\}, \]

\( L \) being the Epanechnikov kernel. From each \( P^{*ib} \) a sample \( S^{*ib} \) of size \( n \) was then drawn without replacement, \( \hat{F}^{*ib}(t) \) was calculated as in eq.17, and the mean square error of \( \hat{F}^i \) was estimated by

\[ MSE^*_i(t) \approx \frac{1}{B} \sum_{b=1}^{B} \left( \hat{F}^{*ib}(t) - F^{*ib}_N(t) \right)^2, \]

where \( F^{*ib}_N \) is the distribution function of \( y \) in \( P^{*ib} \). The bootstrap mean square error was calculated as

\[ MSE_i(t) \approx \frac{1}{I} \sum_{i=1}^{I} MSE^*_i(t) = \frac{1}{I} \frac{1}{B} \sum_{i=1}^{I} \sum_{b=1}^{B} \left( \hat{F}^{*ib}(t) - F^{*ib}_N(t) \right)^2. \] (23)

The results show a good behaviour of our bootstrap resampling method: \( MSE_i(t) \) mimics \( MSE(t) \) in all points and different degrees of heteroscedasticity, this illustrates the result of Theorem 3.

4.2 Application to real data

We finally wished to compare the performance of the new estimators with that of the original Chambers-Dunstan estimator as it is generally used in practice, i.e. assuming the errors of the superpopulation model to be homoscedastic even in the face of evidence to the contrary. To this end we applied \( \hat{F}, \hat{F}_L \) and that of Chambers et al. (1992) (and also, for further comparison \( F_n \)) to a set of \( I = 1000 \) without-replacement samples \( S^i \) of size \( n = 85 \) taken from a population \( P = \{(x_k, Y_{1k}, Y_{2k}, Y_{3k})\} \) comprising data obtained in a 1982 survey of 338 Queensland sugar cane plantations and used by Chambers and Dunstant (1986). Here \( x_k \) (auxiliary information) is the area of plantation \( k \), \( Y_{1k} \) is its total crop yield, \( Y_{2k} \) is the gross income obtained by the crop, and \( Y_{3k} \) is the total cost incurred in obtaining the crop. Figure 1 shows that for each of the \( Y \) it is reasonable to adopt a heteroscedastic linear regression model of its dependence on \( x \), and hence to use \( \hat{F} \) or \( \hat{F}_L \) to estimate its distribution function.
Figure 1: Plots of crop yield, gross income and costs against plantation area for 85 Queensland sugar cane plantations.
Table II. Mean square error (MSE $\times 10^4$) of $F_n$, $F_{cd}$, $\hat{F}$ and $\hat{F}_L$ when applied to one thousand 85-member samples drawn from the data for 338 Queensland sugar cane plantations used by Chambers and Dunstan (1986), for the quartiles of $F_N(t_q)$.

For approximately the same values of $q = F_N(t_q)$ as in Section 4.1, Table II lists, for each $Y_*$, the mean square errors of the above estimators, defined by

\begin{align}
MSE_{cd}(t) &\approx \frac{1}{T} \sum_{i=1}^{T} \left( \hat{F}_{cd}^i(t) - F_N(t) \right)^2, \\
MSE(t) &\approx \frac{1}{T} \sum_{i=1}^{T} \left( \hat{F}^i(t) - F_N(t) \right)^2, \\
MSE_L(t) &\approx \frac{1}{T} \sum_{i=1}^{T} \left( \hat{F}_L^i(t) - F_N(t) \right)^2, \\
MSE_n(t) &\approx \frac{1}{T} \sum_{i=1}^{T} \left( \hat{F}_n^i(t) - F_N(t) \right)^2, \\
\end{align}

(24)

where $\hat{F}_{cd}^i$ is the estimator of Chambers et al. (1992), $F_N$ is the distribution of the $Y_*$ in question defined by $P$, and $\hat{F}^i$, etc., are the estimators as derived from $S^i$. For each of the three $Y_*$, all the estimators have greatest mean square error at the second quartile, but the variance of the empirical estimate is everywhere considerably greater than those of the other estimators. Even though the variance of the $Y_*$ only grows like $\sqrt{T}$ (Chambers and Dunstan 1986), the fact that $\hat{F}_{cd}$ ignores heteroscedasticity is reflected by $MSE_{cd}$ being always greater than $MSE$ and almost always greater than $MSE_L$. This shows the sensibility of the estimator of Chambers-Dunstan type to the model misspecification.
5 Conclusions

The above results show that both \( \hat{F} \) and \( \hat{F}_L \) can be considerably more accurate than the empirical estimator \( F_n \) when the population being studied conforms reasonably well to the model of eq.1, and that \( \hat{F} \) tends to become more accurate than \( \hat{F}_L \) as the heteroscedasticity of this model increases. The estimation errors of both \( \hat{F} \) and \( \hat{F}_L \) have asymptotically normal distributions of mean zero and variances that are complex generalizations of the expression obtained by Chambers et al. (1992) for homoscedastic superpopulation regression models. In particular applications, the error variance of \( \hat{F} \) and \( \hat{F}_L \) may be estimated by the model-consistent bootstrap method described in Section 3, which in the simulations of Section 4 proved to be reasonably accurate.

In short, we have generalized the Chambers-Dunstan estimator to linear superpopulation regression models with unknown heteroscedastic error variance, and have developed a model-consistent resampling scheme allowing estimation of the error variance and other properties of the new distribution function estimators.

6 Acknowledgements

This work was supported through Grants BFM2002-03213 and PGIDIT03PXIC20702PN by the Spanish Ministry of Science and Technology, in collaboration with the European Regional Development Fund, and Xunta de Galicia, respectively.

APPENDIX

A Preliminaries

The dummy indices \( i \) and \( j \) will identify elements of \( S \) and \( P - S \), respectively, in relation to \( P \), and of \( S^* \) and \( P - S^* \) in relation to a bootstrap population \( P^* \). Orders of probability under \( \xi \) are denoted \( O_p \) and \( o_p \), and orders of probability under \( \xi^* \) (i.e. conditional on the sample \( S \)) by \( O_{p^*} \) and \( o_{p^*} \). The
weighted sample mean of a variable $U$ is $\bar{u}_w = \left[\sum_{i \in S} w_i\right]^{-1} \sum_{i \in S} w_i u_i$, where $w_i = 1/\sigma^2(u_i)$, and the
weighted sample variance is $s^2_{w u} = \left[\sum_{i \in S} w_i\right]^{-1} \sum_{i \in S} w_i (u_i - \bar{u}_w)^2$; $\bar{u}_{w0}$ and $s^2_{w0 u}$ are the corresponding
quantities calculated from the bootstrap sample $S^*$ with weights $w_{0i} = 1/\sigma^2_{h0}(u_i)$; and $\bar{u}_{w*}$ and $s^2_{w* u}$ are the corresponding quantities calculated from the bootstrap sample $S^*$ with weights $w^*_i = 1/\sigma^2_{h*}(u_i)$. The
following abbreviations are used:

\[
\begin{align*}
  t_j &= t - \alpha - \beta x_j, \\
  \hat{t}_j &= t - \hat{\alpha}_T - \hat{\beta}_T x_j, \\
  \Delta_\alpha &= \hat{\alpha}_T - \alpha, \\
  \Delta_\beta &= \hat{\beta}_T - \beta, \\
  \hat{D}_{ij} &= \frac{\bar{h}(x_j)}{\sigma(x_i)} \left[ \frac{t_j}{\bar{h}(x_j)} + \Delta_\alpha \left( \frac{1}{\bar{h}(x_i)} - \frac{1}{\bar{h}(x_j)} \right) + \Delta_\beta \left( \frac{x_i}{\bar{h}(x_i)} - \frac{x_j}{\bar{h}(x_j)} \right) \right].
\end{align*}
\]

It is to be noted that

\[
\begin{align*}
  \hat{G} \left( \frac{t - \hat{\alpha}_T - \hat{\beta}_T x_j}{\bar{h}(x_j)} \right) &= n^{-1} \sum_{i \in S} I \left( \frac{Y_i - \hat{\alpha}_T - \hat{\beta}_T x_i}{\bar{h}(x_i)} \leq \frac{t - \hat{\alpha}_T - \hat{\beta}_T x_j}{\bar{h}(x_j)} \right) \\
  &= n^{-1} \sum_{i \in S} I \left( \varepsilon_i \leq \hat{D}_{ij} \right), \\
  \Delta_\beta &= \sum_{i \in S} (x_i - \bar{x}_w) \varepsilon_i \sigma^{-1}(x_i), \\
  \Delta_\alpha &= \sum_{i \in S} w_i \sigma(x_i) \varepsilon_i \left[ 1 - (x_i - \bar{x}_w) \sum_{i \in S} w_i (x_i - \bar{x}_w)^2 \right].
\end{align*}
\]

Note also that the asymptotic distributions of the variables $\sqrt{n} \Delta_\beta$ and $\sqrt{n} \Delta_\alpha$ are both normal, with
mean zero and variances $W^{-1} \tau_{wx}^2$ and $W^{-1} (1 + \tau_{wx}^2 \mu_{wx}^2)$, respectively. The following known or easily
verifiable results are used.

**Lemma 4.** (Bahadur 1966; see also Serfling, 1980, pp.97-99)

Let $a_n = C_0 n^{-q}$, with $q \in (0, 1/2)$ a constant. If $G$ has a bounded first derivative in $\Theta$, then

\[
\sup_{|x| \leq a_n} \left| \{G_n(u + x) - G_n(u)\} - \{G(u + x) - G(u)\} \right| \leq R_n,
\]

where $R_n = o(n^{-1/2})$ is independent of $u \in \Theta$ and $G_n$ is the empirical distribution.

**Lemma 5.** (Carroll, 1982)

\[
n^{-1/4} \sup |\hat{\sigma}_h^2(x) - \sigma^2(x)| \to_{p} 0.
\]
where the supremum is taken over $\Gamma$.

**Lemma 6.** (Carroll, 1982)

For both $U = \sqrt{n}(\hat{\beta}_E - \beta)$ and $U = \sqrt{n}(\hat{\beta}_T - \beta)$,
\[
\sqrt{n}U \to_d N \left(0, \sqrt{W^{-1}r_\infty^2} \right).
\]

**Lemma 7.** If $\{g_n(x)\}$, $\{h_n(x)\}$, $\{\tilde{g}_n(x)\}$ and $\{\tilde{h}_n(x)\}$ are sequences of positive real functions defined on $\Gamma \in \mathbb{R}$, with $\{\tilde{g}_n(x)\}$ bounded above by $A > 0$ and $\{\tilde{h}_n(x)\}$ bounded below by $a > 0$, and such that $\sup |h_n(x) - \tilde{h}_n(x)| \to 0$, then $\exists n_0$ such that $\forall n > n_0$
\[
\sup \left| \frac{g_n(x)}{h_n(x)} - \frac{\tilde{g}_n(x)}{\tilde{h}_n(x)} \right| \leq \frac{\sup |g_n(x) - \tilde{g}_n(x)| + A \sup |h_n(x) - \tilde{h}_n(x)|}{a - \sup |h_n(x) - \tilde{h}_n(x)|}.
\]

**B  Proofs**

**B.1 Theorem 1**

In view of Lemma 6, the asymptotic distribution of the error $[\hat{F}(t) - F_N(t)]$ is the same as if $\hat{F}$ were constructed using $\hat{\alpha}_T$ and $\hat{\beta}_T$ instead of $\hat{\alpha}_E$ and $\hat{\beta}_E$. For the purposes of this theorem we may therefore redefine $r_i$, $\hat{G}$ and $\hat{F}$ as follows
\[
r_i = Y_i - \hat{\alpha}_T - \hat{\beta}_Tx_i,
\]
\[
\hat{G}(u) = n^{-1} \sum_{i \in S} I \left( \frac{r_i}{\hat{\sigma}_h(x_i)} \leq u \right),
\]
\[
\hat{F}(t) = N^{-1} \left\{ \sum_{i \in S} I(Y_i \leq t) + \sum_{j \in P-S} \hat{G} \left( \frac{t-\hat{\alpha}_T-\hat{\beta}_Tx_i}{\hat{\sigma}_h(x_i)} \right) \right\},
\]
so that by eq.2
\[
\hat{F}(t) - F_N(t) = (1 - n/N) \left\{ (N-n)^{-1} \sum_{j \in P-S} \hat{G}(i_j/\hat{\sigma}_h(x_j)) - (N-n)^{-1} \sum_{j \in P-S} I(\varepsilon_j \leq t_j/\sigma(x_j)) \right\}
\]
\[
= (1 - n/N) \left\{ (N-n)^{-1} \sum_{j \in P-S} \hat{G}(i_j/\hat{\sigma}_h(x_j)) - (N-n)^{-1} \sum_{j \in P-S} G(t_j/\sigma(x_j)) \right\} - (1 - n/N) \left\{ (N-n)^{-1} \sum_{j \in P-S} I(\varepsilon_j \leq t_j/\sigma(x_j)) - (N-n)^{-1} \sum_{j \in P-S} G(t_j/\sigma(x_j)) \right\}
\]
\[
= (1 - n/N) \left\{ (N-n)^{-1} \sum_{j \in P-S} I(\varepsilon_j \leq t_j/\sigma(x_j)) - (N-n)^{-1} \sum_{j \in P-S} G(t_j/\sigma(x_j)) \right\},
\]
where $\eta_{N,n} = \sum_{j \in P-S} [I(\varepsilon_j \leq t_j/\sigma(x_j)) - G(t_j/\sigma(x_j))]$. 

20
Since the first term on the right-hand side of eq.28 depends on the sample data while the second, 
\( (1-n/N)(N-n)^{-1} \eta_{N,n} \), depends only on the population not included in the sample, they are independent of each other and may therefore be dealt with separately. The second is easily handled, since by the Central Limit Theorem for double arrays of random variables,

\[
\frac{\sum_{j \in P-S} \left[ I(\varepsilon_j \leq t_j/\sigma(x_j)) - G(t_j/\sigma(x_j)) \right]}{\left( \sum_{j \in P-S} G(t_j/\sigma(x_j))[1-G(t_j/\sigma(x_j))] \right)^{1/2}} \sim d N(0,1),
\]
whence

\[
(n^{1/2}/N) \eta_{N,n} \sim d N \left( 0, \left\{ f(1-f) \int \left[ G \left( \frac{t-\alpha-\beta x}{\sigma(x)} \right) - G \left( \frac{t-\alpha-x}{\sigma(x)} \right) \right] d(x) dx \right\}^{1/2} \right). \tag{29}
\]

To deal with the first term on the right-hand side of eq.28 we likewise split it in two:

\[
(1-n/N)(N-n)^{-1} \sum_{j \in P-S} \left\{ \hat{G}(t_j/\sigma_{h}(x_j)) - G(t_j/\sigma(x_j)) \right\} = \Delta_1 + \Delta_2 \tag{30}
\]

where

\[
\Delta_1 = (1-n/N)(N-n)^{-1} \sum_{j \in P-S} n^{-1} \sum_{i \in S} \left\{ I(\varepsilon_i \leq \hat{D}_{ij}) - G \left( \hat{D}_{ij} \right) \right\} - \left[ I(\varepsilon_i \leq t_j/\sigma(x_j)) - G(t_j/\sigma(x_j)) \right]
\]

and

\[
\Delta_2 = (1-n/N)(N-n)^{-1} \sum_{j \in P-S} \left\{ n^{-1} \sum_{i \in S} G \left( \hat{D}_{ij} \right) - G(t_j/\sigma(x_j)) \right\} + \left[ n^{-1} \sum_{i \in S} I(\varepsilon_i \leq t_j/\sigma(x_j)) - G(t_j/\sigma(x_j)) \right].
\]

To show that the first term, \( \Delta_1 \), is negligible, we first define

\[
a_n = \max_{i \in S} \left| \frac{\hat{h}(x_i)}{\sigma(x_i)} \frac{t_j}{\sigma(x_j)} - \frac{\hat{h}(x_i)}{\sigma(x_i)} \right| \Delta_\alpha \left( \frac{1}{\hat{h}(x_i)} - \frac{1}{\sigma(x_i)} \right) + \Delta_\beta \left( \frac{x_i}{\hat{h}(x_i)} - \frac{x_i}{\sigma(x_i)} \right)
\]

and we note that \( a_n = O_p(n^q) \) with \( q \in (0,1/2) \) (by H4). Also,

\[
n^{-1} \sum_{i \in S} I(\varepsilon_i \leq (t_j/\sigma(x_j)) - a_n) \leq n^{-1} \sum_{i \in S} I(\varepsilon_i \leq \hat{D}_{ij}) \leq n^{-1} \sum_{i \in S} I(\varepsilon_i \leq (t_j/\sigma(x_j)) + a_n),
\]

which implies that \( n^{-1} \sum_{i \in S} I(\varepsilon_i \leq \hat{D}_{ij}) = n^{-1} \sum_{i \in S} I(\varepsilon_i \leq (t_j/\sigma(x_j)) + c_n) \) with \( |c_n| \leq a_n \); and simi-
larly, $G(\hat{D}_{ij}) = G([t_j/\sigma(x_j)] + c_n)$. Therefore, by Lemma 4,

$$n^{-1} \sum_{i \in S} \left[ I \left( \varepsilon_i \leq \hat{D}_{ij} \right) - G \left( \hat{D}_{ij} \right) \right] - n^{-1} \sum_{i \in S} \left[ I \left( \varepsilon_i \leq t_j/\sigma(x_j) \right) - G \left( t_j/\sigma(x_j) \right) \right]
\leq n^{-1} \sum_{i \in S} I \left[ \varepsilon_i \leq (t_j/\sigma(x_j)) + c_n \right] - n^{-1} \sum_{i \in S} I \left[ \varepsilon_i \leq t_j/\sigma(x_j) \right]
- \{G([t_j/\sigma(x_j)] + c_n) - G(t_j/\sigma(x_j))\} \leq R_n = o_p(n^{-1/2}).$$

To tackle 2 we define

$$S_{ni} = (N - n)^{-1} \sum_{j \in P - S} I \left( \varepsilon_i \leq t_j/\sigma(x_j) \right),$$

$$a_{ni} = \frac{1 - \bar{x}_w^2 s_{w2}^2 (x_i - \bar{x}_w)}{n^{-1} \sum_{j \in S} w_i},$$

$$b_{ni} = \frac{(x_i - \bar{x}_w)/\sigma(x_i)}{n^{-1} s_{w}^2 \sum_{j \in S} w_i},$$

and consider the sequence of independent random random vectors

$$\mathbf{Z}_{ni} = \begin{pmatrix} S_{ni} \\ a_{ni} \varepsilon_i \\ b_{ni} \varepsilon_i \end{pmatrix}$$

which have means $\mu_n$ and covariance matrices $\Sigma_{ni}$ given by

$$\mu_n = \begin{pmatrix} (N - n)^{-1} \sum_{j \in P - S} G \left( t_j/\sigma(x_j) \right) \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma_{ni} = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \sigma_{i13} \\ \sigma_{i12} & \sigma_{i22} & \sigma_{i23} \\ \sigma_{i13} & \sigma_{i23} & \sigma_{i33} \end{pmatrix}.$$
As \( n \to \infty \),

\[
n^{-1} \sum_{i \in S} \Sigma_{n_i} \to \Sigma
\]

where

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{22} & \sigma_{23} \\
\sigma_{33}
\end{pmatrix},
\]

with

\[
\begin{align*}
\sigma_{11} &= \int \int \left\{ G \left( \frac{t - \alpha - \beta u}{\sigma(u)} \right) - G \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \right\} d(u)d(v)dudv, \\
\sigma_{22} &= W^{-1} \left( 1 + \tau_{wx}^2 \mu_{wx}^2 \right), \\
\sigma_{33} &= W^{-1} \tau_{wx}^2, \\
\sigma_{12} &= W^{-1} \int \left[ 1 - \tau_{wx}^2 \mu_{wx}(x - \mu_{wx}) \right] \sigma(x)^{-1} d(x)dx \int \int \frac{(t - \alpha - \beta u)/\sigma(u)}{\sigma(u)} zg(z) d(u) dz, \\
\sigma_{13} &= W^{-1} \tau_{wx}^2 \int (x - \mu_{wx}) \sigma(x)^{-1} d(x)dx \int \int \frac{(t - \alpha - \beta u)/\sigma(u)}{\sigma(u)} zg(z) d(u) dz, \\
\sigma_{23} &= -W^{-1} \tau_{wx}^2 \mu_{wx}.
\end{align*}
\]

Also, since \( E|\epsilon|^{2+\delta} = O(1) \) for some \( \delta > 0 \) (H5), the inequality \( E|X + Y|^s \leq 2^{s-1}(E|X|^s + E|Y|^s) \), (Rohatgi 1976, p.164), with \( s = 2 + \delta \), implies that for all \( \nu > 0 \), as \( n \to \infty \)

\[
n^{-1} \sum_{i \in S} \int_{||\mathbf{Z}_n - \mu_n|| > \nu \sqrt{n}} \mathbf{Z}_n - \mu_n \|^2 dG_i \to 0.
\]

Hence, by the vector Central Limit Theorem (Serfling 1980, p.30),

\[
n^{-1} \sum_{i \in S} \mathbf{Z}_{n_i} - \mu_n \to_d N(\mu, n^{-1} \Sigma),
\]

(31)

We also note that as \( N \to \infty \) and \( n \to \infty \) in such a way that \( n/N \to f \in (0, 1) \), then \( C_n^t \to_p C^t \), where \( C_n^t = (C_{n1}, C_{n2}, C_{n3}) \) with

\[
C_1 = 1, \\
C_2 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{1}{\sigma(u)} - \frac{1}{\sigma(v)} \right) d(u)d(v)dudv, \\
C_3 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{u}{\sigma(u)} - \frac{v}{\sigma(v)} \right) d(u)d(v)dudv.
\]

23
and \( C^t = (C_1, C_2, C_3) \) with

\[
C_1 = 1, \\
C_2 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{1}{\sigma(u)} - \frac{1}{\sigma(v)} \right) d(u) d(v) du dv, \\
C_3 = \int \int g \left( \frac{t - \alpha - \beta v}{\sigma(v)} \right) \left( \frac{u}{\sigma(u)} - \frac{v}{\sigma(v)} \right) d(u) d(v) du dv.
\] (32)

Finally, in view of the results of Section A, expanding \( G \) to second order about \( t_j / \sigma(x_j) \) affords

\[
G \left( \hat{D}_{ij} \right) - G \left( t_j / \sigma(x_j) \right) = g \left( t_j / \sigma(x_j) \right) \left( \hat{D}_{ij} - t_j / \sigma(x_j) \right) + o_p(n^{-1/2}).
\] (33)

which together with eqs.31 and 32 implies that, by the Crámer-Wold Theorem (Serfling 1980 pp.17 and 26),

\[
n^{1/2} \Delta_2 \to_d N(0, V')
\] (34)

where

\[
V' = \left\{ (1 - f)^2 \left[ I_1 + I_2^2 W^{-1} (1 + \mu_{wx}^2 \tau_{wx}^{-2}) + I_2^3 W^{-1} \tau_{wx}^{-2} - I_2 I_4 W^{-1} \mu_{wx} \tau_{wx}^{-2} \\
+ I_4^2 \int \left( 1 - \tau_{wx}^{-2} \mu_{wx} (x - \mu_{wx}) \right) \sigma^{-1}(x) d(x) dx + I_4 I_3 \tau_{wx}^{-2} \int \left( x - \mu_{wx} \right) \sigma^{-1}(x) d(x) dx \right] \right\}^{1/2}.
\] (35)

This completes the proof of Theorem 1. \( \square \)

B.2 Lemma 1

The dominant terms of \( \hat{\sigma}_{\theta}^2(x) \) and \( \hat{\alpha}_{\theta}^2(x) \) are the numerator of eq.4 (with \( h = h_0 \)) and the numerator of eq.18, respectively. In the latter case, expanding the square affords

\[
(nh)^{-1} \sum_{i \in S^*} k \left\{ (x - x_i) / h \right\} \left( Y_i^* - \hat{\alpha}_L^* - \hat{\beta}_L^* x_i \right)^2 = \hat{V}_n^*(x)
\]

where \( \hat{V}_n^*(x) = \hat{V}_n^*_{h_1}(x) - 2 \hat{V}_n^*_{h_2}(x) + \hat{V}_n^*_{h_3}(x) \) with

\[
\hat{V}_n^*_{h_1}(x) = (nh)^{-1} \sum_{i \in S^*} k \left\{ (x - x_i) / h \right\} \hat{\sigma}_{\theta_0}^2(x_i) Z_i^2,
\]

\[
\hat{V}_n^*_{h_2}(x) = (nh)^{-1} \sum_{i \in S^*} k \left\{ (x - x_i) / h \right\} \hat{\sigma}_{\theta_0}(x_i) Z_i \left( \hat{\alpha}_L^* - \hat{\alpha}_E + (\hat{\beta}_L^* - \hat{\beta}_E) x_i \right),
\]

\[
\hat{V}_n^*_{h_3}(x) = (nh)^{-1} \sum_{i \in S^*} k \left\{ (x - x_i) / h \right\} \left( \hat{\alpha}_L^* - \hat{\alpha}_E + (\hat{\beta}_L^* - \hat{\beta}_E) x_i \right)^2.
\] (36)
Since \( k \) and \( \Gamma \) are bounded, and since \( \hat{\beta}_L^* = \hat{\beta}_E + O_{p_1}(n^{-1/2}) \) (\( h_0 \) satisfies the conditions for \( \hat{\sigma}_{h_0}^2(x) \to_p \sigma^2(x) \)), then

\[
n^{1/4} \sup |\hat{V}_{h_0}^*(x)| = o_p(1).
\]

Also,

\[
n^{1/4} \sup |\hat{V}_{h_2}^*(x)| = o_p(1)
\]

(by \( K_2 \), since \( E_* \hat{V}_{h_2}^*(x) = O_p(n^{-1}) \) and \( Var_* \hat{V}_{h_2}^*(x) = O_p(n^{-2}(nh)^{-1}) \)); and,

\[
E_* \{ \hat{V}_{h_1}^*(x) \} = (nh)^{-1} \sum_{i \in S^*} k \{(x - x_i)/h\} \hat{\sigma}_{h_0}^2(x_i)
\]

and

\[
n^{1/4} \sup |\hat{V}_{h_1}^*(x) - E_* \{ \hat{V}_{h_1}^*(x) \}| = o_p(1),
\]

by \( B_1 \) and \( B_2 \).

Finally, by \( K_1, K_2 \) and \( B_2 \), expanding \( \hat{\sigma}_{h_0}^2(x) \) and \( d \) to second order as Taylor series about \( x \) shows that

\[
n^{1/4} \sup |E_* \{ \hat{V}_{h_1}^*(x) \} - \hat{\sigma}_{h_0}^2(x)| = o_p(1)
\]

and hence that

\[
n^{1/4} \sup |\hat{V}_h^*(x) - \hat{\sigma}_{h_0}^2(x)| \leq n^{1/4} \left\{ \sup |\hat{V}_{h_1}^*(x) - E_* \{ \hat{V}_{h_1}^*(x) \}| + 2 \sup |\hat{V}_{h_2}^*(x)| + \sup |\hat{V}_{h_3}^*(x)| + \sup |E_* \{ \hat{V}_{h_1}^*(x) \} - \hat{\sigma}_{h_0}^2(x)| \right\} = o_p(1).
\]

\( \square \)
B.3 Lemma 2

Writing $\Sigma_0$ for $\sum_{i \in S^*} w_{0i} \left(x_i - \bar{x}_{w_0}\right)^2$ and $\Sigma^*$ for $\sum_{i \in S^*} w_i^* \left(x_i - \bar{x}_{w_0}\right)^2$,

$$n^{1/2} \left( \hat{\beta}_E - \hat{\beta}_T \right) = n^{1/2} \left\{ \Sigma_0^{s-1} \sum_{i \in S^*} w_i^* \left(x_i - \bar{x}_{w_0}\right) \hat{\delta}_{h_0}(x_i) Z_i - \Sigma_0^{s-1} \sum_{i \in S^*} w_{0i} \left(x_i - \bar{x}_{w_0}\right) \hat{\delta}_{h_0}(x_i) Z_i \right\}$$

$$= n^{-1/2} \left\{ (n^{-1} \Sigma^*)^{-1} \sum_{i \in S^*} \left(w_i^* - w_{0i}\right) \hat{\delta}_{h_0}(x_i) Z_i \left(x_i - \bar{x}_{w_0}\right) \right.$$ (37)

$$+ \left((n^{-1} \Sigma^*)^{-1} - (n^{-1} \Sigma_0)^{-1}\right)^{-1} \sum_{i \in S^*} w_{0i} \hat{\delta}_{h_0}(x_i) Z_i \left(x_i - \bar{x}_{w_0}\right)$$

$$+ (n^{-1} \Sigma^*)^{-1} \sum_{i \in S^*} w_i^* \hat{\delta}_{h_0}(x_i) Z_i \left(\bar{x}_{w_0} - \bar{x}_{w_0}^*\right) \right\}.$$ 

But by the weak law of large numbers and Lemmas 1 and 7, and since $\hat{\sigma}_{h_0}^2(x) \rightarrow_p \sigma^2(x)$,

$$n^{-1} \Sigma_0 \rightarrow_p W \tau_{w_0}^2,$$

$$n^{-1} \left( \Sigma^* - \Sigma_0 \right) \rightarrow_p 0,$$

where $W$ and $\tau_{w_0}^2$ are as defined in Section 2.1.1, and each of the terms on the right-hand side of eq.37 is therefore $o_p(1)$. □

B.4 Lemma 3

Expanding $\hat{\beta}_E$ and $\hat{\beta}_T$ affords

$$n^{1/2} \left( \hat{\beta}_T - \hat{\beta}_E \right) = \left(n^{-1} \tau_{w_0}^2 \sum_{i \in S^*} w_{0i} \right)^{-1} \sum_{i \in S^*} w_{0i} \hat{\delta}_{h_0}(x_i) Z_i \left(x_i - \bar{x}_{w_0}\right).$$

Hence, since $\hat{\sigma}_{h_0}^2(x) \rightarrow_p \sigma^2$, the lemma follows from the weak law of large numbers and the Central Limit Theorem for independent identically distributed random variables. □

References


26


