EFFECTS OF OUTLIERS ON ESTIMATION OF FINITE POPULATION TOTALS

Muga Zablon Maua*

George Otieno Orwa**

Romanus Odhiambo Otieno***

Leo Odiwuor Odongo

Abstract

In this paper, attempt to study effects of outliers on two estimators of finite population total theoretically and by simulation is made. We compare the ratio estimate with the local linear polynomial estimate of finite population total given different finite populations. Both classical and the non parametric estimator based on the local linear polynomial produce good results when the auxiliary and the study variables are highly correlated. It is however noted that in the presence of outlying observations the local linear polynomial perform better with respect to design mean square error (MSE). The non parametric regression estimator based on the local linear polynomial emerges as a better estimator than the ratio estimator in most cases.

* Bondo University College, School of Mathematics and Actuarial Science, Bondo

** Jomo Kenyatta University, Department of Statistics and Actuarial Science, Nairobi

*** Kenyatta University, Department of Mathematics, Nairobi

1. Introduction. A statistician working on any set of data needs to test outliers with a view of rejecting, accommodating or incorporating them. Rejection of an outlier depends on a variety of factors relating to the statistician's interest in the practical situation which may call for the removal or replacement of the discordant data after which he proceeds to analyze the residual or modified data on the original model. It must be borne in mind that once an outlier is rejected, the sample is no longer random but censored.

If however the existence of the auxiliary information is discovered in the course of study, then it can be made use of by proposing an estimator. Let $\bar{Y}$ and $\bar{X}$ be the population means of Y_i and X_i respectively, then R is the ratio of Y to X in the population such that:

$$R = \frac{Y}{X} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{N\bar{Y}}{N\bar{X}} = \frac{\bar{Y}}{\bar{X}} \qquad 3.1.1.1$$

For the sample we have;

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y}{x} = \frac{n\bar{y}}{n\bar{x}} = \frac{\bar{y}}{\bar{x}} \qquad 3.1.1.2$$

where, $\hat{R}$ is an estimator of R, y_i and x_i are the study variable for the i^{th} unit in the sample and auxiliary variable for the i^{th} unit in the sample respectively.

The ratio estimator of the population mean is;

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi} \bar{X} = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R}\bar{X} \qquad 3.1.1.3$$

For population totals the ratio estimators are;

$$\hat{Y}_R = N\hat{\bar{Y}}_R = \hat{R}X \qquad 3.1.1.4$$

The ratio estimator is normally a biased estimator and its bias is determined by;

$$Bias\left(\hat{\bar{Y}}\right) = E\left(\hat{\bar{Y}}\right) - \bar{Y} = E\left(\hat{R}\bar{X} - R\bar{X}\right) = \bar{X}E\left(\hat{R} - R\right) = \bar{X}Bias\hat{R}$$

To obtain bias $\hat{R}$, one may proceed as follows.

Let $$\hat{R}-R=\frac{\bar{y}}{\bar{x}}-R=\frac{\bar{y}-R\bar{x}}{\bar{x}}=\frac{1}{\bar{x}}\left(\bar{y}-R\bar{x}\right) \qquad 3.1.1.5$$

But $$\frac{1}{\bar{x}}=\frac{1}{\bar{X}+\bar{x}-\bar{X}}=\frac{1}{\bar{X}}\left(\frac{1}{1+\frac{\bar{x}-\bar{X}}{\bar{X}}}\right)=\frac{1}{\bar{X}}\left(1+\frac{\bar{x}-\bar{X}}{\bar{X}}\right)^{-1}$$

By Taylor series expansion we get

$$\frac{1}{\bar{x}}\approx\frac{1}{\bar{X}}\left(1-\frac{\bar{x}-\bar{X}}{\bar{X}}\right)$$

Hence

$$\hat{R}-R\approx\frac{\bar{y}-R\bar{x}}{\bar{X}}-\frac{\left(\bar{y}-R\bar{x}\right)\left(\bar{x}-\bar{X}\right)}{\bar{X}^2}$$

Assuming simple random sampling,

$$E\left(\bar{y}-R\bar{x}\right)=\bar{Y}-R\bar{X}=0,\ \text{also}\ E\left(\bar{y}-R\bar{x}\right)\left(\bar{x}-\bar{X}\right)=E\left(\bar{y}-\bar{Y}\right)\left(\bar{x}-\bar{X}\right)-RE\left(\bar{x}-\bar{X}\right)^2$$

Due to simple random sampling without replacement,

$$E\left(\bar{y}-\bar{Y}\right)\left(\bar{x}-\bar{X}\right)=\frac{N-n}{nN}\sum_{i=1}^{N}\frac{\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{N-1}$$

and

$$E\left(\bar{x}-\bar{X}\right)^2=\text{var}\left(\bar{x}\right)=\frac{N-n}{nN}\sum_{i=1}^{N}\frac{\left(x_i-\bar{X}\right)^2}{N-1}=\frac{N-n}{N-1}S_X^2$$

Let $f=\frac{n}{N}$ be the sampling fraction and $\frac{N-n}{nN}$ be the finite population correction factor.

Then;

$$E\left(\hat{R}-R\right)=\frac{1-f}{n\bar{X}^2}\left(RS_x^2-S_{xy}\right)=\frac{1-f}{n\bar{X}^2}\left(RS_x^2-\rho S_xS_y\right)=Bias\ \left(\hat{R}\right) \qquad 3.1.1.6$$

Where the correlation coefficient ρ between y_i and x_i in the finite population is defined by the equation

$$\rho = \frac{E\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)}{\sqrt{E\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)^2}} = \frac{\sum_{1=i}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)}{\left(N-1\right)S_y S_x}$$

and $S_{yx} = \rho S_y S_x$ is the covariance between y_i and x_i.

Hence

$$Bias\left(\hat{\bar{Y}}_R\right) = \bar{X}\ Bias\left(\hat{R}\right) = \frac{1-f}{n\bar{X}}\left(RS_x^2 - \rho S_x S_y\right) \qquad 3.1.1.7$$

2. Assumptions for proof of theoretical results. We outline the assumptions and prove the theoretical properties

3.1.1 Distribution of the errors under ξ: the errors ε_i are independent and have mean zero, variance $v\left(x_i\right)$ and compact support, uniformly for all N.

3.1.2 For each N, the x_i are considered fixed with respect to the super population model ξ. The x_i are independent and identically distributed $F\left(x\right) = \int_{-\infty}^{x} f\left(t\right)dt$, where $f\left(.\right)$ is a density with compact support $\left[a_x, b_x\right]$ and $f\ \left(x\right) > 0$ for all $x \in \left[a_x, b_x\right]$.

3.1.3 Mean and variance functions $m\left(.\right)$, $v\left(x\right)$ on $\left[a_x, b_x\right]$: the mean function $m\left(.\right)$ is continuous and has $p+1$ continuous derivatives, and the variance function $v\left(x\right)$ is continuous and strictly positive.

3.1.4 Kernel K: the kernel K has compact support [-1, 1], is symmetric and continuous, and satisfies $\int_{-1}^{1} K\left(u\right)du = 1$

3.1.5 Sampling rate $n_N N^{-1}$ and bandwidth h_N: as $N \to \infty, n_N N^{-1} \to \pi \in \left(0,1\right), h_N \to 0$ and $\frac{Nh_N^2}{\left(\log\ \log N\right)} \to \infty$.

3.1.6 Inclusion probabilities: for all N, $\min_{i \in U_N} \pi_i \geq \lambda^* > 0, \min_{i,j \in U_N} \pi_{ij} \geq \lambda^* > 0$ and $\limsup_{N \to \infty} n_N \max_{i,j \in U_N : i \neq j} \left|\pi_{ij} - \pi_i \pi_j\right| < \infty$.

3.1.7 Additional assumptions for higher order inclusion probabilities:

$$\lim_{N\to\infty} n_N^2 \lim_{(i_1,i_2,i_3,i_4)\in D_{4,N}} \left| E_P\left[\left(I_{i_1}-\pi_{i_1}\right)\left(I_{i_2}-\pi_{i_2}\right)\left(I_{i_3}-\pi_{i_3}\right)\left(I_{i_4}-\pi_{i_4}\right)\right]\right| < \infty,$$

$$\lim_{N\to\infty} \lim_{(i_1,i_2,i_3,i_4)\in D_{4,N}} \left| E_P\left[\left(I_{i_1}I_{i_2}-\pi_{i_1}\pi_{i_2}\right)\left(I_{i_3}I_{i_4}-\pi_{i_3}\pi_{i_4}\right)\right]\right| = 0,$$

$$\text{and} \limsup_{N\to\infty} n_N \lim_{(i_1,i_2,i_3)\in D_{3,N}} \left| E_P\left[\left(I_{i_1}-\pi_{i_1}\right)^2\left(I_{i_2}-\pi_{i_2}\right)\left(I_{i_3}-\pi_{i_3}\right)\right]\right| < \infty$$

where $D_{t,N}$ denotes the set of all distinct t-turples $(i_1, i_2, ..., i_t)$ from U_N

3. Asymptotic Design Unbiasedness and Consistency. The price for using $\hat{m}$'s in place of m_i 's in the generalized difference estimator (3.2.2.5) is design bias. The estimator $\hat{t}_y$ is, however, asymptotically design unbiased and design consistent under mild conditions, as the following theorem demonstrates;

Theorem 1. Assume 1-7. The local polynomial regression estimator

$$\tilde{t}_y = \sum_{i\in U_N}\left\{(y_i-\hat{m}_i)\frac{I_i}{\pi_i}+\hat{m}_i\right\}$$

is asymptotically design unbiased (ADU) in the sense that

$$\lim_{N\to\infty} E_P\left[\frac{\tilde{t}_y - t_y}{N}\right] = 0$$

with ξ-probability one, and is design consistent in the sense that

$$\lim_{N\to\infty} E_P\left[I_{\{|\tilde{t}_y - t_y| > N\eta\}}\right] = 0$$

with ξ-probability one for all $\eta > 0$.

Proof. By Markov's inequality, it suffices to show that

$$\lim_{N\to\infty} E_P\left[\frac{\tilde{t}_y - t_y}{N}\right] = 0$$

Write

$$\frac{\tilde{t}_y - t_y}{N} = \sum_{i\in U_N}\frac{y_i - m_i}{N}\left(\frac{I_i}{\pi_i}-1\right) + \sum_{i\in U_N}\frac{\hat{m}_i - m_i}{N}\left(1-\frac{I_i}{\pi_i}\right)$$

Then

$$E_P\left|\frac{\tilde{t}_y - t_y}{N}\right| \le E_P\left|\sum_{i\in U_N}\frac{y_i - m_i}{N}\left(\frac{I_i}{\pi_i}-1\right)\right|$$

$$+\left\{E_P\left[\sum_{i\in U_N}\frac{(\hat{m}_i - m_i)^2}{N}\right]E_P\left[\sum_{i\in U_N}\frac{\left(1-\pi_i^{-1}\right)^2}{N}\right]\right\}^{1/2} \quad (3.2.5.1)$$

Under assumptions 1-6 and using the fact that

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}(y_i - m_i)^2 < \infty$$

By Lemma 2(iv), the first term on the right of (3.4.2.1) converges to zero as $N\to\infty$, following the argument of Theorem 1 in Robinson and Sarndal (1983). Under assumption 6,

$$E_P\left[\sum_{i\in U_N}\frac{\left(1-\pi_i^{-1}I_i\right)^2}{N}\right] = \sum_{i\in U_N}\frac{\pi_i(1-\pi_i)}{N\pi_i^2} \le \frac{1}{\lambda}.$$

Combining this with Lemma 4, the second term on the right of (3.2.5.1) converges to zero as $N\to\infty$, and the theorem follows.

3 Asymptotic Mean Squared Error. In this section we derive an asymptotic approximation to the mean squared error of the local polynomial regression estimator and propose a consistent variance estimator.

Theorem 2.

Assume 1-7. Then

$$n_N E_p\left(\frac{\hat{t}_y - t_y}{N}\right)^2 = \frac{n_N}{N^2}\sum_{i,j\in U_N}(y_i - m_i)(y_j - m_j)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} + 0(1)$$

Proof

Let

$$a_N = n_N^{1/2}\sum_{i\in U_N}\frac{y_i - m_i}{N}\left(\frac{I_i}{\pi_i}-1\right) \text{ and } b_N = n_N^{1/2}\sum_{i\in U_N}\frac{m_i - \hat{m}_i}{N}\left(\frac{I_i}{\pi_i}-1\right).$$

Then

$$E_p\left[a_N^2\right]=\frac{n_N}{N^2}\sum_{i,j\in U_N}\left(y_i - m_i\right)\left(y_j - m_j\right)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}$$

$$\leq\left(\frac{1}{\lambda}+\frac{n_N \max_{i,j\in U_N:i\neq j}\left|\pi_{ij}-\pi_i\pi_j\right|}{\lambda^2}\right)\sum_{i\in U_N}\frac{\left(y_i-m_i\right)^2}{N}$$

so that

$\limsup_{N\to\infty} E_p\left[a_N^2\right]<\infty$ by 6.

By Lemma 5,

$E_p\left[b_N^2\right]=0(1)$,

so that

$$E_p\left[a_N b_N\right]\leq\left(E_P\left[a_N^2\right]E_P\left[b_N^2\right]\right)^{1/2}=0(1)$$

hence $$n_N E_p\left(\frac{\hat{t}_y-t_y}{N}\right)^2=E_p\left[a_N^2\right]+2E_p\left[a_N b_N\right]+E_p\left[b_N^2\right]=E_p\left[a_N^2\right]+0(1)$$

which completes the proof.

Next we show that the asymptotic mean squared error in theorem 2 can be estimated consistently under mild assumptions.

Theorem 3.

Assume 1-7. Then $\lim_{N\to\infty} n_N E_P\left|\hat{V}\left(N^{-1}\hat{t}_y\right)-AMSE\left(N^{-1}\hat{t}_y\right)\right|=0$

where

$$\hat{V}\left(N^{-1}\hat{t}_y\right)=\frac{1}{N^2}\sum_{i,j\in U_N}\left(y_i - \hat{m}_i\right)\left(y_j - \hat{m}_j\right)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\frac{I_i I_j}{\pi_{ij}}$$

and

$$AMSE\left(N^{-1}\hat{t}_y\right)=\frac{1}{N^2}\sum_{i,j\in U_N}\left(y_i - m_i\right)\left(y_j - m_j\right)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}$$

Therefore, $\hat{V}\left(N^{-1}\hat{t}_y\right)$ is asymptotically design unbiased and design consistent for $AMSE\left(N^{-1}\hat{t}_y\right)$

Proof of theorem 3

Write

$$A_N = n_N E_P \left| \frac{1}{N^2} \sum_{i,j\in U_N} (y_i - m_i)(y_j - m_j) \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \frac{I_i I_j - \pi_{ij}}{\pi_{ij}} \right|$$

Now

$$n_N^2 E_P \left(\frac{1}{N^2} \sum_{i,j\in U_N} (y_i - m_i)(y_j - m_j) \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \frac{I_i I_j - \pi_{ij}}{\pi_{ij}} \right)^2$$

$$= n_N^2 \sum_{i,k\in U_N} \frac{1-\pi_i}{\pi_i} \frac{1-\pi_k}{\pi_k} \frac{(y_i - m_i)^2 (y_k - m_k)^2}{N^4} \frac{\pi_{ik} - \pi_i\pi_k}{\pi_i\pi_k}$$

$$+ 2n_N^2 \sum_{i\in U_N} \sum_{k,\ell\in U_N: k\neq\ell} \frac{1-\pi_i}{\pi_i} \frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_k\pi_\ell} \frac{(y_i - m_i)^2 (y_k - m_k)(y_\ell - m_\ell)}{N^4} \times E_P \left[\frac{I_i - \pi_i}{\pi_i} \frac{I_k I_\ell - \pi_{k\ell}}{\pi_{k\ell}} \right]$$

$$+ n_N^2 \sum_{i,j\in U_N: i\neq j} \sum_{k,\ell\in U_N: k\neq\ell} \frac{\pi_{ik} - \pi_i\pi_j}{\pi_i\pi_j} \frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_k\pi_\ell} \frac{(y_i - m_i)(y_j - m_j)(y_k - m_k)(y_\ell - m_\ell)}{N^4}$$

$$\times E_P \left[\frac{I_i I_j - \pi_{ij}}{\pi_{ij}} \frac{I_k I_\ell - \pi_{k\ell}}{\pi_{k\ell}} \right]$$

$$= a_{1N} + a_{2N} + a_{3N}$$

But

$$a_{1N} \leq n_N^2 \sum_{i\in U_N} \frac{(y_i - m_i)^4}{\lambda^3 N^4} + n_N^2 \sum_{i,k\in U_N: i\neq k} \frac{(y_i - m_i)^2 (y_k - m_k)^2 \left|\pi_{ik} - \pi_i\pi_k\right|}{\lambda^4 N^4}$$

$$\leq \left(\frac{1}{N\lambda^3} + \frac{n_N \max_{i,k\in U_N: i\neq k} \left|\pi_{ik} - \pi_i\pi_k\right|}{N\lambda^4} \right) \sum_{i\in U_N} \frac{(y_i - m_i)^4}{N}$$

which goes to zero as $N \to \infty$, and

$$a_{3N} \leq \frac{\left(n_N \max_{i,k\in U_N: i\neq k} \left|\pi_{ik} - \pi_i\pi_k\right|\right)^2}{\lambda^4 \lambda^{*2}}$$

$$\times \sum_{i,j\in U_N:i\neq j}\ \sum_{k,\ell\in U_N:k\neq \ell} \frac{\left|(y_i-m_i)(y_j-m_j)(y_k-m_k)(y_\ell-m_\ell)\right|}{N^4}$$

$$\times \left|E_P\left[\frac{I_iI_j-\pi_{ij}}{\pi_{ij}}\frac{I_kI_\ell-\pi_{k\ell}}{\pi_{k\ell}}\right]\right|$$

$$\leq 0\left(N^{-1}\right)+\frac{\left(n_N \max_{i,k\in U_N:i\neq k}\left|\pi_{ik}-\pi_i\pi_k\right|\right)^2}{\lambda^4\lambda^{*2}}$$

$$\times \max_{(i,j,\ell)\in D_{4,N}}\left|E_P\left[\frac{I_iI_j-\pi_{ij}}{\pi_{ij}}\frac{I_kI_\ell-\pi_{k\ell}}{\pi_{k\ell}}\right]\right|\sum_{i\in U_N}\frac{(y_i-m_i)^4}{N}$$

which converges to zero as $N\to\infty$ by 7. The Cauchy-Schwarz inequality may then be applied to show that $a_{2N}\to 0$ as $N\to\infty$, and it follows that $A_N\to 0$ as $N\to\infty$.

4. Outlier Robust Estimation. As the proof of theorem 5 indicates, the local polynomial regression estimate is robust in the sense of asymptotic attainment of the Godambe –Joshi lower bound. However since the estimator is based on least squares, it is susceptible to the effects of observations with unusual response values (outliers). If an observed y_i is sufficiently far from the bulk of the observed responses for nearby values of x, $\hat{m}(x)$ will be drawn towards the unusual response and away from the majority of the points. Alternative criteria Lowess (Cleveland, 1979), and its successor loess (Cleveland and Delvin, 1988) are nearest neighbor-based local polynomial estimators that allow the data analyst to down weight the effect of unusual observations. This is done through an iterative process. An ordinary local polynomial estimate is first calculated. Observations then have weights $\{\delta_1,\ldots,\delta_2\}$ attached to them, where the weights decrease smoothly as the absolute residual from the loess fit increases. The updated estimate is then the local polynomial estimate with weights ΔW, where $\Delta = diag\{\delta_1,\ldots,\delta_2\}$.The process is then iterated several times. Unfortunately, as Machler (1989) noted, since the original residuals are based on the ordinary non robust loess fit, the robust version can still be sensitive to outliers. The pth order local polynomial regression is based on minimizing

$$\sum_{i=1}^{n}\left[y_i - \beta_0 - \ldots - \beta_p(x_i-x)^p\right]^2 K\left(\frac{x_i-x}{h}\right) \qquad 3.2.9.1$$

Some authors have suggested the related approach of using a local version of M-estimation which attempts to achieve robustness by replacing 3.4.1 by

$$\sum_{i=1}^{n}\rho\left[y_i - \beta_0 - ... - \beta_p\left(x_i - x\right)^p\right]K\left(\frac{x_i - x}{h}\right) \qquad 3.2.9.2$$

$\rho(.)$ is chosen to down weight outliers. This is accomplished by choosing $\rho(.)$ to be symmetric with a unique minimum at zero, so that its derivative $\varphi(.)$ is bounded. Minimization of 3.4.2 requires iterative procedure which is stopped after one or two steps, (Fan and Jiang, 1999). However since the iterations start at the least squares local polynomial estimator, the estimator is still potentially sensitive to outliers. True robustness requires an estimator that is not based on the least squares estimator.

Wang et al (1994) investigated the Least Absolute Values (LAV) version of 3.4.2 estimating $m(.)$ by minimizing

$$\sum_{i=1}^{n}\left|y_i - \beta_0 - ... - \beta_p\left(x_i - x\right)^p\right|K\left(\frac{x_i - x}{h}\right) \qquad 3.2.9.3$$

We investigate the conditional breakdown of 3.2.9.3 and its robustness

3.8. Determining the Conditional Breakdown . From definitions l and m : **Breakdown of an estimator** is the smallest fraction of outliers that can force the estimators to arbitrary values, and is thus the measure of the resistance of the estimator to unusual values. **Breakdown point of an estimator** τ is defined to be $\alpha^* = \min\left[\frac{m}{n}; bias\left(m;\tau,y,X\right) is \text{ inf } inite\right]$

where $bias\left(m;\tau,y,X\right)$ is the maximum bias that can be caused by replacing any m of the original data points by arbitrary values, (Donoho and Huber, 1983). Any estimator that is not at all resistant to outliers, such as one based on least squares has breakdown $\frac{1}{n}$. Since the local polynomial regression estimate $\hat{m}(.)$ is implemented by solving many local regression problems, each centered at an evaluation point x, its breakdown properties are defined on a local level as well. We restrict ourselves to kernel functions $K(.)$ that are positive on a bounded interval [-1, 1]. Conditional breakdown implies that unlike for parametric models, the breakdown point changes depending on the evaluation point of x.

3.8.2. Breakdown based on least absolute values-LAV. In order to describe the breakdown properties of local LAV regression estimators, we first must consider the breakdown point of the weighted LAV regression problem involving the observations for which the weights are positive. The weights that are used in each of the local regression problems are determined by the selected kernel function and bandwidth, i.e. $w_i = h^{-1}K\left(\frac{x_i - x}{h}\right)$. Since the breakdown is based on a set of weighted LAV regressions, it depends at any evaluation point on both the local distribution of the predictor values and the kernel used.

While the local distribution of predictors is typically beyond the control of the data analyst, the choice of kernel is not, leaving open the possibility that it might be chosen in such a way as to make the estimator as robust as possible. At an evaluation point, the bandwidth used determines the set of observations within the local regression. This suggests that the bandwidth could be chosen so as to maximize robustness in some sense, but this is a mistaken conclusion. Wang and Scott derived the bandwidth that minimizes the asymptotic average mean squared error of $\hat{m}$ showing that it satisfies

$$h_{opt} = \left(\frac{36}{f(0)^2 \int m''(x)^2\, dx}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \qquad 3.2.9.4$$

Where f is the density of the errors (taking x to be uniform on [0, 1] and assuming constant variance for the errors).

Thus the optimal choice of h depends on the curvature of m and the density of ε, and cannot be set arbitrarily so as to ensure robustness. Equation 3.5.2.1 assumes use of a uniform kernel. If a different kernel is used, the bandwidth must be adjusted. Wang and Scott showed that the equivalent bandwidth when using a kernel K_2 rather K_1 than satisfies

$$h_{opt}(K_2) = h_{opt}(K_1)\left[\frac{V(K_2)}{V(K_1)}\right]^{\frac{1}{2}} \qquad 3.2.9.5$$

Where $V(K)$ is the variance of the kernel, $\int x^2K(x)dx$. The table below is a list of commonly used kernels. The interpretation of the table is that, for example, if bandwidth h yields an appropriate amount of smoothing when using uniform kernel, the bandwidth 1.291 h is the appropriate choice when using a quadratic kernel. Thus, any comparisons of robustness across kernels corrects for this scale effect by using equivalent bandwidths.

Table 1. Multipliers to give equivalent bandwidths for different kernels

Kernel	Formula	Variance	Multiplier
Uniform	$\frac{1}{2}$	$\frac{1}{3}$	1.000
Quadratic	$\frac{3}{4}(1-x^2)$	$\frac{1}{5}$	1.291
Biweght	$\frac{15}{16}(1-x^2)^2$	$\frac{1}{7}$	1.528
Triweight	$\frac{35}{32}(1-x^2)^3$	$\frac{1}{9}$	1.732
Tricube	$\frac{70}{81}(1-\|x\|^3)^3$	0.1440329	1.521

The robustness of any kernel choice at an evaluation point is evaluated in two ways. First we use the breakdown value, the smallest number of observations that can force the estimator to arbitrary values. Here we do not use the breakdown point, (the proportion of observations in the span of the kernel that can force the estimator to arbitrary values).This is because the number of observations in the span depends on the appropriate multiplier for the bandwidth for the chosen kernel. Suppose the bandwidth used at evaluation point x using a uniform kernel includes $\alpha_u(x)$ observations, with breakdown point $\alpha_u(x)$, then the smallest number of observations that could possibly break down the estimate at x using the uniform kernel is $[n_u(x)\alpha_u(x)]$.

On the other hand, if a quadratic kernel was used, the bandwidth would be 29.1% larger at x yielding $n_q(x)$ observations in the span of the kernel, with $n_q(x)$ probably larger than $n_u(x)$.

The smallest number of observations that could possibly break down the estimate at x using the quadratic kernel is $\left[n_q(x)\alpha_q(x)\right]$,where $\alpha_q(x)$, is the breakdown point at x when using the quadratic kernel. The choice of kernel is up to the data analyst, so the preferred choice on the basis of breakdown would be the one with larger value of $n(x)\alpha(x)$, (the breakdown value), not larger than $\alpha(x)$,(the breakdown point). This argument shows that breakdown value is not sufficient to describe resistance in the nonparametric regression context.

Since the breakdown value is an increasing function of the number of observations in the span of the kernel, kernels with larger equivalent bandwidths (such as the triweight) have advantage over kernels with smaller equivalent bandwidths (such as the uniform) in terms of breakdown value. For this reason we examine a second measure of breakdown. For a given kernel, say there are $n(x)$ observations in the span of the kernel at evaluating point x, and the breakdown value at that point is $b(x)$. The estimator cannot break down at x if the number of outliers within the span of the kernel is less than $b(x)$, so the probability that the estimator will not break down at x is

pr(Estimator cannot break down at x)

$$= \sum_{j=0}^{b(x)-1} P\left(j \text{ of the observations in the span are outliers}\right)$$

Say there are k outliers in the sample, and they are spread randomly over the observations in the sample. Then the probability that j of the observations in the span of the kernel is outliers is hypergeometric,

$$P\left(j \text{ of the observations in the span are outliers}\right) = \frac{\binom{k}{j}\binom{n-k}{n(x)-j}}{\binom{n}{n(x)}}, \text{ with } 0 \le j \le k .$$

Note that if $k < b(x)$, the estimator cannot possibly break down at x, but as k gets larger, the probability of having too many outliers in the span of the kernel increases, decreasing the probability that the estimator cannot beak down. We note that a smaller bandwidth makes it more likely that the estimator cannot break down, since there are fewer observations in the span of the kernel, implying an advantage for kernels with smaller equivalent bandwidths. These two measures quantify a tradeoff between choosing kernels using smaller bandwidths and those using larger bandwidths.

4. Empirical Study

We have considered a natural population taken from the Kenya National Bureau of Statistics, Statistical Abstract 2002. The data is based on wage employment by industry for 100 units. The data for the year 2000 provided the auxiliary variable while that of the year 2001 provided the study variable. The population data with outliers is as a result of deliberate key punch errors on a number of data points from this natural population. For these two sets of data, we have drawn random samples of size 20 and 30, and replicated the experiments 500 times.

We used three different bandwidth $h = 0.1, h = 0.25$ and $h = 0.5$ for the non parametric estimator. The first band width is equal to the post stratum width while the second is based on an ad hoc rule of $\frac{1}{4}$th the data range. The third band width has been incorporated to help verify robustness.

4.3 Results

4.3.6 Relative MSE of the ratio estimator to the local linear polynomial estimator (Natural population and Population with outliers)

		Relative Mean Square Error	
	h	**n=20**	**n=30**
Natural	**0.1**	0.4629	0.8762

population	**0.25**	0.5784	0.7925
	0.5	0.8646	0.9676
Population with outliers	**0.1**	18.747	13.44
	0.25	4.196	2.805
	0.5	1.1324	0.5281

4.4 Discussion of results. It is clear that the ratio estimator performs better than the local linear polynomial estimator when the population is natural irrespective of the variance used. The local linear polynomial regression estimator turns out to be a better estimator when the population contains outliers. It becomes even more robust when the sample size is increased giving room for the presence of more outliers. The relative mean square errors decrease as the bandwidths increase from 0.1 to 0.5 which implies robustness of the local linear polynomial regression estimator when the Quadratic Kernel is combined with a smaller bandwidth.

4.5. Conclusion. From the results obtained we conclude that the choice of an appropriate estimator of finite population totals is important. The ratio estimator will be very useful when the variables are highly correlated such that their graph is a line through the origin. The local linear polynomial is a more appropriate estimator of population total when combined with a smaller bandwidth than the parametric ratio estimator.

References

Augustyns, I. (1997). Local polynomial smoothing of sparse multinomial data. Unpublished Ph.D thesis, Limburgs University.

Brewer, K.R.W.(1963). Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process. Australian Journal of statistics 5,93-105.

Cassel, C-M., Sarndal, C-E., and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. Wiley, New York.

Chambers, R.L. (1986), Outlier robust finite population estimation, JASA, Vol.81, No. 396, pp. 1063-1069.

Horvitz, D.G. and D.J. Thompson. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.

Chen, J. and Qin, J.(1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. Biometrika 80, 107-116.

Cleveland, W.S. (1979). Locally weighted regression and smoothing scatter plots. Journal of the American Statistical association 74, 829-836.

Cleveland, W.S. and Devlin,S.J.(1988). Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association, 83,596-610.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. Wiley, New York.

Cook, R.D. Holschuch, N. and Weisberg, S. (1982), A note on an alternative outlier model, J.R.S.S., B, 44, No.3, pp. 370-376.

Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In: Bickel, P., Doksum, K. and Hodges, J.L. pp. 157-184.

Dorfman, A.H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. Annals of statistics 21, 1452-1475.

Draper, N.R and J.A.John (1981). Influential observations and outliers in regression.Technometrics, 23, 21-26.

Fan, J., Hu,T.C and Truong, Y.K.(1994), Robust non –parametric function estimation. Scandinavian Journal of Statistics, 21,433-446.

Fan, J. and Gijbels, I.(1996). Local Polynomial Modeling and it's applications, Chapman and Hall, London.

Giloni, A. and Padberg, M. (2004). The finite sample breakdown point for local linear regression polynomial. Journal of optimization, 14, 1028-1042.

Godambe, V.P. and Joshi,V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1.Annals of Mathematical Statistics 36, 1707-1722.

Hidiroglou, M.A. and Srinath, K.P. (1981), Some estimators of a population total from simple random samples containing large units, JASA, Vol. 78, No. 375, pp. 690-695.

Horvitz, D.G. and D.J.Thompson. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.

John Wiley and Sons (1996). Outliers in statistical data. New York. 10058-0012.USA.

Opsomer, J.-., and Ruppert,D. (1997). Fitting a bivariate additive model by local polynomial regression. Annals of Statistics 25, 186-211.

Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, Vol.9, No.4, pp. 705-724.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. Biometrika 57, 377-387.

Ruppert, D., and Wand, M.P. (1994). Multivariate locally weighted least squares regression. Annals of Statistics 22, 1346-1370.

Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing, Chapman and Hall, London.

Wang, F.T. and Scott, D.W. (1994), The L_1 method for robust nonparametric regression.

Journal of the A American Statistical Association.