Statistics and Applications {**ISSN 2452-7395** (online)}
Volume 17, No. 2, 2019 (New Series), pp 37-45

Robust Nonparametric Covariance Technique

C.T. Jose[1], K.P. Chandran[2], K. Muralidharan[2], S. Sujatha[1] and B.Ismail[3]
[1]ICAR- Central Plantation Crops Research Institute, Regional Station, Karnataka
[2]ICAR- Central Plantation Crops Research Institute, Kerala
[3]Department of Statistics, Mangalore University, Karnataka

Received: May 27, 2019; Revised: July 01, 2019; Accepted: December 26, 2019

Abstract

Outlier detection and robust estimation are the integral part of data mining and has attracted much attention recently. Generally, the data contain abnormal or extreme values either due to the characteristics of the individual or due to the errors in tabulation, data entry etc. The presence of outliers may badly affect the data modeling and analysis. Analysis of semi-parametric regression with design matrix as the parameter component and covariate as the nonparametric component is considered in this paper. The regression estimate and the cross validation technique can behave very badly in the presence of outliers in the data or when the errors are heavy-tailed. The cross-validation technique to estimate the optimum smoothing parameter will also be affected badly by the presence of outliers. A robust method, which is not influenced by the presence of outliers in the data, is proposed to fit the semi-parametric regression with design matrix as the parameter component and covariate as the nonparametric component. Robust M- kernel weighted local linear regression smoother is used to fit the regression function. The cross-validation technique to estimate the optimum smoothing parameter will also be affected badly by the presence of outliers. A robust cross-validation technique is proposed to estimate the smoothing parameter. The proposed method is useful to compare the treatments after eliminating the covariate effect. The method is illustrated through simulated and field data.

Key words: Nonparametric, robust inference, covariance.

1. Introduction

The fundamental objective of statistical data analysis is to obtain data systematically and to make inferences or appropriate decisions based on the data. Presence of outliers or extreme values in the experimental data is a major concern for data analysis. Outlier is an observation that appears to be inconsistent with the remainder of the observations in the data set. Agricultural field experimental data may contain abnormal or extreme values due to various reasons such as genetical variations (super trees/very low yielders), loss of yield due to pest/ disease infestation, errors in tabulation, data entry etc. These extreme values or outliers, generally increase the experimental error in data analysis. Detection of outliers and the possible remedies are very important in data analysis. These outliers are nuisance for the data analysts. A robust method is proposed for the analysis of semi-parametric regression model in the presence of outliers.

Corresponding Author: C.T. Jose
Email: ctjos@yahoo.com

2. Model Settings and Estimators

The semi-parametric or the non-parametric covariance model considered for the study is of the form

$$\mathbf{Y}=\mathbf{X\beta}+\phi(\mathbf{U})+\boldsymbol{\varepsilon} \tag{1}$$

where, $\mathbf{Y}$ is the observation vector, $\mathbf{m}=\mathbf{X\beta}+\phi(\mathbf{U})$, is the regression function, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of treatment effect, $\phi(\mathbf{U})$ is the non-parametric function representing the relationship between $\mathbf{Y}$-$\mathbf{X\beta}$ and the covariate $\mathbf{U}$ which is assumed to be a smooth function and $\boldsymbol{\varepsilon}$ is the error term assumed to be *iid* with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Backfitting algorithm (Buja *et al.*, 1989) is used to estimate the treatment vector and covariate effect in the regression model and estimates are given by

$$\widehat{\boldsymbol{\beta}}=[\mathbf{X}^T(\mathbf{I}-\mathbf{S})\mathbf{X}]^{-1}\mathbf{X}^T\,(\mathbf{I}-\mathbf{S})\mathbf{Y},\quad \widehat{\boldsymbol{\phi}}=\mathbf{S}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\ \text{ and }\ \widehat{\mathbf{m}}=\mathbf{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\phi}}$$

where, $\mathbf{S}$ is the smoothing matrix derived using local linear regression (Ruppert and Wand, 1994). Let $\mathbf{S}_i$ be the ith row of the smoother matrix, then

$$\mathbf{S}=[\mathbf{S}_1\ \dots\ \mathbf{S}_n]^T$$
$$\mathbf{S}_i^T=\mathbf{e}_1^T\left(\mathbf{Z}_{u_i}^T\mathbf{W}_{u_i}\mathbf{Z}_{u_i}\right)^{-1}\mathbf{Z}_{u_i}^T\mathbf{W}_{u_i}$$

where,

$$\mathbf{Z}_{u_i}=\begin{bmatrix}1 & (u_1-u_i)\\ \vdots & \vdots\\ 1 & (u_n-u_i)\end{bmatrix},\quad \mathbf{e}_1^T=[1\ 0\ 0]$$

and $\mathbf{W}_{u_i}=diag\left\{K\left(\frac{u_1-u_i}{h}\right),\dots,K\left(\frac{u_n-u_i}{h}\right)\right\}$ for some kernel functions K and bandwidth h. The properties of the estimates are provided by Jose and Ismail (2001) and Rupert and Wand (1994). Cross validation (leave-one-out) technique is generally used to estimate the optimum bandwidth h. The cross validation score is given by

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\widehat{m}_{(-i)h}\right]^2$$

where, y_i, i=1,…,n are the observations and $\widehat{m}_{(-i)h}$ is the leave-one-out estimate (estimated value of m_i without using the i^{th} observation) with h as bandwidth. The optimum bandwidth is the value of h which minimizes the cross validation score $CV(h)$. The estimate, $\widehat{\boldsymbol{\beta}}$ is asymptotically unbiased and its asymptotic variance is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ which is same as the fully parametric model (Opsomer and Ruppert, 1999). Cleveland and Devlin (1988) and Hastie and Tibshirani (1990) discussed the estimation of error variance in linear regression smoothers. An approximate estimate of the error variance is given by

$$\widehat{\sigma}^2=\frac{1}{[n-p-2trace(\mathbf{S})+trace(\mathbf{S}^T\mathbf{S})]}\left[\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\phi}}\right]^T\left[\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\phi}}\right]$$

The variance of $\widehat{\boldsymbol{\beta}}$ is estimated by

$$\widehat{V}\left(\widehat{\boldsymbol{\beta}}\right) = diag(\mathbf{P}\mathbf{P}^T)\hat{\sigma}^2$$

where, $\mathbf{P}= (\mathbf{X}^T(\mathbf{I}\text{-}\mathbf{S})\ \mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}\text{-}\mathbf{S})$. The significance of the covariate effect $\boldsymbol{\phi}$ can be tested using the lack of fit statistic or by comparing the mean residual sum of squares (Hart, 1997; Jose, *et al.*, 2009). Under the null hypothesis that the covariate effect $\phi(\mathbf{U})= \mathbf{0}$, the mean residual sum of squares obtained by fitting the model (1) is given by

$$\hat{\sigma}_0^2 = \mathbf{Y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}]^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}]\mathbf{Y}/(n-p-1)$$

The lack of fit statistic is given by

$$R = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$$

The statistic R asymptotically follows an F distribution with (n-p-1), [n-p-2*trace*($\mathbf{S}$)+ *trace*($\mathbf{S}^T\mathbf{S}$)] degrees of freedom and it can be used for testing the significance of the covariate effect.

3. Analysis of Data in the Presence of Outliers

The regression estimate and the cross validation technique can behave very badly in the presence of outliers in the data or when the errors are heavy-tailed (Leung, D., 2005). One remedy is to remove the influential observations from the data. Another approach is to use robust smoother, which is not as vulnerable as the usual smoothing technique. A robust M-type estimate $\widehat{m}$ of the regression function can be obtained by minimizing the objective function

$$\sum_{i=1}^{n} \rho\left[\frac{y_i - \widehat{m}_i}{s}\right] \tag{2}$$

where, $\rho(.)$ is an even function with bounded first derivative $\psi(.)$ and a unique minimum at zero. The derivative $\psi(x) = \frac{d\rho(x)}{dx}$ is called the influence function and $w(x) = \frac{\psi(x)}{x}$ is the corresponding weight function. Several M-type estimators have been discussed in literature using different types of influence functions (Huber, 1981; Rey, 1983; Hampel *et al.*, 1986; Tukey, 1977). Tuckey's biweight robust function is very popular and it is considered in this paper. The ρ, ψ and w functions corresponding to the Tuckey's robust estimator is given by

$$\rho(x) \quad = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{x}{c}\right)^2\right)\right]^3 & |x| \le c \\ \frac{c^2}{6} & |x| > c \end{cases}$$

$$\psi(x) = \begin{cases} x\left[\left(1 - \left(\frac{x}{c}\right)^2\right)\right]^2 & |x| \le c \\ 0 & |x| > c \end{cases}$$

$$w(x) = \begin{cases} \left[\left(1 - \left(\frac{x}{c}\right)^2\right)\right]^2 & |x| \leq c \\ 0 & |x| > c \end{cases}$$

The turning constant c is picked to give reasonably high efficiency. When the errors are normal and x is the standardized residual, then c=4.685 produce 95% efficiency.

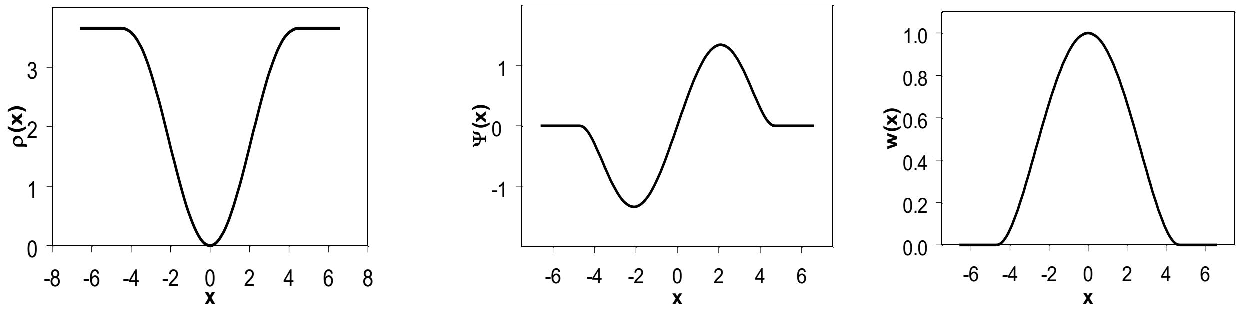

Figure 1: ρ, ψ **and** w **functions of Tuckey's biweight robust estimate**

Iterated reweighted least squares technique is used to solve the minimization problem in eq. (2) to obtain the robust estimate of the regression function. The estimate of the regression function in the k^{th} iteration is given by

$$\widehat{\mathbf{m}}_{(k)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(k)} + \widehat{\boldsymbol{\Phi}}_{(k)}$$

$$\widehat{\boldsymbol{\beta}}_{(k)} = \left[\mathbf{X}^T\left(\mathbf{I} - \mathbf{S}_{(k)}\right)\mathbf{X}\right]^{-1}\mathbf{X}^T\left(\mathbf{I} - \mathbf{S}_{(k)}\right)\boldsymbol{Y}$$

$$\widehat{\boldsymbol{\Phi}}_{(\mathbf{k})} = \mathbf{S}_{(k)}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{(k)}\right)$$

where, $\mathbf{S}_{(k)}$ is the smoothing matrix in the k^{th} iteration derived using robust local linear regression. Let $\mathbf{S}_{i(k)}$be the i^{th} row of the smoothing matrix in the k^{th} iteration, then

$$\mathbf{S}_{(k)} = \left[\mathbf{S}_{1(k)} \dots \mathbf{S}_{n(k)}\right]^T$$

$$\mathbf{S}_{i(k)}^T = \boldsymbol{e}_1^T\left(\mathbf{Z}_{u_i}^T\mathbf{W}_{u_i(k)}^*\mathbf{Z}_{u_i}\right)^{-1}\boldsymbol{Z}_{u_i}^T\boldsymbol{W}_{u_i(k)}^*$$

$$\mathbf{W}_{u_i(k)}^* = diag\left\{w_i^*\left(r_{1(k-1)}\right), \dots, w_i^*\left(r_{n(k-1)}\right)\right\}$$

$$w_i^*\left(r_{j(k-1)}\right) = \frac{K\left(\frac{u_i - u_j}{h}\right)w\left(r_{j(k-1)}\right)}{\sum_{l=1}^n K\left(\frac{u_i - u_l}{h}\right)w\left(r_{l(k-1)}\right)}, \qquad j = 1, \dots, n$$

where $w\left(r_{j(k-1)}\right)$is the value of the robustness weight function corresponding to y_j in the k^{th} iteration and $r_{(k-1)j} = \frac{\left[y_j - \widehat{m}_{j(k-1)}\right]}{s_{(k-1)}}$ is the standardized residual of the j^{th} datum in the (k-1)th iteration with $\widehat{m}_{j(k-1)}$ as the estimated value and $r_{(0)i}$=0 for i=1,…,n. The Median of Absolute Deviation from median (*MAD*) is used for computing a robust estimate for the scale factor s and

$$s_{(k-1)} = \frac{\underset{i}{median}\left|e_{(k-1)i} - \underset{j}{median}(e_{(k-1)j})\right|}{0.6745}$$

where $e_{(k-1)i} = y_i - \widehat{m}_{i(k-1)}$

The estimate of the regression function in the k^{th} iteration is written as

$$\widehat{\mathbf{m}}_{(k)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(k)} + \mathbf{S}_{(k)}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{(k)}\right)$$

Iteration is continued till there is no significant improvement in the estimated values and the final estimate of the regression function is written as

$$\widehat{\mathbf{m}}^* = \mathbf{X}\widehat{\boldsymbol{\beta}}^* + \mathbf{S}^*(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^*)$$
$$\widehat{\boldsymbol{\beta}}^* = [\mathbf{X}^T(\mathbf{I} - \mathbf{S}^*)\mathbf{X}]^{-1}\mathbf{X}^T\,(\mathbf{I} - \mathbf{S}^*)\mathbf{Y}$$
$$V(\widehat{\boldsymbol{\beta}}^*) = diag(\mathbf{P}^*\mathbf{P}^{*T})\hat{\sigma}^{*2}$$

Where $\mathbf{S}^*$ is the smoothing matrix of the final iteration, $\widehat{\mathbf{m}}^*$, $\widehat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}^*$ are the final estimates of the regression function, treatment vector and scale factor respectively and

$$\mathbf{P}^* = (\mathbf{X}^T(\mathbf{I}\text{-}\mathbf{S}^*)\,\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}\text{-}\mathbf{S}^*)$$

Optimum bandwidth: Let $w_i^{\#}$ be the final robustness weight assigned to y_i and $\widehat{m}_{i(h)}^{\#}$ be the estimated value of m_i with bandwidth *h*. The Mean Squared Error (*MSE*) of the estimated value corresponding to the bandwidth *h* is given by

$$MSE(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{m}_{i(h)}^{\#}\right)^2$$

The cross validation score *CV* (*h*) does not work well for the robust smoothers because the *CV* function itself is strongly influenced by the outliers (Wang and Scott, 1994). The cross validation score is the sum of squares of the prediction errors of the smoother at each of the design points. When there are outliers, the prediction errors corresponding to the outliers will be uncharacteristically extreme and these extreme prediction errors will inflate the *CV*(*h*). Therefore, similar to robust smoothing technique, the influence of extreme prediction errors should be minimized. A robust cross validation score *RCV* (*h*) is defined as

$$RCV(h) = \frac{\sum_{i=1}^{n} w_i^{\#}\left(y_i - \widehat{m}_{(-i)(h)}^{\#}\right)^2}{\sum_{i=1}^{n} w_i^{\#}}$$

where, $w_i^{\#}$ is the final robustness weight defined earlier, $\widehat{m}_{(-i)(h)}^{\#}$is the robust estimate of y_i with *h* as bandwidth and without using the i^{th} observation y_i. The value of *h* which minimizes the robust cross validation score *RCV* (*h*) will be the optimum bandwidth. In the computation of *RCV*(*h*), the effect of outliers is controlled by taking weighted sum of squares of the prediction errors of the smoother at each of the design points with the robustness weight $w_i^{\#}$.

4. Simulation Study

A simulation study was conducted to evaluate the performance of the proposed method. The semi-parametric regression model considered for the simulation study is given by

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\phi(\mathbf{U})+\boldsymbol{\varepsilon} \tag{3}$$

where **Y** is the n x 1 observation vector, $\mathbf{m}=\mathbf{X}\boldsymbol{\beta}+\phi(\mathbf{U})$, is the regression function, **X** is the nxk design matrix, $\boldsymbol{\beta}$ is the k x 1 treatment effect vector which is taken as $\boldsymbol{\beta}^T$=[-2 -2 0 4], $\phi(u)=1+2sin(\pi u)$ and the random error vector $\boldsymbol{\varepsilon}$ follows $N(\mathbf{0}, \sigma^2\mathbf{I})$ and uϵ[0,1]. Based on the above, 100 sets of data are simulated for different values of n (100, 200, 400) and σ (1.0, 2.0) with 0%, 4% and 8% outliers. To generate data with specific percentage of outliers, the required number of random numbers between 0 to n are generated and the value of the regression function m corresponding to the data points are replaced with $m+6\sigma$. The Epanechnikov kernel function $K(u)=0.75(1-u^2)$ is employed in the study. The treatment effect vector $\boldsymbol{\beta}^T=[\beta_1\ \beta_2\ \beta_3\ \beta_4]$, the nonparametric function ϕ and the error variance σ^2 are estimated using the method given in Section 2. Tuckey's biweight function with the turning point c=4.685 is used as the robustness function. The Average Mean Squared Errors (*AMSE*) of the estimated values of σ, $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and **m** with the true values of 100 sets of simulated data for different values of n (100, 200, 400) and σ (1.0, 2.0) are given in Table 2. The *AMSE* of the estimated parameters are calculated as follows:

$$AMSE \text{ of } \hat{\sigma}=\frac{1}{100}\sum_{i=1}^{100}\left(\sigma-\hat{\sigma}_{(i)}\right)^2,$$

$$AMSE \text{ of } \hat{\boldsymbol{\beta}}=\sum_{j=1}^{4}\frac{1}{100}\sum_{i=1}^{100}\left(\beta_j-\hat{\beta}_{j(i)}\right)^2$$

$$AMSE \text{ of } \hat{\boldsymbol{\phi}}=\frac{1}{100}\sum_{i=1}^{100}\frac{1}{n}\sum_{j=1}^{n}\left[\phi(u_j)-\hat{\phi}_{(i)}(u_j)\right]^2,$$

$$AMSE \text{ of } \hat{\mathbf{m}}=\frac{1}{100}\sum_{i=1}^{100}\frac{1}{n}\sum_{j=1}^{n}\left[m-\hat{m}_{(i)}(u_j)\right]^2$$

where, $\hat{\sigma}_{(i)}$, $\hat{\beta}_{j(i)}$, $\hat{\phi}_{(i)}$ and $\hat{m}_{(i)}$ are the estimated values of σ , β_j, ϕ and the regression function m corresponding to the i^{th} simulated data set. The bias of the point estimates of $\hat{\sigma}, \hat{\beta}_j$, j=1,…,4 are calculated as follows

$$\text{Bias of } \hat{\sigma}=\frac{1}{100}\sum_{i=1}^{100}\left(\sigma-\hat{\sigma}_{(i)}\right)$$

$$\text{Bias of } \hat{\beta}_j=\frac{1}{100}\sum_{i=1}^{100}\left(\beta_j-\hat{\beta}_{j(i)}\right),\ j=1,\dots,4$$

The *AMSE* of the estimates are converging to zero as n increases or in other words, the estimated values are converging to the true values as n increases. Note that the bias of the point estimates $\hat{\sigma}, \hat{\beta}_j$, j = 1,…,4 are also negligible as n increases (Table 2). This indicates the consistency of the estimates. The *MSE* varies with change in the choice of bandwidths. The optimum bandwidth (bandwidth corresponds to the minimum *MSE*) depends on the curvature of the function. The optimum bandwidth for estimating the parameters of the model was obtained based on the robust cross validation technique given in Section 2.

The comparison of Average Mean Squared Errors (*AMSE*) of the estimated values of σ, $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and **m** with the true values of 100 sets of simulated data for different values of *n* (100, 200, 400) and σ (1.0, 2.0) showed that in the presence of outliers (4% and 8%) the robust method performs much better than the non-robust method. In the absence of outliers the performance of both the robust and non-robust methods is almost the same. The value of *AMSE* decreases as *n* increases or in other words the estimates converges to the true value.

5. Application

The proposed method is used to analyze the yield data (both weight of nuts and number of nuts) of arecanut recorded in an experiment to evaluate the effects of organic and inorganic fertigation in arecanut + cocoa systems conducted at Central Plantation Crops Research Institute, Vittal, India. The experiment consists of 6 levels of nutrition (denoted as $\beta_1, \ldots, \beta_6$) applied randomly to 12 year old arecanut palms to evaluate their effect on the yield of arecanut. Treatments were applied to a total of 253 arecanut palms. Pre-treatment yield was taken as the covariate to control the error due to palm to palm variation. The yield obtained after a gap of two year from the start of the experiment was taken as the response variable. The estimated treatment effects and its standard errors using the linear covariance technique, nonparametric covariance technique as well as the robust nonparametric covariance technique with pre-treatment yield as covariate for both weight of nuts and number of nuts are given in Table 3 and 4 respectively. Even though, there is not much difference in the estimated value of the treatment effects employing different methods, the standard error of the estimates are comparatively lower in the case of the proposed robust technique. Since the outliers are present in both the extremes (high and low) and the number of observations is also high, the difference in treatment effects estimated using robust and other methods are very less. The estimated value of σ corresponding to the linear, nonparametric and robust nonparametric covariance technique for weight of nuts and number of nuts are also given in Table 3 and 4 respectively. The standard errors of the estimates and the estimated value of σ are less in the proposed robust method than that of the linear and nonparametric covariance technique.

6. Conclusion

Linear covariance technique is generally used for analysing the designed experiments having covariates, assuming a linear relationship between response and covariate. A more flexible semi-parametric model is used when the relationship is not linear or unknown. The experimental data particularly, those from field experiments generally contains some extreme values or outliers due to large plant to plant variations and their presence very badly affect the analysis and generate distorted results. In the present study, a robust method is proposed to analyse the semi-parametric regression model in the presence of outliers. The proposed method is useful when the data contains extreme values or outliers and there is no advance information about the relationship between the response variable and covariate.

References

Buja, A. Hastie, T.J and Tibshirani, R.J. (1989). Linear smoothers and additive models. *Annals of Statistics*, **17,** 453-555.

Cleveland, William S. and Devlin, Susan J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**(**403**), 596–610.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics-The Approach Based on Influence Functions*. New York: John Wiley and Sons.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Verlag, New York.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall, London.

Huber, P.J. (1981). *Robust Statistics*. John Wiley and Sons, New York.

Jose, C.T. and Ismail, B. (2001). Nonparametric inference on jump regression surface. *Journal of Nonparametric Statistics*, **13**, 791-813.

Jose, C.T., Bhat Ravi, Ismail, B. and Jayasekhar, S. (2009). Spatial smoothing technique in field experiments. *Journal of the Indian Society of Agricultural Statistics*, **63**, 151-157.

Leung, D. (2005). Cross-validation in nonparametric regression with outliers. *The Annals of Statistics* **33**, 2291-2310.

Opsomer, J.D. and Ruppert, D. (1999). A root-*n* consistent estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, **8**, 715-732.

Rey, W.J.J. (1983). *Introduction to Robust and Quasi-robust Statistical Methods*. Springer-Verlag, Heidelberg.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346-70.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Appendix

Table 1: Optimum bandwidth ad *AMSE* of the estimates in the simulation study

σ	Outliers (%)	n	h	$AMSE(\hat{\boldsymbol{\beta}})$		$AMSE(\hat{\boldsymbol{\phi}})$		$AMSE(\hat{\mathbf{m}})$		$AMSE(\hat{\sigma})$	
				SP	Robust SP	SP	Robust SP	SP	Robust SP	SP	Robust SP
1	0	100	0.30	0.1486	0.1471	0.0796	0.0541	0.0731	0.0664	0.0064	0.0182
		200	0.20	0.0720	0.0740	0.0309	0.0256	0.0398	0.0389	0.0030	0.0078
		400	0.15	0.0315	0.0341	0.0167	0.0148	0.0215	0.0218	0.0011	0.0033
	4	100	0.30	0.3712	0.1741	0.0946	0.0608	0.1835	0.0810	0.3194	0.0143
		200	0.25	0.1656	0.0849	0.0522	0.0308	0.1041	0.0408	0.3089	0.0080
		400	0.15	0.0774	0.0384	0.0459	0.0162	0.0944	0.0241	0.2948	0.0038
	8	100	0.25	0.5138	0.1793	0.2440	0.0594	0.4018	0.0846	0.8825	0.0180
		200	0.25	0.2285	0.0985	0.1704	0.0285	0.2723	0.0439	0.8511	0.0133
		400	0.25	0.1250	0.0548	0.1440	0.0179	0.2222	0.0231	0.8331	0.0121
2	0	100	0.30	0.5399	0.5672	0.1913	0.1591	0.2631	0.2588	0.0188	0.0644
		200	0.30	0.3009	0.3142	0.1144	0.0887	0.1451	0.1452	0.0101	0.0415
		400	0.25	0.1463	0.1530	0.0543	0.0401	0.0689	0.0674	0.0043	0.0125
	4	100	0.30	1.4690	0.6348	0.3745	0.1797	0.7428	0.2981	1.2803	0.0432
		200	0.30	0.5670	0.2865	0.2134	0.0931	0.3958	0.1386	1.2292	0.0281
		400	0.20	0.3239	0.1418	0.2455	0.0509	0.3576	0.0810	1.1912	0.0224
	8	100	0.30	1.7699	0.6486	1.0668	0.1936	1.6062	0.3138	3.3692	0.0716
		200	0.30	0.9476	0.3266	0.6910	0.0900	1.0818	0.1394	3.3647	0.0584
		400	0.25	0.4338	0.1461	0.7617	0.0442	0.9799	0.0709	3.3693	0.0576

SP: Semi- parametric

Table 2: Bias of the robust point estimates in the simulation study

σ	Outliers (%)	n	h	Bias of				
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}$
1	0	100	0.30	–0.0184	–0.0187	0.0201	0.0169	–0.0610
		200	0.20	–0.0182	–0.0157	0.0155	0.0185	–0.0282
		400	0.15	0.0026	–0.0053	0.0127	–0.0100	–0.0187
	4	100	0.30	–0.0232	0.0100	–0.0010	0.0143	0.0440
		200	0.25	–0.0161	0.0112	–0.0002	0.0051	0.0346
		400	0.15	0.0033	0.0019	–0.0004	–0.0048	0.0294
	8	100	0.30	–0.0002	0.0166	–0.0077	–0.0088	0.0851
		200	0.25	–0.0008	0.0093	–0.0084	0.0000	0.0789
		400	0.20	0.0010	0.0039	–0.0058	0.0010	0.0745
2	0	100	0.30	–0.0060	–0.0079	0.0055	0.0085	–0.1037
		200	0.25	–0.0090	0.0234	–0.0156	0.0013	–0.0586
		400	0.20	–0.0030	0.0040	–0.0060	0.0051	–0.0398
	4	100	0.30	0.0222	–0.0271	–0.0120	0.0168	–0.0558
		200	0.25	0.0131	–0.0135	–0.0110	0.0114	0.0490
		400	0.20	–0.0060	0.0184	–0.0088	–0.0037	0.0445
	8	100	0.30	–0.0132	0.0294	–0.0131	–0.0031	0.1248
		200	0.30	–0.0143	0.0292	–0.0061	–0.0089	0.1160
		400	0.25	–0.0117	0.0202	–0.0106	0.0020	0.1119

Table 3: Estimated values with standard errors (weight of nuts) of the field data

Parameter	Linear		Semi-parametric		Robust Semi-parametric	
	Estimate	SE	Estimate	SE	Estimate	SE
$\mu + \beta_1$	9.969	0.683	9.924	0.622	9.925	0.548
$\mu + \beta_2$	9.414	0.683	9.570	0.626	9.573	0.552
$\mu + \beta_3$	10.029	0.638	9.949	0.594	9.950	0.524
$\mu + \beta_4$	9.883	0.675	9.994	0.617	9.991	0.543
$\mu + \beta_5$	9.922	0.691	9.918	0.636	9.916	0.560
$\mu + \beta_6$	10.767	0.630	10.758	0.587	10.758	0.517
σ	4.317		4.312	-	3.803	-

μ: Overall mean

Table 4: Estimated values with standard errors (number of nuts) of the field data

Parameter	Linear		Semi-parametric		Robust Semi-parametric	
	Estimate	SE	Estimate	SE	Estimate	SE
$\mu + \beta_1$	328.80	22.71	331.96	20.71	330.80	16.83
$\mu + \beta_2$	307.84	22. 70	308.87	20.70	308.85	16.83
$\mu + \beta_3$	331.13	21.12	334.87	19.62	336.45	15.95
$\mu + \beta_4$	332.86	22.40	336.57	20.45	337.32	16.63
$\mu + \beta_5$	324.32	22.91	315.21	21.27	313.69	17.29
$\mu + \beta_6$	370.57	20.87	374.34	19.35	374.71	15.72
σ	143.06		142.96		116.16	

μ: Overall mean