Working Paper 95-04
Statistics and Econometrics Series 02
February 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

ON THE BEHAVIOUR OF RESIDUAL PLOTS IN REGRESSION

Santiago Velilla*

Abstract

Properties of least squares versus robust regression residual plots are compared under a common set of assumptions.

Key Words

Bias of a robust estimator; Extreme points; Partially modified residuals; Outliers.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

1. INTRODUCTION

Consider $z_i = (y_i, x_i')'$, $i = 1, \ldots, n$, independent and identically distributed observations, where y_i is a scalar and x_i is a pxl vector. The classical linear regression model with random regressors assumes that

$$y_i = \alpha + x_i'\beta + u_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where u_i is an error independent of x_i and $(\alpha, \beta')'$ is a $(p+1)$xl vector of unknown parameters. The observations z_i arise from a "central" model $H_0(y,x)$ characterized by a marginal distribution $G_0(x)$ for the x_i, and a zero mean marginal distribution $F_0(u/\sigma_0)$ for the u_i, where σ_0 is a scale parameter. For example $F_0 \equiv N(0,\sigma_0)$.

Write $B = (\alpha, \beta')'$ and let $\hat{B} = (\hat{\alpha}, \hat{\beta}')' = \underset{B}{\operatorname{argmin}} \sum_{i=1}^{n} e_i^2(B)$ be the least squares estimator of B, where $e_i(B) = y_i - \alpha - x_i'\beta$. The least squares residuals

$$\hat{e}_i = e_i(\hat{B}) = y_i - \hat{\alpha} - x_i'\hat{\beta}, \tag{1.2}$$

$i = 1, \ldots, n$, are used to check the adequacy of the model and to detect unusual patterns in the data. The behaviour of the residuals is studied conditioning on the observed value of the nxp design matrix $X = (x_1, \ldots, x_n)'$. Residual analysis is generally conducted in a graphic way. The idea is to plot the residuals against any other quantity orthogonal to them, generally the fitted values $\hat{y}_i = \hat{\alpha} + x_i'\hat{\beta}$, $i = 1, \ldots, n$, such that, under the null of a correctly specified parametric model, the expected behaviour of the plot contains no visible pattern. Observed patterns are then attributed to inappropriate assumptions. See Cook and Weisberg (1982), Weisberg (1984) or Atkinson (1985) for details.

It is well known that $\hat{B}$ is very sensitive to both outliers in the nxl vector of responses $Y = (y_1, \ldots, y_n)'$ and extremes in the rows of X and, therefore, several alternative robust estimators $\tilde{B}_n = (\tilde{\alpha}_n, \tilde{\beta}_n')'$ have been proposed. Plotting the residuals

$$\tilde{e}_i = \tilde{e}_i(\tilde{B}_n) = y_i - \tilde{\alpha}_n - x_i'\tilde{\beta}_n, \tag{1.3}$$

is advocated by some authors, among others Rousseeuw and Leroy (1987, p. 92-93) as an after-fit diagnostic tool. The question is if residual plots based in robust estimators can be interpreted in the same form as the standard least squares residual plots. This seems to be the case for R estimators and for M estimators with monotone ψ function as shown by McKean,

Sheather and Hettmansperger (1990, 1993). However, some complications arise when using high breakdown point estimators. This is reported by Cook, Hawkins and Weisberg (1992) and McKean et al. (in press a) for the case of the least median of squares (LMS) estimator of Rousseeuw (1984), and by McKean et al. (1993) for the case of GM estimators.

Properties of least squares residuals are studied under the assumption that the carriers $\mathbf{X}$ are fixed or equivalently, as stated above, conditioning on $\mathbf{X}$ in the model (1.1). An assumption in robust regression is that (1.1) holds only approximately and, therefore, the joint distribution $H(y,x)$ of the observations is different but close, in some way, to the central model $H_0(y,x)$. As a consequence, while in linear least squares the carriers $\mathbf{x}_i$ are considered as known constants, the robust approach takes into account the stochastic nature of the $\mathbf{x}_i$ in the appearance of dubious data. Comparison of the properties of the vectors of residuals $\hat{\mathbf{e}} = (\hat{e}_1, \ldots, \hat{e}_n)'$ and $\tilde{\mathbf{e}} = (\tilde{e}_1, \ldots, \tilde{e}_n)'$ is then at issue because each vector is obtained under a different set of premises. The aim of this paper is to study, under a common framework of assumptions, some aspects of the comparison between plots based on $\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}$. Notation is established in section 2 while section 3 contains the main results. Section 4 illustrates the theory with two real data examples and section 5 concludes with some final comments.

2. BACKGROUND AND MOTIVATION

2.1 Model

For the purpose of comparison of this paper, observations z_i, $i = 1, \ldots, n$, are assumed to be of the form

$$z_i = (1-\varepsilon_i)z_{i0} + \varepsilon_i z_{i1}, \tag{2.1}$$

where $z_{i0} = (y_{i0}, x_{i0}')'$ and $z_{i1} = (y_{i1}, x_{i1}')'$ are $(p+1)\times 1$ random vectors, and ε_i is a random variable taking values 0 and 1 with probabilities $1-\varepsilon$ and ε $(0 \le \varepsilon \le .5)$. The triplets $(\varepsilon_1, z_{i0}, z_{i1})$ are i.i.d. and unobservable and ε_i is independent of the pair (z_{i0}, z_{i1}). z_{i0} arises from the "central" model $H_0(y,x)$ of (1.1) and z_{i1} is an outlying component with distribution $H_1(y,x)$ that, in principle, can have any form. To get meaningful results, it will be assumed that $H_1(y,x)$ is characterized, as $H_0(y,x)$ in (1.1), by a marginal distribution $G_1(x)$ for the x_i and a relation between the response and the regressors of the form

$$y = \alpha + x'\beta + u, \tag{2.2}$$

where (x,u) are independent and u has a zero mean distribution $F_1(u/\sigma_1)$ depending on an unknown scale parameter σ_1. The setup (2.1)-(2.2) is quite flexible and allows to model data anomalies only in the response ($G_1 \equiv G_0$), only in the predictors ($F_1 \equiv F_0$, $\sigma_1 = \sigma_0$), or on both parts of the observation at the same time. Recall that α and β identical for H_0 and H_1 so that both the central model and the perturbation keep the *same linear structure* in the conditional mean. Anomalous points are "identified" by the nonzero coordinates of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. (2.1) is inspired by a general contamination model in time series proposed by Martin and Yohai (1986).

It is clear that, under (2.1)-(2.2), the common distribution H of the z_i is a mixture of the form

$$H(y,x) = (1-\varepsilon)H_0(y,x) + \varepsilon H_1(y,x), \tag{2.3}$$

which is a particular case of distribution in the ε-contamination neighborhood of Tukey

$$\mathcal{H}_\varepsilon = \{G:\ G = (1-\varepsilon)H_0 + \varepsilon H^*,\ H^* \text{ arbitrary}\},$$

under which much work in robust regression has been done. The following results clarify the meaning of (2.1)-(2.2). All proofs are given in the appendix.

Proposition 2.1. Consider model (2.1)-(2.2) for i.i.d. observations $z_i = (y_i, x_i')'$, $i = 1, \ldots, n$, and suppose that both $G_0(x)$ and $G_1(x)$ have densities denoted, respectively, by $g_0(x)$ and $g_1(x)$ with respect to a common dominating measure m(dx) in $\mathbb{R}^p$. Then:

a) Conditionally on $\mathbf{X} = (x_1, \ldots, x_n)'$, the components of $\mathbf{Y} = (y_1, \ldots, y_n)'$ are independent with conditional distribution functions

$$F(y_i \mid \mathbf{X}) = F(y_i \mid x_i) = [1-\varepsilon(x_i)]F_0(q_i/\sigma_0) + \varepsilon(x_i)F_1(q_i/\sigma_1),$$

where

$$\varepsilon(x_i) = \frac{\varepsilon g_1(x_i)}{(1-\varepsilon)g_0(x_i) + \varepsilon g_1(x_i)}$$

and

$$q_i = y_i - \alpha - x_i'\beta.$$

The $\varepsilon(x_i)$ are i.i.d. with mean ε and variance bounded by $\varepsilon(1-\varepsilon)$;

b) $E[\mathbf{Y} \mid \mathbf{X}] = \mathbf{WB}$, where $\mathbf{W} = (1_n \mid \mathbf{X})$;

c) $V[\mathbf{Y} \mid \mathbf{X}] = D(\mathbf{X})$, where $D(\mathbf{X}) = \mathrm{diag}(d_1(\mathbf{X}), \ldots, d_n(\mathbf{X}))$ is an nxn diagonal matrix such that $E[D(\mathbf{X})] = \sigma^2 I_n$, with $\sigma^2 = (1-\varepsilon)\sigma_0^2 + \varepsilon\sigma_1^2$.

The main implications of proposition 2.1 can be summarized as follows.

Proposition 2.2. Under (2.1)-(2.2):

a) $(q_1, \ldots, q_n)'$ are i.i.d. with zero mean and variance σ^2;

b) Conditionally on $\mathbf{X}$, $(q_1, \ldots, q_n)'$ are: (i) independent with zero mean but (ii) not identically distributed if G_0 and G_1 are distinct. Therefore, the models (1.1) and (2.1)-(2.2) are, in general, different.

Remark. It might seem tempting trying to generalize model (2.1)-(2.2) by assuming an structure for $H_j(y,x)$ of the form

$$H_j(y,x) = \int_{u \le x} F_j[(y - \alpha_j - u'\beta_j)|\sigma_j]\, G_j(du), \tag{2.4}$$

for $j = 0,1$, where $(\alpha_0, \beta_0')'$ differs from $(\alpha_1, \beta_1')'$. (2.4) implies

$$E[y_i|x_i] = \alpha(x_i) + x_i'\beta(x_i),$$

with intercept and slope parameters

$$\alpha(x_i) = [1-\varepsilon(x_i)]\alpha_0 + \varepsilon(x_i)\alpha_1,$$

$$\beta(x_i) = [1-\varepsilon(x_i)]\beta_0 + \varepsilon(x_i)\beta_1.$$

Also, $\mathrm{var}[y_i|x_i] = d_i(\mathbf{X}) + [1-\varepsilon(x_i)]\,(\alpha_0 + x_i'\beta_0)^2 + \varepsilon(x_i)\,(\alpha_1 + x_i'\beta_1)^2 - [\alpha(x_i) + x_i'\beta(x_i)]^2$. The intuitive basis of (2.4) is then counterbalanced by the untractability of the latter expressions which suggests following the discussion under the framework (2.1)-(2.2).

2.2 Estimators

Estimators considered in this paper will be of the form

$$\tilde{\mathbf{B}}_n = T_n[z_1, \ldots, z_n] = T[H_n], \tag{2.5}$$

where T is a functional defined on a space of distributions in $\mathbb{R}^{p+1}$ and H_n is the empirical distribution function of the sample $z_1, z_2, \ldots, z_n$. If the estimator is both consistent and Fisher consistent at the central model then, under H_0, $\tilde{\mathbf{B}}_n$ converges to $T[H_0] = B$. However, for general H in $\mathcal{H}_\varepsilon$, $T[H]$ will be different from B. In large samples, a suitable measure of the robustness of $\tilde{\mathbf{B}}_n$ over the neighborhood $\mathcal{H}_\varepsilon$ is the curve of maximum asymptotic bias (Martin, Yohai and Zamar (1989))

$$B(T,H_0,\varepsilon) = \max\{b_M(T,H):\ H \in \mathcal{H}_\varepsilon\}, \tag{2.6}$$

where $b_M(T,H) = \|T[H]-B\|_M$, $\|a\|_M = (a'Ma)^{1/2}$ and M is a $(p+1)\times(p+1)$ positive definite matrix properly chosen. An example illustrates the concept of asymptotic bias $T[H] - B$.

Example 2.1 Let $\tilde{H}(y,x)$ be a general distribution in $\mathbb{R}^{p+1}$. Following Hinkley (1977), the functional defining the least squares estimator of B is

$$LS[\tilde{H}] = \begin{pmatrix} 1 & \mu' \\ \mu & S \end{pmatrix}^{-1} E[\begin{pmatrix} y \\ xy \end{pmatrix}],$$

where $S = E[xx']$ and $\mu = E[x]$ (All expectations are taken with respect to $\tilde{H}$). For $H(y,x)$ as in (2.1)-(2.2), proposition 2.1 yields $E[y] = \alpha + E[x']\beta$ and $E[xy] = E[x]\alpha + E[xx']\beta$ and, therefore,

$$LS[H] = \begin{pmatrix} 1 & E[x'] \\ E[x] & E[xx'] \end{pmatrix}^{-1} \begin{pmatrix} 1 & E[x'] \\ E[x] & E[xx'] \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = B,$$

so the least squares estimator has no asymptotic bias. Observe also that, by the same arguments, $LS[H_0] = LS[H_1] = B$ and, in particular, the least squares estimator is also Fisher consistent. The functional for an M estimator is defined implicitly by the solution of the equation

$$E[w\psi[(y-w'M[\tilde{H}])/\sigma]] = 0,$$

where ψ is an odd and bounded function, $w = (1, x')'$ and σ is a scale parameter. If F_0 is symmetric about the origin, $M[H_0] = B$ and the M estimator is Fisher consistent. However, if $F_1(u/\sigma_1)$ is not symmetric about zero, the expectation

$$E[w\psi[(y-w'B)/\sigma]] = E[m\varepsilon(x)\begin{pmatrix} 1 \\ x \end{pmatrix}] = m\varepsilon\begin{pmatrix} 1 \\ \int xG_1(dx) \end{pmatrix},$$

where $m = \int\psi(u/\sigma)F_1(du/\sigma_1)$, is, in general, not null and, as a consequence, M[H] is different from B. ■

For the estimators (2.5), it will be assumed that the representation below holds:

$$\tilde{B}_n = T[H_n] = T[H] + n^{-1}\sum_{j=1}^{n} EIC(z_j) + o_p(n^{-1/2}), \qquad (2.7)$$

where $EIC(z_j)$ is an empirical version of the influence curve of the estimator at z_j (see, for example, Hampel, Ronchetti, Rousseeuw and Stahel 1986, p. 85 and 93). The framework (2.5)-(2.7) includes least squares estimators, M estimators and GM estimators. However, it excludes the high breakdown point LMS estimator whose asymptotic convergence rate is $n^{1/3}$.

3. RESIDUALS AND ASYMPTOTIC BIAS

3.1 Least squares residuals

The least squares estimator is $\hat{\mathbf{B}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}$, where, as in proposition 2.1.b), $\mathbf{W} = (1_n|\mathbf{X})$. Write the vector of least squares residuals $\hat{\mathbf{e}} = (\hat{e}_1, \ldots,\hat{e}_n)'$ and the vector of fitted values $\hat{\mathbf{Y}} = (\hat{y}_1, \ldots,\hat{y}_n)' = \mathbf{W}\hat{\mathbf{B}}$ in the form

$$\hat{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \qquad \hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}, \tag{3.1}$$

where $\mathbf{H}$ is the nxn orthogonal projection matrix $\mathbf{H} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$. From proposition 2.1 it easily follows that, under (2.1)-(2.2):

a) $E[\hat{\mathbf{e}}|\mathbf{X}] = 0$, $V[\hat{\mathbf{e}}|\mathbf{X}] = (\mathbf{I}_n - \mathbf{H})D(\mathbf{X})(\mathbf{I}_n - \mathbf{H})$;
b) $E[\hat{\mathbf{Y}}|\mathbf{X}] = \mathbf{W}\mathbf{B}$, $V[\hat{\mathbf{Y}}|\mathbf{X}] = \mathbf{H}D(\mathbf{X})\mathbf{H}$, and (3.2)
c) $C[(\hat{\mathbf{e}},\hat{\mathbf{Y}})|\mathbf{X}] = (\mathbf{I}_n - \mathbf{H})D(\mathbf{X})\mathbf{H}$.

where $D(\mathbf{X})$ is as in proposition 2.1.c). A first approximation for the $\varepsilon(x_i)$ is $\varepsilon(x_i) \cong E[\varepsilon(xi)] = \varepsilon$ or, equivalently, $D(\mathbf{X}) \cong \sigma^2\mathbf{I}_n$. From the group of expressions (3.2), this yields

$$V[\hat{\mathbf{e}}|\mathbf{X}] \cong \sigma^2(\mathbf{I}_n - \mathbf{H}) \quad , \quad C[(\hat{\mathbf{e}},\hat{\mathbf{Y}})|\mathbf{X}] \cong 0$$

and, for practical purposes, the usual properties and interpretation of plots of least squares residuals versus fitted values should be expected to hold under the setup (2.1)-(2.2).

3.2 Robust residuals

Properties of least squares residuals can be derived explicitly because of the tractability of expressions (3.1). In contrast, for a general robust estimator in the framework (2.5)-(2.7), an approximate method of analysis seems necessary. Put $w_i = (1, x_i')'$. For $\tilde{\mathbf{B}}_n = T[H_n]$, the residuals $\tilde{e}_i = y_i - w_i'\tilde{\mathbf{B}}_n$ and the fitted values $\tilde{y}_i = w_i'\tilde{\mathbf{B}}_n$ can be decomposed

$$\tilde{e}_i = y_i - w_i'\tilde{\mathbf{B}}_n = q_i - w_i'(\tilde{\mathbf{B}}_n - T[H]) - w_i'(T[H] - B),$$

$$\tilde{y}_i = w_i'\tilde{\mathbf{B}}_n = w_i'(\tilde{\mathbf{B}}_n - T[H])) + w_i'T[H], \tag{3.3}$$

where the $q_i = y_i - w_i'B$ are the variables of proposition 2.1. Inserting the representation (2.7) into (3.3) then, to first order,

$$\tilde{e}_i \cong q_i - w_i'[n^{-1}\sum_{j=1}^{n} EIC(z_j)] - w_i'(T[H] - B),$$

$$\tilde{y}_i \cong w_i'[n^{-1}\sum_{j=1}^{n} EIC(z_j)] + w_i'T[H]. \tag{3.4}$$

Observe that the (eventually) nonnull asymptotic bias term $T[H] - B$ appears

in the right hand side of expressions (3.3) and (3.4). Put $\tilde{\mathbf{Y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$. The next result follows from proposition 2.1.

Proposition 3.1. Under (2.1)-(2.2), approximation (3.4) yields:

a) $E[\tilde{\mathbf{e}}|\mathbf{X}] \cong -\mathbf{W}a_n$ and $E[\tilde{\mathbf{Y}}|\mathbf{X}] \cong \mathbf{W}(B+a_n)$, where $a_n = n^{-1}\sum_{j=1}^{n} E[EIC(z_j)|\mathbf{X}] + (T[H]-B)$;

b) $V[\tilde{\mathbf{Y}}|\mathbf{X}] \cong \tilde{\mathbf{H}}_n = \mathbf{W}V[n^{-1}\sum_{j=1}^{n} E[EIC(z_j)|\mathbf{X}]\mathbf{W}'$;

c) $V[\tilde{\mathbf{e}}|\mathbf{X}] \cong [D(\mathbf{X})-\tilde{\mathbf{H}}_n] + \tilde{\mathbf{C}}_n$, where $\tilde{\mathbf{C}}_n = 2\tilde{\mathbf{H}}_n-(\tilde{\Gamma}_n+\tilde{\Gamma}_n')$ and $\tilde{\Gamma}_n$ is an nxn matrix of the form $\tilde{\Gamma}_n = n^{-1}\sum_{j=1}^{n}\tilde{\Gamma}_{nj}$, for $\tilde{\Gamma}_{nj}=C[(\mathbf{Y},EIC(z_j))|\mathbf{X}]\mathbf{W}'$, $j = 1, \ldots,n$;

d) $C[(\tilde{\mathbf{e}},\tilde{\mathbf{Y}})|\mathbf{X}] \cong \tilde{\Gamma}_n - \tilde{\mathbf{H}}_n$.

Remark. It is easy to see that application of approximation (3.4) in the case of the least squares estimator goes back to expressions (3.1). Following Hinkley (1977), an empirical version of the influence curve for the least squares estimator LS[H] is $EIC(z_j) = n(\mathbf{W}'\mathbf{W})^{-1}w_j(y_j - w_j'B)$. Since LS[H] = B and $\mathbf{W} = (w_1, \ldots,w_n)'$ the right hand sides of (3.4) are simply $\mathbf{Y} - \mathbf{HY} = \hat{\mathbf{e}}$ and $\mathbf{HY} = \hat{\mathbf{Y}}$.

Some relevant comments on (3.4) and proposition 3.1 are: (*i*) From (3.4), one can write

$$\tilde{\mathbf{e}} \cong \hat{\mathbf{e}} + \mathbf{W}\gamma_n, \tag{3.5}$$

where $\gamma_n = \hat{\mathbf{B}}_n - [n^{-1}\sum_{j=1}^{n} EIC(z_j)] - T[H]$. (3.5) shows how in the sampling decomposition

$$\mathbf{Y} = \mathbf{W}\tilde{\mathbf{B}}_n + \tilde{\mathbf{e}},$$

the residuals might retain information on the carriers **X**, a phenomenon not desirable in residual analysis; (*ii*) $E[\tilde{\mathbf{e}}|\mathbf{X}]$ is, approximately, a nonnull vector in the linear manifold spanned by the matrix **W**. This vector depends on the asymptotic bias T[H] - B; (*iii*) $C[(\tilde{\mathbf{e}},\tilde{\mathbf{Y}})|\mathbf{X}]$ is, in general, not zero and, therefore, the interpretation of a residual plot based on robust estimators might be complicated by a nonorthogonal association pattern between $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{Y}}$; (*iv*) $V[\tilde{\mathbf{e}}|\mathbf{X}]$ is obtained from the matrices $\tilde{\mathbf{H}}_n$ and $\tilde{\Gamma}_n$ whose expressions depend on concrete specifications of both the model (2.1)-(2.2)

and the functional estimator T[.]. A suitable standarization of the residuals is difficult to define.

3.3 Simulation results

A small simulation study is conducted to illustrate the comments above. Besides the least squares estimator (LS) $\hat{\mathbf{B}}$, three different robust estimators of **B** are used:

a) The M-estimator $\tilde{\mathbf{B}}_M$ based on the Huber ψ function (Huber 1973), defined as solution of the estimating equation

$$\sum_{i=1}^{n} \psi[(y_i - w'B)/\sigma]x_i = 0, \tag{3.6}$$

where $\psi(t) = t \min[1, k/|t|]$ and $k = 1.345$. In (3.6), the scale parameter σ is replaced by the median absolute deviation (MAD) $\tilde{\sigma}_0 = 1.483 \underset{i}{\text{med}} \left| \tilde{e}_{0,i} - \underset{j}{\text{med}}\, \tilde{e}_{0,j} \right|$ computed from an initial set $\tilde{e}_{0,i} = y_i - w_i'\tilde{\mathbf{B}}_0$ of residuals;

b) The three-step estimator $\tilde{\mathbf{B}}_{GMM}$ of Simpson, Ruppert and Carroll (1992) based on the Hampel ψ function with bends at $a = 1.5$, $b = 3$ and $c = 8$, and constructed by iterating the one-step scoring relation

$$\tilde{\mathbf{B}} = \tilde{\mathbf{B}}_0 + \mathbf{M}_0^{-1}\mathbf{g}_0,$$

where $\tilde{\mathbf{B}}_0$ is the LMS estimator, $\mathbf{g}_0 = \tilde{\sigma}_0 \sum_{i=1}^{n} \omega_i \psi(\tilde{e}_{0,i}/\tilde{\sigma}_0)x_i$, $\mathbf{M}_0 = n^{-1} \sum_{i=1}^{n} \psi'(\tilde{e}_{0,i}/\tilde{\sigma}_0)\mathbf{W}'\Omega\mathbf{W}$, and the matrix Ω is an nxn diagonal matrix of weights $\text{diag}(\omega_1, \ldots, \omega_n)$. The weights are given by

$$\omega_i = \min\left\{ 1, \left[\frac{c}{(x_i - \tilde{\mu}_n)'\tilde{\Sigma}_n(x_i - \tilde{\mu}_n)} \right] \right\}, \tag{3.7}$$

where c is the 95% percentile of a χ^2_p distribution and the pair $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ is taken from the minimum volume ellipsoid (MVE) estimator for the sample of regressors $x_1, \ldots, x_n$. In this paper, the exact algorithm proposed in Cook et al. (1993) is used to compute the MVE. $\tilde{\mathbf{B}}_{GMM}$ has 50% breakdown point and bounded influence function in the x and y spaces.

c) The one-step estimator $\tilde{\mathbf{B}}_{GMS}$ of Coakley and Hettmansperger (1993) based on the Huber ψ function of (3.6) and given by

$$\tilde{\mathbf{B}} = \tilde{\mathbf{B}}_0 + \mathbf{M}_S^{-1}\mathbf{g}_S,$$

where $\tilde{\mathbf{B}}_0$ is as in (3.7), $\mathbf{g}_S = \tilde{\sigma}_0 \sum_{i=1}^{n} \omega_i \psi(\tilde{e}_{0,i}/(\omega_i\tilde{\sigma}_0))x_i$, $\mathbf{M}_S = \mathbf{W}'\Pi\mathbf{W}$, where $\Pi = \text{diag}(\pi_1, \ldots, \pi_n)$, and $\pi_i = \psi'(\tilde{e}_{0,i}/(\omega_i\tilde{\sigma}_0))$. The weights (ω_i) are as in

(3.7). $\tilde{\mathbf{B}}_{GMS}$ has 50% breakdown point and bounded influence function in both x and y.

To simplify matters, the discussion is centered on a simple linear regression model

$$E[y|x] = \alpha + x\beta,$$

where both the response and the regressor are scalar so that p = 1. N = 200 samples of size n = 40 are generated through a model of the form

$$H(y,x) = (1-\varepsilon)H_0(y,x) + \varepsilon H_1(y,x),$$

where $H_j(y,x) = G_j(x)F_j[(y-w'B)|\sigma_j]$, $w = (1, x)'$, $B = (\alpha , \beta)' = (1, 1)'$, $G_0 \equiv F_0 \equiv N(0,1)$ $(\sigma_0 = 1)$, $G_1 \equiv F_1 \equiv N(0,16)$ $(\sigma_1 = 4)$, and the contamination proportion is $\varepsilon = .25$. Figure 3.1 displays the sample means of the residuals (left) and the correlation coefficients between residuals and fitted values (right) for the simulated samples. The box plots for the estimators GMM and GMS show a clear deviation from the behaviour of the LS estimator. Nonnull, both positive and negative, sample means and sample correlation coefficients occur in a rather symmetric fashion. These phenomenon is not unexpected in the light of comments *ii)* and *iii)* at the end of section 3.2 and might alterate the visual perception of the standard residual plots. Figure 3.2 is a 2x3 matrix array of plots describing the situation for a particular sample. The plot in cell (1,1) is the scatter cloud of the data (y_i,x_i), i = 1, ...,40. Cases 3 and 39 are identified as extreme points with the largest values of x. Cell (1,2) contains the plot of the studentized least squares residuals $r_i = \hat{e}_i/[\hat{\sigma}(1-h_{ii})^{1/2}]$ versus the fitted values $\hat{y}_i$, where $\hat{\sigma}$ is the least squares estimate of the standard deviation σ and h_{ii} is the ith diagonal element of the nxn orthogonal projection matrix **H**. Cases 3 and 39 stand out along the horizontal axis. The plot in cell (1,3) is the plot of r_i versus case number. Cases 3 and 39 lie within a not very noticeable cloud of residuals. The situation changes, however, in the second row, where raw residuals based on the GMS estimator are used to construct plots (2,1), (2,2) and (2,3). In (2,1) and (2,2), cases 3 and 39 are detected in both the horizontal and vertical axis, and in (2,3), cases 3 and 39 have the largest associated residuals $\tilde{e}_{i,GMS}$. This is an illustration on how robust residuals might retain harmful information on the regressor variable, in the spirit of comment *i)* at the end of section 3.2.

Figure 3.1

Figure 3.2

Remark. Estimators $\tilde{\mathbf{B}}_M$, $\tilde{\mathbf{B}}_{GMM}$ and $\tilde{\mathbf{B}}_{GMS}$ are used by McKean, Sheather and Hettmansperger (1993) in a previous work on the properties of residuals obtained from both least squares and robust estimation techniques. A comparison of their results with the ones obtained in this paper is given in section 5 below.

3.4 (Partially) Modified residuals

A possible way to overcome (*i*)-(*iv*) of section 3.2 is to construct plots based on a suitable modified vector of residuals. Given a set of p regressors $\mathbf{x}_1$, ..., $\mathbf{x}_p$ indexed by $I = \{1, ..., p\}$, denote by $I_k = \{i_1, ..., i_k\}$ $(1 \le i_1 < ... < i_k \le p)$ a selected subgroup and consider the matrix $\mathbf{W}_k = (1_n | \mathbf{X}_{I_k})$, where $\mathbf{X}_{I_k}$ is the nxk matrix formed with the regressors in I_k. The class of partially modified residuals $\tilde{\mathbf{e}}_{M,k}$ is defined by

$$\tilde{\mathbf{e}}_{M,k} = (\mathbf{I}_n - \mathbf{H}_k)\tilde{\mathbf{e}}, \tag{3.8}$$

where $\mathbf{H}_k = \mathbf{W}_k(\mathbf{W}_k'\mathbf{W}_k)^{-1}\mathbf{W}_k'$ is the orthogonal nxn projection matrix on the linear manifold spanned by $\mathbf{W}_k$. Notice that, since $\tilde{\mathbf{e}}_{M,p} = (\mathbf{I}_n - \mathbf{H}_p)\tilde{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})(\mathbf{Y} - \mathbf{W}\tilde{\mathbf{B}}_n) = \hat{\mathbf{e}}$, the fully modified residuals are identical to the least squares residuals. The idea behind the modified residuals is, in some way, similar to the idea of projected residuals in nonlinear regression suggested by Cook and Tsai (1985), although in a different context.

For plotting purposes, the use of $\tilde{\mathbf{e}}_{M,k}$ can be implemented as follows. Once that an anomaly is detected in a plot based on $\tilde{\mathbf{e}}$, construct the family of plots

$$(\tilde{\mathbf{e}}, \mathbf{x}_r), \tag{3.9.a}$$

for r = 1, ...,p. The goal is to detect directions in the x-space where the nonorthogonal association between $\tilde{\mathbf{e}}$ and $\mathbf{X}$ is most harmful. Construct then, sequentially, the families

$$\begin{gathered}(\tilde{\mathbf{e}}_{M,\{i_1\}}, \mathbf{x}_r), \\ (\tilde{\mathbf{e}}_{M,\{i_1,i_2\}}, \mathbf{x}_r),\end{gathered} \tag{3.9.b}$$

...., until reaching an index subset $I_k = \{i_1, ...,i_k\}$ where the associated group of plots have no visible pattern and then conduct a residual analysis based on the class $\tilde{\mathbf{e}}_{M,k}$. From (3.5), one obtains

$$\tilde{e}_{M,k} \cong (I_n - H_k)(\hat{e} + W\gamma_n) = \hat{e} + h_k,$$

where $h_k = (I_n - H_k)W_{(k)}\gamma_{n,(k)}$, $W_{(k)}$ is the nx[(p+1)-k] matrix of the regressors not in I_k and $\gamma_{n,(k)}$ is formed with the corresponding coordinates of γ_n. If no pattern is observed, the following (heuristic) results follow if h_k is taken to be approximately constant:

$$V[\tilde{e}_{M,k} | X] \cong V[\hat{e} | X] \cong \sigma^2(I_n - H), \tag{3.10.a}$$

and

$$\tilde{e}'_{M,k}W \cong (\hat{e} + h_k)'W = (0 | h'_k W_{(k)}). \tag{3.10.b}$$

(3.10.a) suggests replacing the modified residuals $\tilde{e}_{M,k,i}$ by standardized versions

$$\tilde{r}_{M,k,i} = \tilde{e}_{M,k,i}/[\tilde{\sigma}(1-h_{ii})^{1/2}], \tag{3.11}$$

where $\tilde{\sigma}$ is a robust estimate of the scale of the $q_i = y_i - w'B$. From the sequential construction of I_k above, it is reasonable to expect a small magnitude for the row vector $h'_k W_{(k)}$. Since the modified residuals have sample mean equal to zero, the plot of the standardized $\tilde{r}_{M,k} = (\tilde{r}_{M,k,i})$ versus the fitted values $\tilde{Y}$ should be a corrected version of the initial of $\tilde{e}$ versus $\tilde{Y}$.

4. EXAMPLES

Two well-known real data examples are used to illustrate both the results in sections 3.1 and 3.2, and the methodology based on the partially modified residuals introduced in section 3.4. The data are assumed to be generated by an appropriate setup of the form (2.1)-(2.2).

4.1 Gesell adaptive score data

The Gesell adaptive score data are n = 21 observations on scalar variables y and x whose meaning and description can be found, for example, in Cook and Weisberg (1982). It is accepted that case 18 is extreme with the largest value of x, and case 19 is an outlying response. For this example, the residual plots based on both M and GMS estimates had a similar behaviour to the plots based on the LS residuals. However, graphical displays based on GMM residuals presented an anomaly as shown in the 3x3 matrix array of figure 4.1. Cells (1,2) and (1,3) are the standard LS residual plots where case 18 has a small studentized residual and case 19 has the largest

residual as it should be. The second row presents the problem. Plots in cells (2,1) and (2,2) of the raw GMM residuals versus, respectively, $\mathbf{x}$ and the fitted values suggest linear trends as remarked by the superimposed least squares lines. Residuals based on the GMM estimator retain information on $\mathbf{x}$ and cases 18 and 19 have now the largest residuals in absolute magnitude. Plots in row 3 correspond to the standardized modified GMM residuals using the MAD as the estimate $\tilde{\sigma}$ for σ. Cells (3,2) and (3,3) identify properly the character of cases 18 and 19.

Figure 4.1

4.2 Salinity data

The data and previous analyses of them are described in Rousseeuw and Leroy (1987 p. 82). There are n = 28 cases and p = 3 regressors. It seems to be common agreement about the fact that cases 3, 5 and 16 are extreme rows of the matrix $\mathbf{X}$. Case 16 has associated an outlying response as well. The analysis given in this paper is supported by the three graphical arrays in figures 4.2.a), 4.2.b) and 4.2.c). In this example, residuals based on M estimators behaved as least squares residuals. The situation was different for the case of GMM and GMS residuals which in turn, shared a similar pattern. GMM residuals are chosen for illustration purposes. The first row in 4.2.a) contains the plots of the least squares residuals and the relative position of cases 3, 5 and 16. As shown in the second row, the relative position of 3, 5 and 16 changes when the GMM residuals are used instead since now cases 5 and 16 have the largest residuals. The third row contains the plots $(\tilde{\mathbf{e}}_{GMM}, \mathbf{x}_r)$ and suggests a problem associated with the regressor $\mathbf{x}_3$. Figure 4.2.b) displays the family $(\tilde{\mathbf{e}}_{GMM,\{i_1\}}, \mathbf{x}_r)$ and supports this impression since the residual for case 5 stands out unless the GMM residuals are projected onto the third regressor. From 4.2.c), $I_2 = \{1,3\}$ seems to produce a set of modified residuals with a reasonable degree of orthogonality versus the fitted values. As seen in the third row, case 16 has the largest residual and case 5 remains unnoticed.

Figure 4.2.a)

Figure 4.2.b)

Figure 4.2.c)

5. DISCUSSION AND COMPARISON WITH PREVIOUS WORK

In this paper, properties of residual plots for both least squares and robust estimators are compared under a common set of assumptions. Theoretical results suggest that for high breakdown but biased robust estimators, the residuals retain information on the regressors, and this phenomenon might produce misleading residual plots. A remedial action is suggested in section 3.4 based on the class of partially modified residuals $\tilde{\mathbf{e}}_{M,k}$ associated to a properly built subset $I_k = \{i_1, \ldots, i_k\}$ of regressors. These are thought to be a flexible compromise between the raw robust residuals $\tilde{\mathbf{e}}$ and their least squares counterparts $\hat{\mathbf{e}}$.

McKean et al. (1993) study the behaviour of residual plots based on M estimators and GM estimators. They consider first-order properties of residuals derived from fitting a model $\mathbf{Y} = \mathbf{WB} + \mathbf{u}$, when the true model is $\mathbf{Y} = \mathbf{WB} + \mathbf{C}\delta_n + \mathbf{u}$, where $\mathbf{C}$ is a full rank nxq matrix whose columns are independent of the columns of $\mathbf{W}$, and $\delta_n = n^{-1/2}\theta$ for a qx1 vector θ. Their results conclude that plots based on M estimators behave quite similarly to the LS residual plots but the plots based on both GMM and GMS estimators might be misleading as they have negative correlation patterns even when the true model is fitted. McKean et al. (1993) propose specific standarization methods for each class of residuals $\tilde{\mathbf{e}}_M$, $\tilde{\mathbf{e}}_{GMM}$, and $\tilde{\mathbf{e}}_{GMS}$.

Similar conclusions are obtained in this paper regarding the behaviour of the residuals based on M estimators under the setup (2.1)-(2.2). As shown in figure 3.1., the correlation in the plot of residuals versus fitted values will be hardly perceptible. For GMM and GMS estimators, however, figure 3.1. and examples 4.1 and 4.2 show that positive correlations can also occur under (2.1)-(2.2). An advantage of the modified residuals $\tilde{\mathbf{e}}_{M,k,i}$ is that they can be standardized in a natural way that not depends on concrete specifications of both the functional defining the robust estimator and the associated empirical influence function. Recall, finally, that in McKean et al. (1993), the emphasis is more in model misspecification than in general properties as developed here.

A remaining problem in this paper is the study of the behaviour of the LMS residuals. This has been done in McKean et al. (in press b) in the same context of model misspecification as in McKean et al. (1993). The conclusions are similar to the ones relative to the high breakdown estimators GMM and GMS. Plots are misleading since strong negative correlations appear to distort the visual perception of the plot. McKean et

al. (in press b) study also the case of residuals based on the high breakdown LTS estimator of Rousseeuw (1984). Nonetheless, as shown in Atkinson (1986), Rousseeuw and van Zomeren (1990), and Fung (1993) the LMS residuals are a valuable tool in relation to the masking problem associated to the standard least squares diagnostic techniques. In fact, Fung (1993) contains an interesting study of the salinity data based only on the first and third regressor variables where the role of cases 3, 5 and 16 is clarified in detail. Fung (1993) concludes that case 16 is the only outlying response in the data set while case 5 is a slightly high leverage point. Notice that the exploratory sequential graphical method proposed in section 3.4 reaches exactly the same conclusion, and identifies correctly the character of case 5, once the harmful effect of the bias of the estimator on the third variable is eliminated via projection.

ACKNOWLEDGEMENTS

Part of this paper was written while the author was a visiting scholar at New York University. The author is grateful to professor S. Chatterjee for his hospitality and stimulating conversations.

APPENDIX

Proof of proposition 2.1. a) Consider, for $y = (y_1, \ldots, y_p)' \in \mathbb{R}^n$ and $\mathbf{X} = (x_1, \ldots, x_n)'$ of nxp, the function

$$C(y,\mathbf{X}) = \prod_{i=1}^{n} \{[1-\varepsilon(x_i)]F_0(q_i/\sigma_0) + \varepsilon(x_i)F_1(q_i/\sigma_1)\}, \qquad (A.1)$$

Integrating (A.1) with respect the joint of $\mathbf{X} = (x_1, \ldots, x_n)'$ yields, by independence and construction of (2.1) and (2.2),

$$\int_{\mathbf{X}\le\mathcal{X}} C(y,\mathcal{X}) \prod_{i=1}^{n} [(1-\varepsilon)g_0(x_i) + \varepsilon g_1(x_i)]m(dx_i)$$

$$= \prod_{i=1}^{n} [(1-\varepsilon)\int_{x_i\le\alpha_i} F_0(q_i/\sigma_0)\, g_0(x_i)m(dx_i) + \varepsilon\int_{x_i\le\alpha_i} F_1(q_i/\sigma_1)\, g_1(x_i)m(dx_i)]$$

$$= \prod_{i=1}^{n} [(1-\varepsilon)H_0(y_i,\alpha_i) + \varepsilon H_1(y_i,\alpha_i)] = \prod_{i=1}^{n} H(y_i,\alpha_i)$$

$= P[\mathbf{Y} \le y;\ \mathbf{X} \le \mathcal{X}]$, where $\mathcal{X} = (\alpha_1, \ldots, \alpha_n)'$, and the notations $a \le b$, between pxl vectors $a = (a_1, \ldots, a_p)'$, $b = (b_1, \ldots, b_p)'$, and $A \le B$,

between nxp matrices $A = (a_{ij})$, $B = (b_{ij})$, mean, respectively, $a_j \leq b_j$ for j = 1, ...,p and $a_{ij} \leq b_{ij}$ for i = 1, ...,n, j = 1, ...,p. Equations above imply the first part of a). Obviously $\varepsilon(x_i)$ are i.i.d. with expectation $E[\varepsilon(x_i)] = \varepsilon\int_{\mathbb{R}^p} g_1(x_i)m(dx_i) = \varepsilon$. Since $0 \leq \varepsilon(x_i) \leq 1$, $E[\varepsilon(x_i)^2] \leq E[\varepsilon(x_i)] = \varepsilon$, whence $var[\varepsilon(x_i)] \leq \varepsilon - \varepsilon^2 = \varepsilon(1-\varepsilon)$. **b and c)** For all i = 1, ...,n, by part a) and construction of $F_0(u|\sigma_0)$ and $F_1(u|\sigma_1)$, it is easily seen that $E[q_i|\mathbf{X}] = E[q_i|x_i] = 0$, which is equivalent to statement b). By the same arguments, $cov[(y_i,y_j)|\mathbf{X}] = 0$, for $i \neq j$, and $d_i(\mathbf{X}) = var[y_i|\mathbf{X}] = var[y_i|x_i] = var[q_i|x_i] = [1-\varepsilon(x_i)]\sigma_o^2 + \varepsilon(x_i)\sigma_1^2$. Therefore, $V[\mathbf{Y}|\mathbf{X}] = D(\mathbf{X}) = diag(d_1(\mathbf{X}), ...,d_n(\mathbf{X}))$, where $E[d_i(\mathbf{X})] = (1-\varepsilon)\sigma_o^2 + \varepsilon\sigma_1^2 = \sigma^2$. ■

Proof of proposition 2.2. a) $E[q_i] = E[E[q_i|x_i]] = 0$ and $var[q_i] = E[var[q_i|x_i]] + var[E[q_i|x_i]] = E\{[1-\varepsilon(x_i)]\sigma_o^2 + \varepsilon(x_i)\sigma_1^2\} = \sigma^2$. **b)** From part a) of proposition 2.1, the joint of $(q_1, ...,q_n)'$ given $\mathbf{X}$ is

$$\prod_{i=1}^{n}\{[1-\varepsilon(x_i)]F_0(q_i/\sigma_0) + \varepsilon(x_i)F_1(q_i/\sigma_1)\},$$

from where b) follows. ■

REFERENCES

Atkinson, A.C. (1985), *Plots Transformations and Regression*, Oxford, U.K.: Oxford UNiversity Press.

——————— (1986), "Masking Unmasked," *Biometrika*, 73, 533-541.

Coakley, C.W., and Hettmansperger, T.P. (1993), "A Bounded Influence, High Breakdown Efficient Regression Estimator," *Journal of the American Statistical Association*, 88, 872-880.

Cook, R.D., and Weisberg, S. (1982), *Outliers and Influence in Regression*, London: Chapman and Hall.

Cook, R.D., and Tsai, C-L. (1985), "Residuals in Nonlinear Regression," *Biometrika*, 72, 23-29.

Cook, R.D., Hawkins, D. M., and Weisberg, S. (1992), "Comparison of Model Misspecification Diagnostics Using Residuals form Least Mean of Squares and Least Median of Squares Fits," *Journal of the American Statistical Association*, 87, 419-424.

——————— (1993), "Exact Iterative Computation of the Robust Minimum Volume Ellipsoid Estimator,"

Statistics and Probability Letters, 16, 213-218.

Fung, K. Y. (1991), "Unmasking Outliers and Leverage Points: A Confirmation," *Journal of the American Statistical Association*, 88, 515-519.

Hampel, F., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics*, New York: John Wiley.

Hinkley, D.A. (1977), "Jacknifing in unbalanced situations," *Technometrics*, 19, 285-292.

Huber, P..J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics*, 1, 799-821.

Martin, R.D., Yohai, V., and Zamar, R.H. (1989), "Min-Max Robust Regression," *The Annals of Statistics*, 17, 1608-1630.

Martin, R.D., and Yohai, V. (1986), "Influence Functionals for Time Series" (with discussion), *The Annals of Statistics*, 781-855.

McKean, J.W., Sheather, S.J., and Hettmansperger, T.P. (1990), "Regression Diagnostics for Rank-Based Methods," *Journal of the American Statistical Association*, 85, 1018-1028.

_____ (1993), "The Use and Interpretation of Residuals based on Robust Estimation," *Journal of the American Statistical Association*, 88, 1254-1263.

_____ (in press a), "Robust and High Breakdown Fits of Polynomial Models," *Technometrics*.

_____ (in press b), "The interpretability of LMS and LTS Residual Plots."

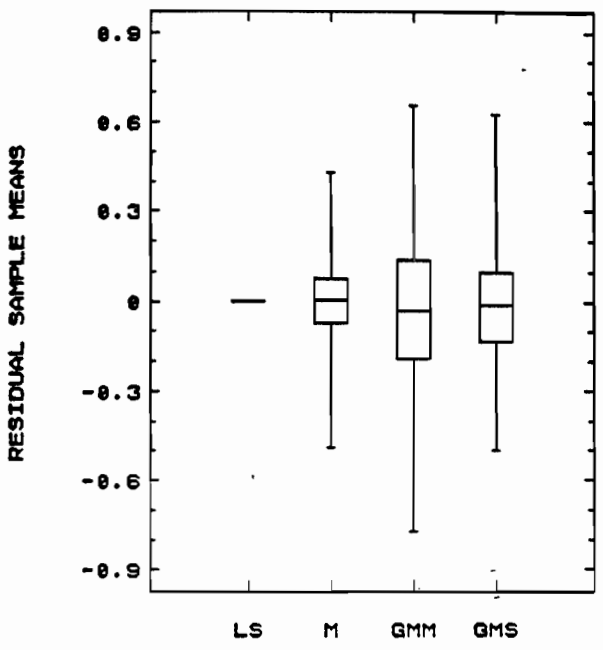
Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P.J., and Leroy, A. (1987), *Robust Regression and Outier Detection*, New York: John Wiley.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points" (with discussion), *Journal of the American Statistical Association*, 85, 633-651.

Simpson, D.G., Ruppert, D., and Carroll, R. (1992), "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Association*, 87, 439-450.

Weisberg, S. (1984), *Applied Linear Regression*, 2nd Edition, New York: John Wiley.


RESIDUAL SAMPLE MEANS
0.9
0.6
0.3
0
-0.3
-0.6
-0.9
LS
M
GMM
GMS

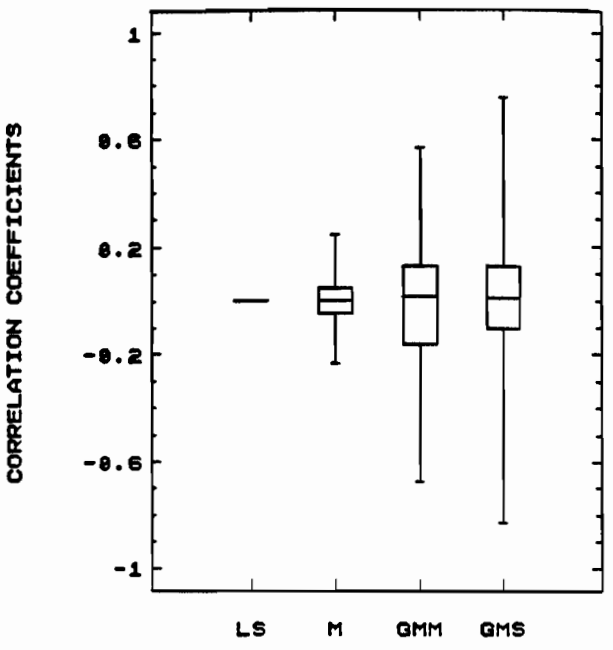

CORRELATION COEFFICIENTS
1
0.6
0.2
-0.2
-0.6
-1
LS
M
GMM
GMS

Figure 3.1

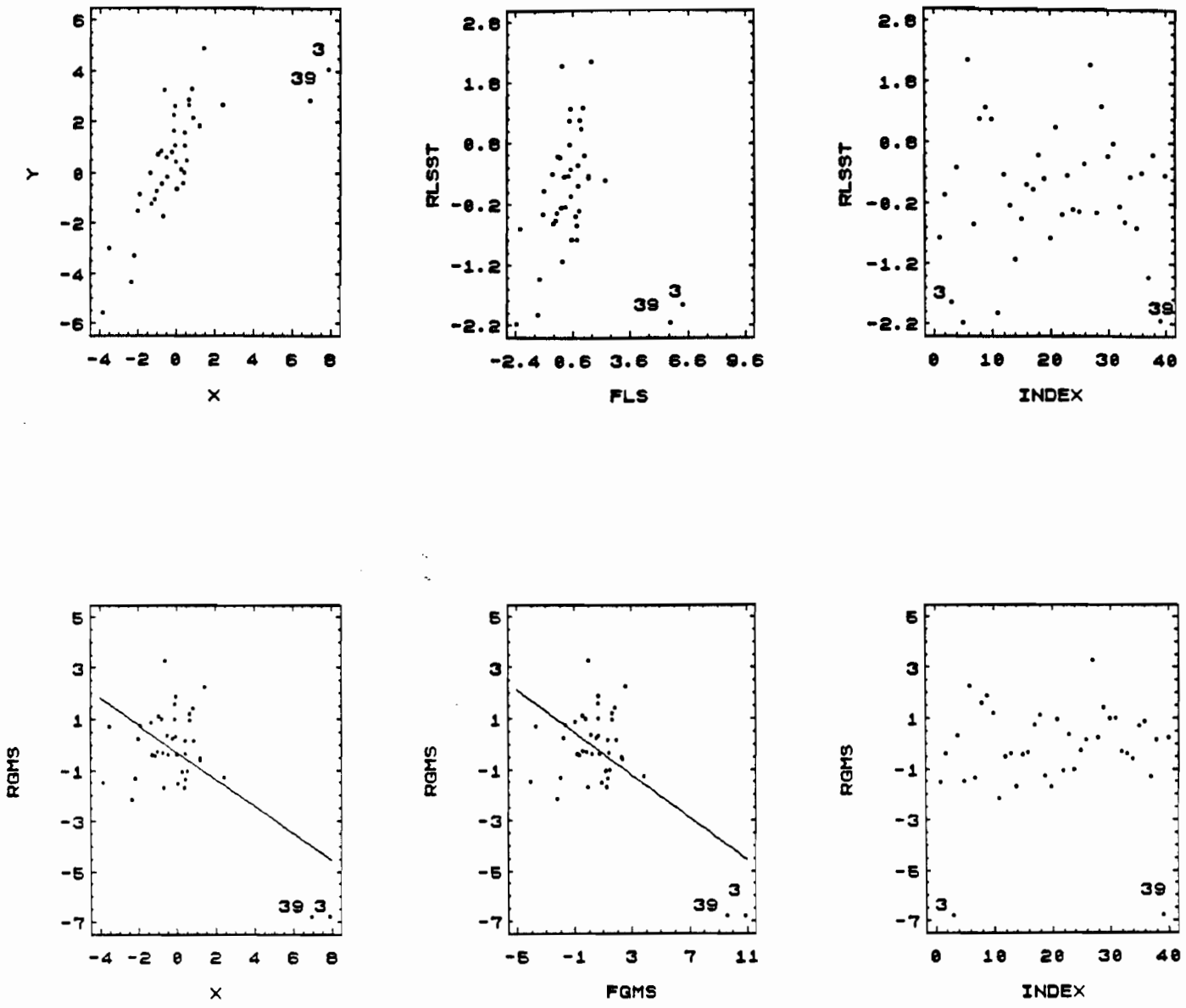
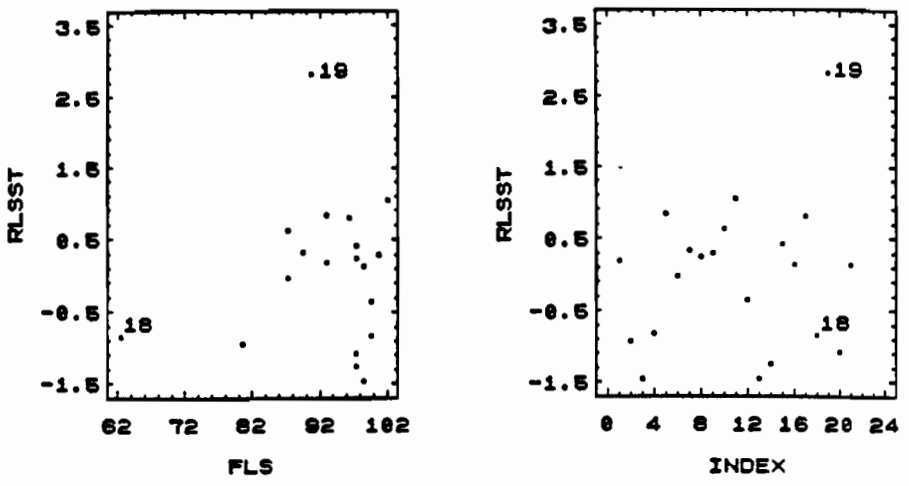
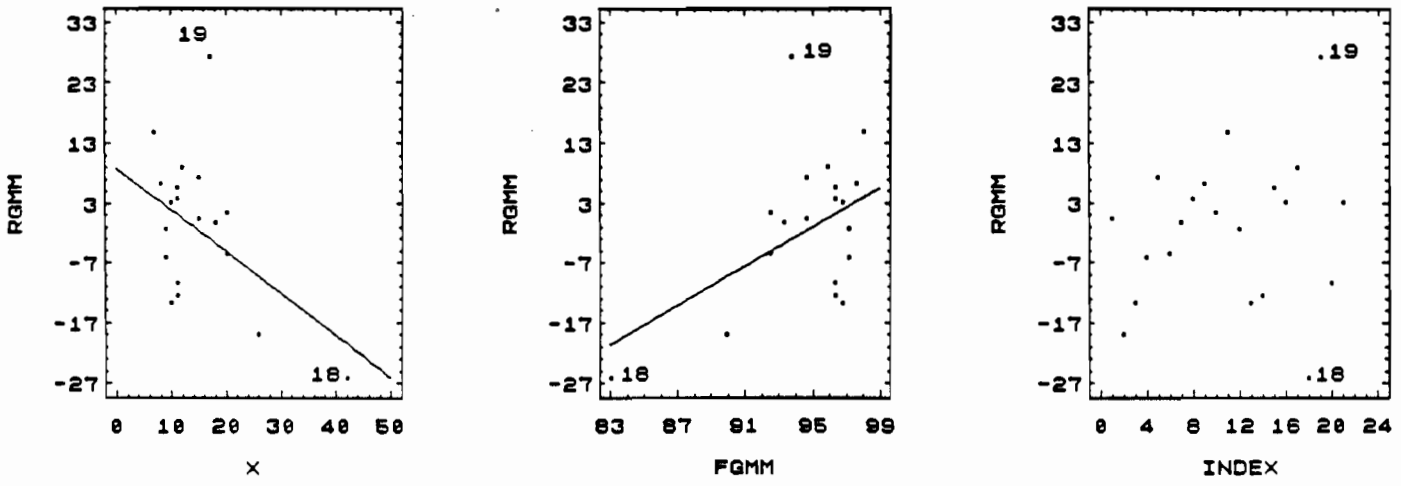


Figure 3.2


RLSST
FLS
INDEX
.19
.18


RGMM
X
FGMM
INDEX
19
18.
.19
.18

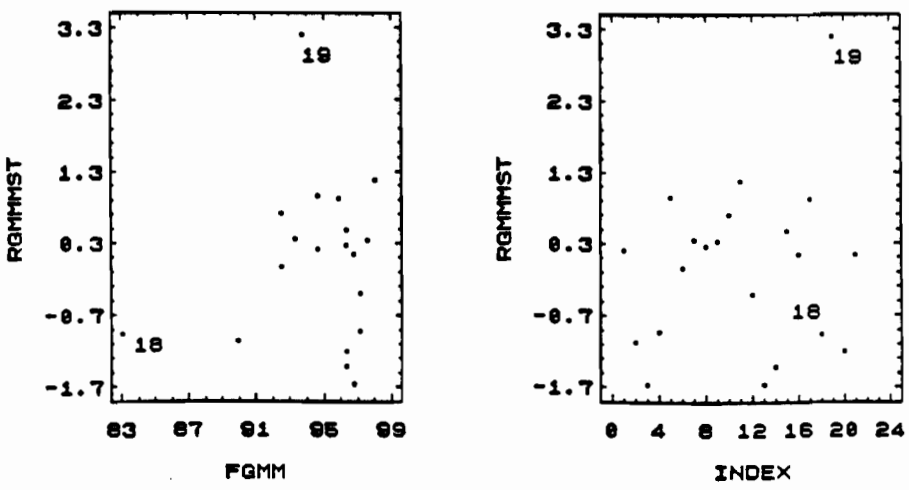
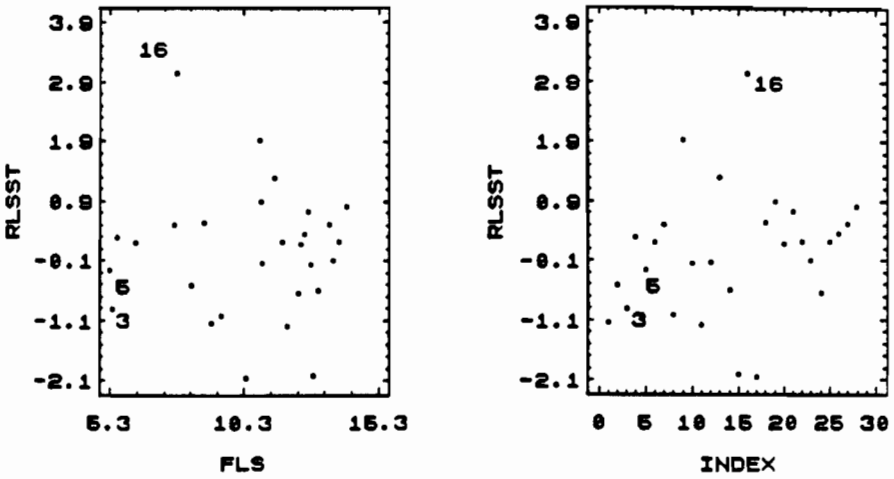
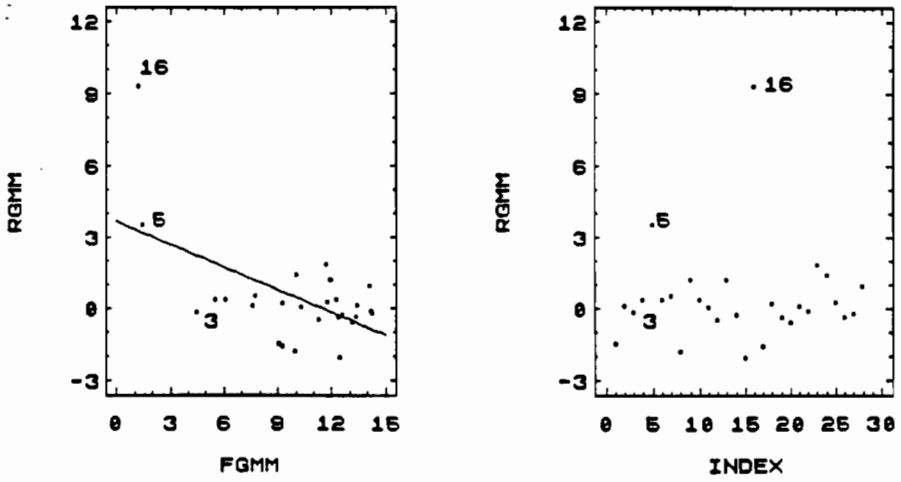

RGMMMST
FGMM
INDEX
19
18

Figure 4.1





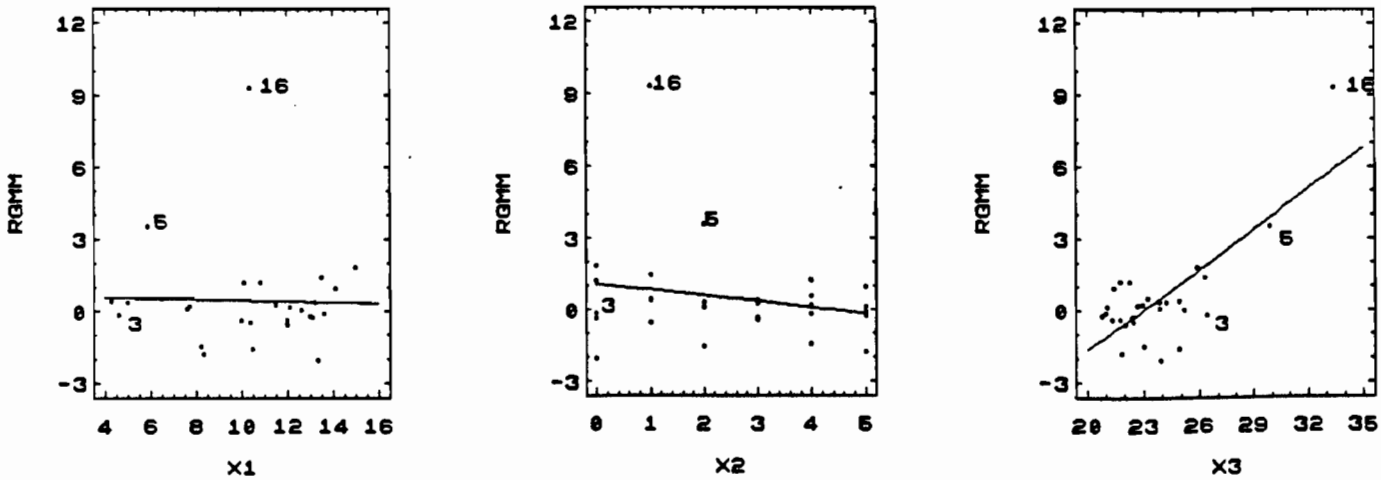


Figure 4.2.a)

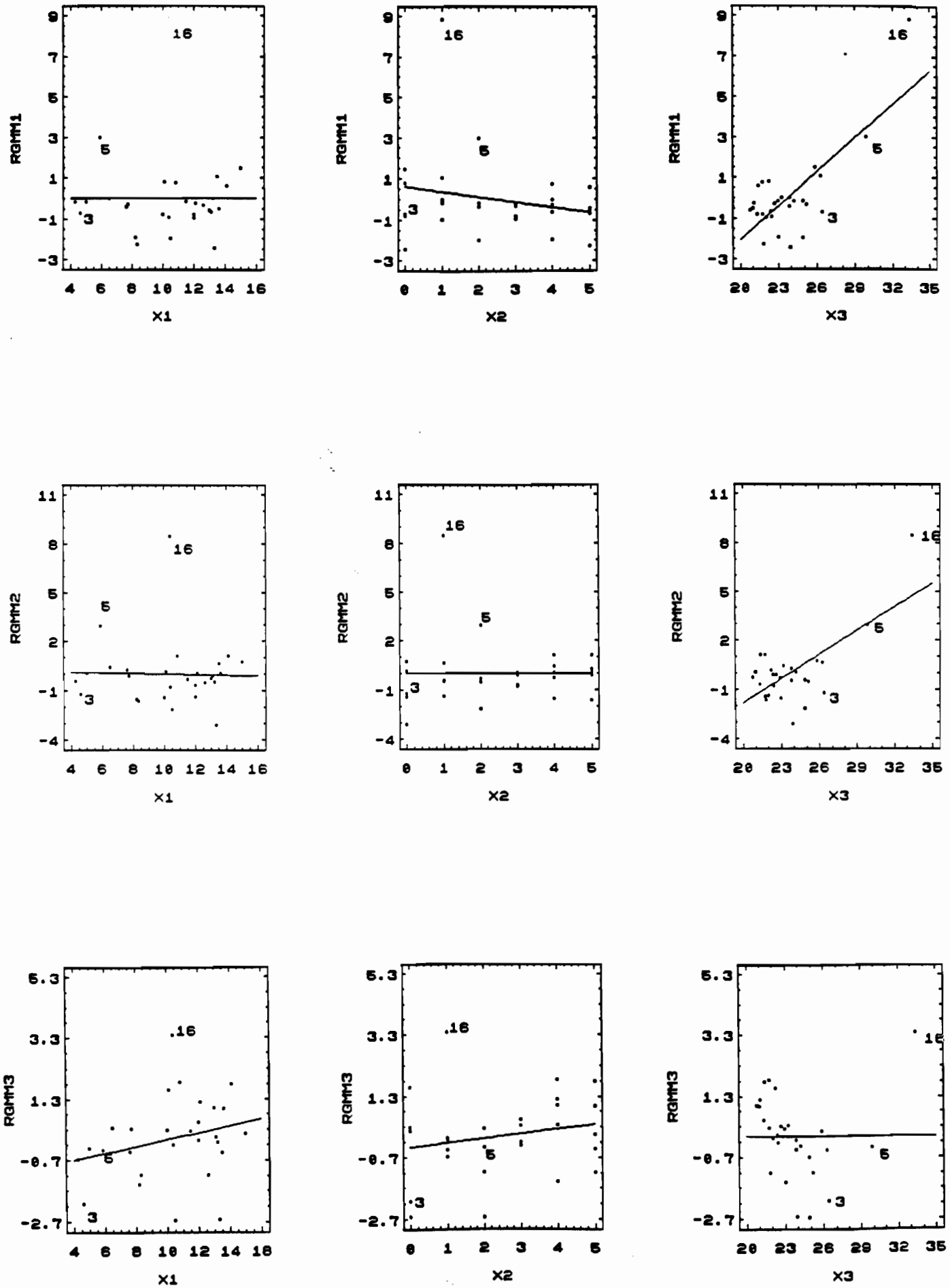


Figure 4.2.b)

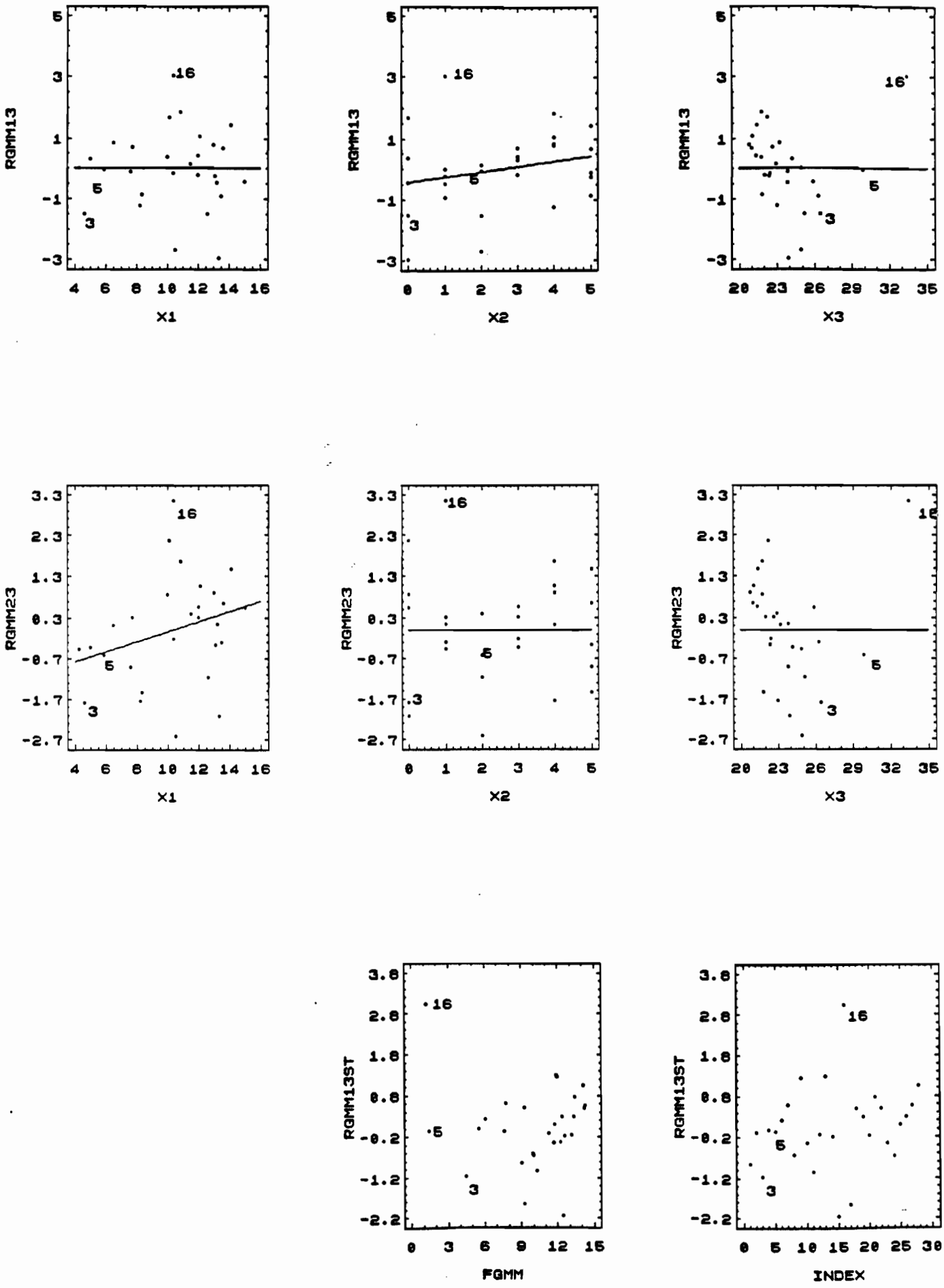


Figure 4.2.c)

CAPTIONS FOR FIGURES

Figure 3.1 Sample means of the residuals (left) and correlation coefficients between residuals and fitted values (right).

Figure 3.2 Residual plots based on the LS and GMS estimators for a particular sample of the simulation study of section 3.3. The notation for the legends on the axes are self explained.

Figure 4.1 Gesell adaptive score data: 3x3 matrix array of residual plots based on the LS and GMM estimators.

Figure 4.2.a) Salinity data: 3x3 matrix array of preliminary residual plots based on the LS and GMM estimators.

Figure 4.2.b) Salinity data: 3x3 matrix array of plots of the one-dimensional partially modified GMM residuals versus the regressors.

Figure 4.2.c) Salinity data: 3x3 matrix array of the two dimensional partially modified GMM residuals versus the regressors. The third row is the final output of the analysis.