Разкриване на несъгласувани наблюдения в регресионни модели

Записки

доц. дн Нейко М. Нейков
НИМХ-БАН

София, 13. 10. 2017 г.

Съдържание

Глава 1

Въведение

Настоящите записки са посветени на разкриването на несъгласувани наблюдения в многомерни масиви от данни, включително на множествената линейна регресия чрез процедури от програмна среда R. Ще бъдат използвани процедурите LTS и MCD от robustbase библиотеката за робастно оценяване. Причината за това е, че тези робастни оценки за параметрите на линейни регресионни модели и ковариационни матрици, притежават възможно най-високата прагова стойност (точка) 1/2, а от друга страна са комбинаторни по наблюденията варианти на Метода на Най-малките Квадрати (МНК) и Метода на Максималното правдоподобие (ММП). Преди да бъде дадена точна дефиниция на прагова точка ще разгледаме една нейна интерпретация на достъпен език. Под прагова точка на една статистика се разбира максималния процент несъгласувани наблюдения, който не промени значимо стойността на статистиката, избрана за оценка на параметъра било то средна, дисперсия, ковариационна матрица, регресионни параметри и други. Пример за статистика с висока прагова стойност от 1/2 (50%) е медианата в едномерния случай, тъй като 50% от наблюденията могат да бъдат заместени с произволни стойности без ни най-малко това да влияе на стойността на тази статистика.

Дефиниция на прагова точка. Основен критерий, чрез който статистическите оценки могат да бъдат класифицирани е по стойността на праговата точка. От многото дефиниции за прагова точка най-подходяща за целите на приложната статистика е дадената дефиниция от Rousseeuw and Leroy (1987), тъй като е ориентирана към крайната извадка. Нека с $\widetilde{\mathrm{X}}_m$ означим извадката, получена чрез заместването на кои да е m наблюдения от оригиналната извадка X с произволни стойности, а T е дадена статистика.

Дефиниция 1.1 *Праговата точка на статистиката T за крайната извадка X се дефинира като*

$$\varepsilon_n^*(T) = \frac{1}{n}\min\{m : \sup_{\tilde{X}_m} \|T(\tilde{X}_m) - T(X)\| < \infty\},$$

където $\|.\|$ е Евклидовата норма.

Супремумът в дефиницията е взет по всевъзможните замърсени извадки от m наблюдения. От дефиницията следва, че праговата точка дава представа за максималния процент наблюдения, чиито стойности могат да бъдат заместени с произволни стойности без това да влоши качествата на оценката. Класическата оценка на средната, едномерна или многомерна, стойност притежава асимптотична прагова точка 0 при неограниченото нарастване на обема на извадката n, тъй като наличието само на едно несъгласувано наблюдение, което се отклонява значително от мажоритарната част на наблюденията, води до значителна изместеност на $\sup_{\tilde{X}_m} \|T(\tilde{X}_m) - T(X)\|$. За разлика от оценката на извадковата средна, оценката на извадковата медиана не се променя при промяната на стойностите на почти половината от наблюденията.

Ще използваме следните съкращения:

1. Least Trimmed Squares estimator (LTS) - оценка по метод на най-малката сума от орязани квадрати;
2. Minimum Covariance Determinant estimator (MCD) - оценка на многомерната средна и ковариационна матрица по k наблюдения от извадка с обем n, чиято детерминанта е минимална.

Във втора глава са разгледани методите и алгоритмите за оценяване в линейни регресионни модели като МНК, LTS, LTS с тегла и робастна процедура за разкриване на регресионни несъгласувани наблюдения в зависимата променлива.

В трета глава са разгледани класически и робастни от тип MCD оценки на многомерната средна и ковариационна матрица, процедури за разкриване на многомерни несъгласувани наблюдения с класическо и робастно от тип MCD разстояние на Махаланобис и процедура за разкриване на несъгласувани наблюдения в линейни регресионни модели чрез съвместното използване на LTS стандартизираните регресионни остатъци и робастните MCD разстояния на Махаланобис.

Литературната справка съдържа използваните заглавия в текста.

Глава 2

Робастни LTS оценки на параметрите в линейната регресия

2.1 Линеен регресионен модел

2.1.1 Дефиниции

Линейния регресионен модел се дефинира като

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i = x_i^T \beta + \epsilon_i, \qquad i = 1, \ldots, n \tag{2.1}$$

където y_i и $x_i^T = (x_{i0}, x_{i1}, \ldots, x_{ip-1})$ са независими наблюдения над зависимата променлива и вектора от наблюдения на предикторните променливи, съответно, $\beta^T = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ е вектор от неизвестни параметри, ϵ_i е случайна грешка на i-то наблюдение, за която се предполага, че има нулево очакване $E(\epsilon_i) = 0$, константна неизвестна дисперсия $\text{var}(\epsilon_i) = \sigma^2$, ϵ_l и ϵ_l са независими за $l \neq m$, т.е. $\text{cov}(\epsilon_l, \epsilon_m) = 0$.

В матрична форма модела приема вида: Във векторна форма модела (2.1) приема вида

$$Y_{n\times 1} = X_{n\times p}\beta + \epsilon_{n\times 1},$$

където

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} y_1 \\ \ldots \\ y_i \\ \ldots \\ y_n \end{pmatrix}, \ \mathbf{X}_{n\times p} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \\ 1 & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \ \epsilon_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \ldots \\ \epsilon_i \\ \ldots \\ \epsilon_n \end{pmatrix}$$

Оценката по МНК $\hat{\theta}$ на β се дефинира като

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

$$\begin{aligned} RSS(\beta) &= (Y - X\beta)^T(Y - X\beta) \\ &= Y^TY - (X\beta)^TY - Y^TX\beta + (X\beta)^TX\beta \\ &= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta \end{aligned}$$

От необходимото условие за екстремум на функцията $RSS(\beta)$

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2Y^TX + 2X^TX\beta = 0$$

следва, че оценката по МНК удовлетворява системата нормални уравнения

$$X^TX\beta = Y^TX$$

Ако матрицата X е с пълен ранг тогава $det(X^TX) \neq 0$, от където следва че

$$\hat{\beta} = \left(X^TX\right)^{-1}X^TY.$$

Матрицата от вторите частни производни на $RSS(\beta)$ е равна на

$$\frac{\partial^2 RSS(\beta)}{\partial \beta \partial \beta^T} = 2X^TX.$$

Следователно $RSS(\beta)$ достига глобален минимум в $\hat{\beta}$.

Предсказаните стойности на Y чрез регресионния модел се дефинират като

$$\hat{Y} = X\hat{\beta} = X\left(X^TX\right)^{-1}X^TY = HY.$$

Матрицата $H = X\left(X^TX\right)^{-1}X^T$ се нарича *hat* матрица, понеже поставя шапка на вектора Y или матрицата на проектиране. Чрез разликата между стойностите на наблюденията y_i и предсказаните стойности $\hat{y}_i$ се дефинират регресионните остатъци (residual), т.е.,

$$\hat{\epsilon}_i = y_i - \hat{y}_i, \quad \hat{\epsilon} = Y - \hat{Y} = (I_n - H)Y$$

Ще отбележим, че както предсказаните стойности така и остатъците са линейни комбинации на наблюденията y_i. Някои свойства на проекционната матрицаH: H е симетрична $H^T = H$; H е идемпотентна $H^2 = H$; $(I_n - H)^T(I_n - H) = (I_n - H)$

2.1.2 Свойства на оценката по МНК

Оценката по МНК $\hat{\beta}$ е неизместена оценка за β. Д-во:

$$\begin{aligned} E(\hat{\beta}) &= E\left[\left(X^TX\right)^{-1}X^TY\right] \\ &= \left(X^TX\right)^{-1}X^TE(Y) \\ &= \left(X^TX\right)^{-1}X^TX\beta = \beta \end{aligned}$$

Ще покажем се дисперсията на $\hat{\beta}$ е $var(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$. Д-во: От дефиницията на дисперсия на векторна случайна величина следва, че

$$\begin{aligned} var(\hat{\beta}) &= var\left[\left(X^T X\right)^{-1} X^T Y\right] \\ &= \left(X^T X\right)^{-1} X^T var(Y) X \left(X^T X\right)^{-1} \\ &= \left(X^T X\right)^{-1} X^T \sigma^2 I_n X \left(X^T X\right)^{-1} \\ &= \sigma^2 \left(X^T X\right)^{-1} X^T I_n X \left(X^T X\right)^{-1} \\ &= \sigma^2 \left(X^T X\right)^{-1} \end{aligned}$$

2.1.3 Оценяване на дисперсията на грешката σ^2

Оценката за дисперсията на грешката σ^2 в регресионния модел се дефинира както следва Нека

$$RSS(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = (Y - \hat{Y})^T (Y - \hat{Y})$$

е сумата от квадратите на регресионните остатъчи. Тогава,

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n - p}$$

е неизместена оценка за σ^2 . Д-во: Да разгледаме вектора от остатъците

$$\hat{\epsilon} = (Y - \hat{Y}) = Y - HY = (I_n - H)Y.$$

От

$$(I_n - H)X = X - X \left(X^T X\right)^{-1} X^T X = 0,$$

следва че

$$\hat{\epsilon} = (I_n - H)Y = (I_n - H)(X\beta + \epsilon) = (I_n - H)\epsilon$$

Следователно,

$$\begin{aligned} RSS(\hat{\beta}) &= (Y - \hat{Y})^T (Y - \hat{Y}) && (2.2) \\ &= \hat{\epsilon}^T (I_n - H)^T (I_n - H)\hat{\epsilon} && (2.3) \\ &= \hat{\epsilon}^T (I_n - H)\hat{\epsilon} && (2.4) \end{aligned}$$

От предположението за свойствата на грешката в модела следва, че $E(\epsilon_i \epsilon_j) = \sigma^2$ ако $i = j$ и 0 в продивен случай. Следователно

$$\begin{aligned} E(RSS(\hat{\beta})) &= E\left[\sum_{i,j} \epsilon_i (I_n - H)_{i,j} \epsilon_j\right] && (2.5) \\ &= \sigma^2 \sum_{i=1}^{n} (I_n - H)_{i,i} && (2.6) \\ &= \sigma^2 (n - tr(H)) = \sigma^2 (n - p), && (2.7) \end{aligned}$$

понеже

$$\begin{aligned} tr(H) &= tr\left[X\left(X^TX\right)^{-1}X^T\right] && (2.8)\\ &= tr\left[\left(X^TX\right)^{-1}X^TX\right] && (2.9)\\ &= tr(I_p) = p && (2.10)\end{aligned}$$

Hence,

$$E\left(\frac{RSS(\hat{\theta})}{n-p}\right) = \sigma^2$$

We remind that $tr(A+B) = tr(A)+tr(B)$, $tr(AB) = tr(BA)$ and $tr(A^TBA) = tr(BAA^T)$.

2.1.4 стандартна грешка на параметрите на модела

Нека $V = \left(X^TX\right)^{-1}$ с елементи V_{ij} Понеже $var(\hat{\beta}) = \sigma^2V$ то стандартната грешка на всеки елемент на този вектор е $\hat{\beta}_i$ на $\hat{\beta}$ е

$$se(\hat{\beta}_i) = \sigma\sqrt{V_{ii}}$$

Понеже σ е неизвестна се използва нейната оценка

$$s = \hat{\sigma} = \sqrt{\frac{RSS(\hat{\beta})}{n-p}}$$

откъдето получаваме

$$se(\hat{\beta}_i) \approx s\sqrt{V_{ii}}$$

2.2 Оценка по метод на най-малките квадрати

Дефиниция 2.1 *Оценката $\hat{\beta}$ на параметъра β на регресионния модел по метода на най-малките квадрати (МНК) се дефинира като*

$$\min_{\beta}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2. \tag{2.11}$$

Знаем, че когато $\text{rank}(X) = p$ оценката по МНК се получава в явен вид

$$\hat{\beta} = (X^TX)^{-1}X^TY. \tag{2.12}$$

Оценката на σ^2 се дефинира като

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2}{n-p}, \tag{2.13}$$

а оценката на ковариационната матрица на $\hat{\beta}$ се дефинира като

$$\mathrm{cov}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}. \tag{2.14}$$

Чрез критерия на Стюдънт

$$t_j = \frac{\hat{\beta}_j}{\sqrt{(\mathrm{cov}(\hat{\beta}))_{jj}}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathrm{X^T X})^{-1}_{\mathrm{jj}}}} \tag{2.15}$$

се проверява хипотезата за значимост $H_o : \hat{\beta}_j = 0$ срещу алтернативната $H_1 : \hat{\beta}_j \neq 0$ за $j = 1, \ldots, p-1$. Отхвърлянето на H_o интерпретираме като значимо влияние на j-та предикторна променлива x_j за $j = 1, \ldots, p-1$ върху зависимата променлива Y.

От изразите за оценките (2.12)-(2.14) следва, че ако в конкретните данни има несъгласувани наблюдения, то тяхното влияние ще се отрази върху крайните резултати. Да допуснем, че някои от несъгласуваните наблюдения са в стойностите на зависимата променлива y_i, като конкретните стойности са много големи, т.е., това са изключително груби грешки в данните y_i. Влиянието на тези наблюдения ще бъде силно изразено в оценката на $\hat{\sigma}^2$ от (2.13), поради вдигането на квадрат на съответните остатъци. Това ще доведе до некоректни стойности на ковариационната матрица (2.14), което автоматично би довело до неправдоподобни резултати за доверителните интервали за β и некоректни стойности на критерия на Стюдънт за значимост на регресионните параметри. Изразите (2.13)-(2.14) са в основата на критерия на Фишер за общата проверка на линейни хипотези за параметрите β, а именно $H_o : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ срещу алтернативната H_1 : поне един от $\beta_j \neq 0$.

Ефекта от влиянието на несъгласувани наблюдения в предикторните променливи $(x_{i1}, \ldots, x_{ip})$, които са твърде отдалечени от мажоритарната част на матрицата X в определени ситуации могат да имат катастрофални последствия за крайните резултати. Този тип наблюдения се наричат (leverage points) наблюдения на разбалансиране, наклоняване. Характерното за този тип несъгласувани наблюдения е, че те наклоняват хиперравнината $X\hat{\beta}$ към тях. Този тип наблюдения се наблюдават често при събирането на твърде нееднородни данни. Разкриването на този тип несъгласувани данни с помощта на стандартни статистически процедури е изключително трудна задача, поради това, че се появяват в околности на мажоритарната част на X данните под формата на клъстери.

2.2.1 Пример - линеен регресионен модел с и без несъгласувани наблюдения.

В този параграф ще бъде разгледан пример на прост линеен регресионен модел, за да демонстрираме ефекта от влиянието на несъгласуваните наблюдения в данните върху оценките на параметрите и останалите статистики на регресионния анализ. За целта са генерирани 111 наблюдения по модел

$$y_i = \mu_i + \varepsilon_i = 1 + 3x_i + \varepsilon_i \qquad \text{за} \qquad i = 1, \ldots, 111, \tag{2.16}$$

Таблица 2.1: Оценка на параметрите на регресионен модел $y = 1 + 3x + \epsilon$:

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	Signif.
(Intercept)	0.96712	0.13155	7.352	5.96e-11	***
x	3.06138	0.08566	35.737	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9168 on 98 degrees of freedom
Multiple R-squared: 0.9287, Adjusted R-squared: 0.928
F-statistic: 1277 on 1 and 109 DF, p-value: < 2.2e-16

в който x и ε са стандартно нормално разпределени $N(0, 1)$. Данните са представени на панела вляво на фиг. 2.1 с малки кръгчета и в Таблица 3.2 на приложението. Резултата от процедурата lm за линеен регресионен анализ по МНК от програмната среда R са дадени в Таблица 2.1. Оценките на параметрите са 0.96712 и 3.06138, съответно, докато оценката на коефициента на корелация е 0.9287. Тези оценки са статистически значими според критерия на Стюдент, тъй като съответните им $p - value$, дадени в последната колона на таблицата са много по-малки от стандартните нива на съгласие $0.05, 0.01, 0.001$. Според F-теста на Фишер хипотезата H_o се отхвърля понеже съответната $p - value < 2.2e - 16$ е много малка в сравнение със стандартните нива на съгласие. Съответната регресионна линия е дадена на панела вляво на фиг. 2.1.

На фиг. 2.2 са дадени четири стандартни диагностични плота за проверка на предположенията за валидност на модела. На плота горе вляво по ординатната ос са дадени регресионните остатъци спрямо предсказаните стойности по модела, дадени на абсцисната ос. Виждаме, че остатъците са случайно разпределени около нулата и се намират в интервала $(-3, 3)$, което потвърждава адекватността на модела на очакването $E(y) = \mu$. Q-Q плота на остатъците, горе вдясно, потвърждава визуално проверката за стандартно нормално разпределение на остатъците. На плота долу вляво са дадени коренуваните стандартизирани остатъци спрямо предсказаните стойности по модела. Не се забелязват големи остатъци, което е индикатор за наличие на несъгласувани наблюдения. На панела долу вдясно са дадени разстоянията на Cook $||\hat{\beta} - \hat{\beta}_{(-i)}||$ за $i = 1, \ldots, n$, като разлика на оценката на параметъра $\hat{\beta}$ и $\hat{\beta}_{(-i)}$, пресметнати съответно по всичките наблюдения и чрез изключване на i-то наблюдение от данните, където $||.||$ е Евклидовата норма.

На панела вляво на фиг. 2.1 с малки кръгчета са дадени замърсените данни, получени чрез заместване на някои от наблюденията, генерирани по модел (2.16) с несъгласувани наблюдения както следва: (a) $5, 6, 21, 54, 100 - 107$ са разбалансирани несъгласувани наблюдения; (b) $108 - 111$ са от тип балансирани наблюдения; (c) $11, 31, 3364$ са несъгласувани (вертикални) наблюдения в зависимата променлива.

Резултата от процедурата lm е даден в Таблица 2.2. Влиянието на несъгласуваните наблюдения върху оценките на параметрите на модел (2.16) е драстично: 5.13216 и -0.30915. Оценката на коефициента на корелация между наблюдаваните и предсказаните по модела стойности е 0.1618. Както оценките на параметрите, така и ко-

ефициента на корелация са статистически значими, което следва от съответните им $p - value$. От плота вляво на фиг. 2.1 се вижда, че прогностичните качества на така оценения модел биха били катастрофални, което би довело до неправдоподобни изводи с непредсказуеми последствия.

Таблица 2.2: Оценка на параметрите на регресионен модел y = 1 + 3x + ϵ по замърсени данни:

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	Signif.
(Intercept)	5.13216	0.63166	8.125	7.59e-13	***
x	-0.30915	0.06739	-4.587	1.21e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.14 on 109 degrees of freedom
Multiple R-squared: 0.1618, Adjusted R-squared: 0.1541
F-statistic: 21.04 on 1 and 109 DF, p-value: 1.206e-05

Стандартните диагностични плотове за проверка на предположенията за валидност на модела по замърсените данни, представени на фиг.2.3 подсказват неадекватност на модела или проблем от несъгласувани наблюдения в данните.

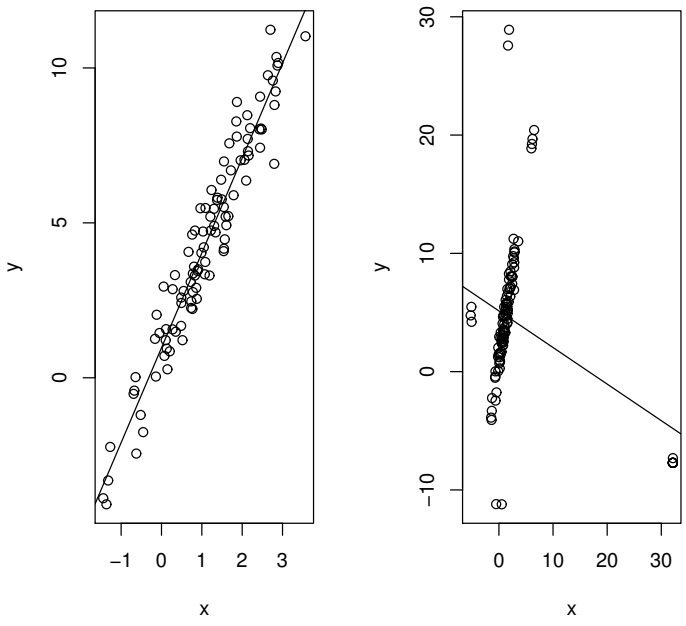
В двумерното и тримерно пространство проблема с несъгласуваните наблюдения е лесно отстраним поради възможност за визуален контрол. Несъгласуваните наблюдения са сериозен проблем в пространства с по-висока размерност, в които броя на предикторните променливи $p > 3$, които са по-често срещаните в практиката. В подобни случаи е целесъобразно използването на робастните методи, които автоматично редуцират (претеглят) влиянието на несъгласуваните наблюдения. Един такъв метод ще разгледаме в следващия параграф.

2.3 LTS регресионни оценки

Дефиниция 2.2 *Оценката $\hat{\beta}_{LTS}$ на параметъра β на регресионния модел по метода най-малката сума от орязани квадрати (the Least Trimmed Squares, LTS, Rousseeuw, 1984) се дефинира като:*

$$\min_{\beta} \sum_{i=1}^{k} (y_{\nu(i)} - x_{\nu(i)}^T \beta)^2. \tag{2.17}$$

където $(y_{\nu(i)} - x_{\nu(i)}^T \beta)^2 \leq \ldots \leq (y_{\nu(k)} - x_{\nu(k)}^T \beta)^2 \leq \ldots \leq (y_{\nu(n)} - x_{\nu(n)}^T \beta)^2$ са наредените стойности на остатъците $(y_i - x_i^T \beta)^2$ при фиксирана стойност на β, $\nu = (\nu(1), \ldots, \nu(n))$ е пермутацията на индексите на наблюденията, която зависи от β, k е параметър на орязване, който се избира в интервала $\lfloor \frac{n+p+1}{2} \rfloor \leq k \leq n$, $\lfloor a \rfloor$ цялата част на a.



Фигура 2.1: Плот вляво: генерирани данни по модел (2.16). Плот вдясно: генерирани данни по модел (2.16) със замърсяване. Правите линии са оценени по МНК с процедурата lm от R.

LTS стандартната грешка на линейния регресионнен модел се дефинира като

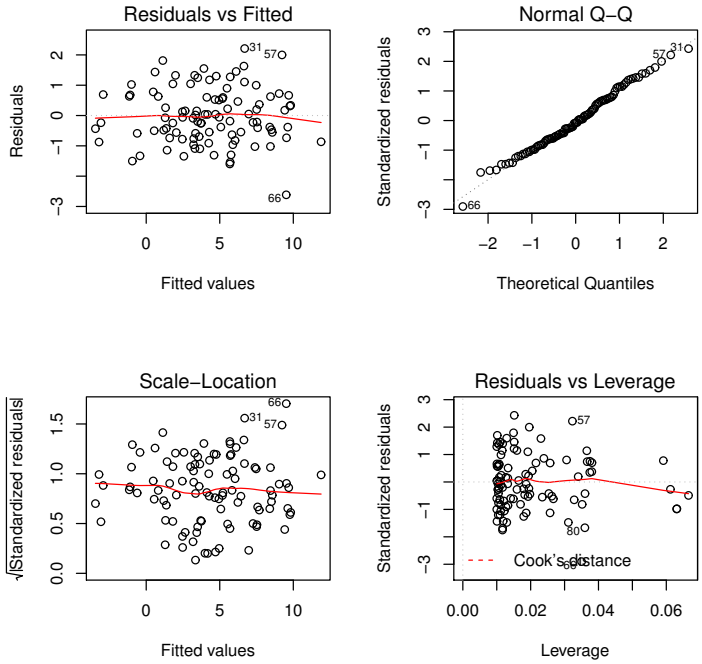
$$\hat{\sigma}_{LTS} = \sqrt{\frac{1}{k-p}\sum_{i=1}^{k} r^2_{\nu(i)}(\widehat{\beta}_{\text{LTS}})}.$$

От дефиницията на LTS следва еквивалентното представяне

$$\min_{\beta}\sum_{i=1}^{k}(y_{\nu(i)} - x^T_{\nu(i)}\beta)^2 = \min_{\beta}\min_{I\in\mathcal{I}_k}\sum_{i\in I}(y_i - x_i^T\beta)^2 = \min_{I\in\mathcal{I}_k}\min_{\beta}\sum_{i\in I}(y_i - x_i^T\beta)^2, \quad (2.18)$$

където $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, а $\mathcal{I}_k$ е множеството от всички подмножества от k индекса.

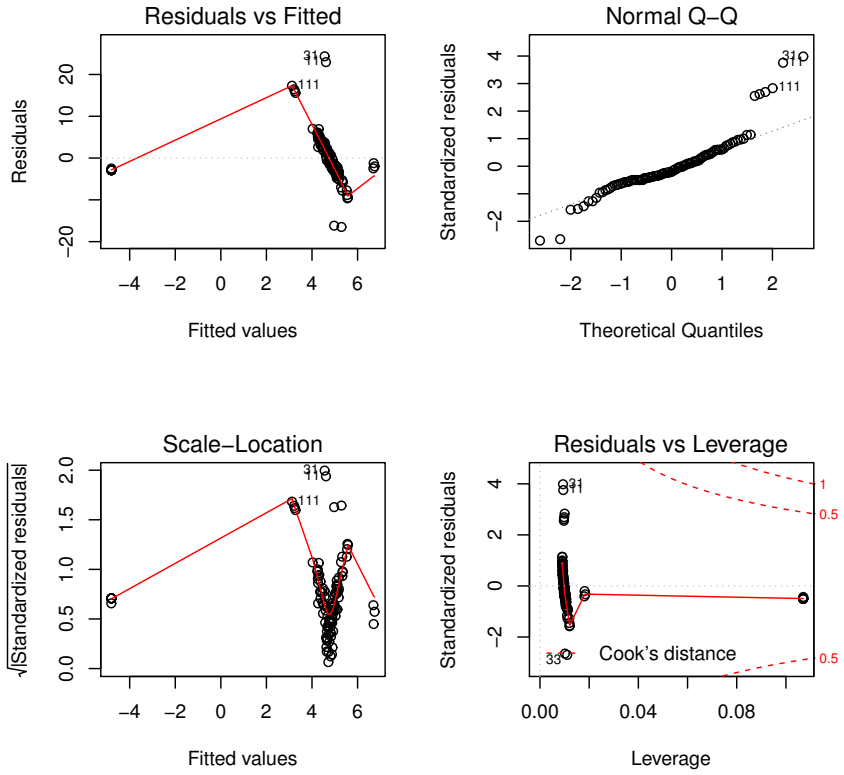
От това представяне на целевата функция следва, че LTS оценката се дефинира като оценка по МНК върху подизвадка с обем k от всевъзможните $\binom{n}{k}$ подизвадки. Това означава, че за малки стойности на n и k е възможно да бъде намерена точната LTS оценка на β. За тази цел е необходимо да бъдат намерени оценките по МНК по всевъзможните $\binom{n}{k}$ подизвадки. При големи стойности на n и k това не е възможно, поради което се търси приближена LTS оценка на β. Rousseeuw and van Drissen (2000a)



Фигура 2.2: Диагностични регресионни плотове, получени чрез процедурата lm за генерираните данни по модел (2.16).

предлагат FAST-LTS алгоритъма за приближено намиране на LTS оценка на β, който ще бъде разгледан по-долу.

Понеже LTS оценката $\hat{\beta}$ е формирана по мажоритарната част от k наблюдения, то останалите $n-k$ наблюдения би трябвало да бъдат третирани като несъгласувани с регресионния модел. LTS оценката ще притежава максимална прагова точка от $1/2$, ако параметъра на орязване k е в границата $\lfloor \frac{n+p+1}{2} \rfloor \leq k \leq \lfloor \frac{n+p+2}{2} \rfloor$. LTS оценката ще бъде със занижена ефективност при този избор на параметъра k, тъй като ще бъде основана почти на $1/2$ от данните. Малко вероятно е $n-k$ от наблюденията да са несъгласувани наблюдения, нещо повече в извадката може изобщо да няма несъгласувани наблюдения в конкретния линеен регресионен модел. По-вероятно е една голяма част от идентифицираните като несъгласувани $n-k$ на брой наблюдения да са близко разположени до мажоритарната част от онези k наблюдения, по които е определена приближената LTS оценка. Твърде голяма неопределеност има в представата ни за "близко разположени до мажоритарната част"в $p+1$-мерното пространство на данните на регресионния модел. Естествената мярка за близост в многомерното пространство на данните се основава на LTS стандартизираните регресионни остатъци $r_i/\hat{\sigma}_{LTS}$ за $i = 1, \ldots, n$.



Фигура 2.3: Диагностични регресионни плотове, получени чрез процедурата lm по замърсените данни по модел (2.16).

2.3.1 Едностъпково подобрение на LTS регресионните оценки

Ако регресионния модел е вярно дефиниран, то стандартизираните LTS регресионни остатъци са симетрично разпределени спрямо нулата, а при голям обем на извадката n са стандартно нормално разпределени. Поради тази причина се пресмятат и използват LTS тегла за разкриване на несъгласувани наблюдения:

$$w_i^{LTS} = \begin{cases} 1 & |r_i/\hat{\sigma}_{LTS}| \leq 2.5 \\ 0 & \text{в противен случай.} \end{cases}$$

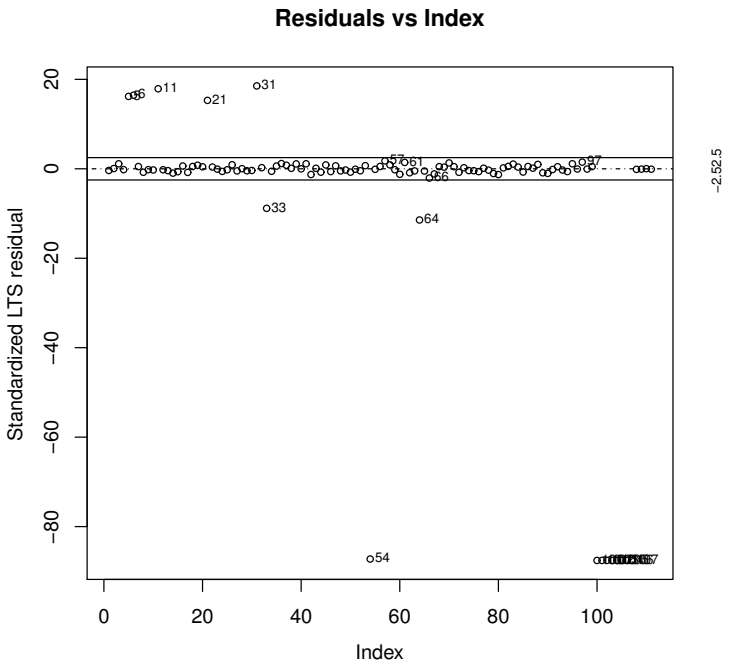
Изборът на константата 2.5 е продиктуван от предположението за стандартно нормално разпределение на регресионните остатъци, тъй като 0.975% от квантилите на $N(0,1)$ се намират в интервала $(-3,3)$.

Дефиниция 2.3 *Подобрената LTS оценка на стандартната грешка на линейния*

регресионнен модел се дефинира като

$$\hat{\sigma}^*_{LTS} = \sqrt{\frac{1}{\sum_{i=1}^n w_i^{LTS} - p} \sum_{i=1}^n w_i^{LTS} r_i^2(\widehat{\beta}_{\mathrm{LTS}})}.$$

Като диагностично средство за разкриване на несъгласувани наблюдения в методологията на LTS регресионния анализ е прието използването на двумерен плот. По числовата ос Ox на този плот се нанасят номерата на наблюденията за $i = 1, \ldots, n$, а на оста Oy са дават стандартизираните LTS регресионни остатъци $r_i/\hat{\sigma}^*_{LTS}$ и правите $y = -2.5$ и $y = 2.5$, както е на Фиг.2.4. Наблюденията, които попадат извън тези две прави се класифицират като LTS несъгласувани наблюдения.



Фигура 2.4: Диагностичен плот: номер на наблюдения на абсцисната ос Ox срещу стандартизирани LTS регресионни остатъци $r_i(\widehat{\beta}_{\mathrm{LTS}})/\hat{\sigma}^*_{LTS}$ на ординатната ос Oy.

Понеже е малко вероятно процента на несъгласуваните наблюдения в реални данни да бъде почти 50%, то е разумно провеждането на LTS регресионен анализ с максимална прагова точка от $1/2$ само с разузнавателна цел. Ако след провеждането на едностъпковия LTS претеглен МНК процента на несъгласувани наблюдения остава твърде висок, то е уместно използването на друг тип регресионни модели, които използват смес от разпределения на регресионни модели или клъстерни регресионни модели за анализ на данни.

Ако диагностичния LTS двумерен плот на стандартизираните остатъци показва нисък процент α на несъгласуваните наблюдения, тогава е препоръчително параметърът на орязване да бъде дефиниран като $k = n(1 - \alpha)$, където $0 \leq \alpha < 1/2$. Така

например ако $\alpha = 0.1$ тогава $k = 0.9n$, т.е., LTS оценката ще бъде построена по 90% от данните.

Разглеждайки LTS теглата w_i^{LTS} за $i = 1, \ldots, n$ като априорни тегла е целесъобразно и уместно провеждането на допълнителен анализ на оригиналните данни, по методологията на претеглените МНК. Чрез този анализ се цели подобряване на ефективността на LTS оценката на неизвестните параметри, β и σ на регресионния модел, вследствие на което доверителни интервали и проверка на линейни хипотези относно β ще бъдат по-надеждни и ефективни. В робастната статистика е прието тази процедура да се нарича едностъпков LTS претеглен регресионен анализ по МНК.

Дефиниция 2.4 *Оценката по едностъпков претеглен МНК на параметъра β чрез LTS теглата се дефинира като*

$$\min_{\beta} \sum_{i=1}^{n} w_i^{LTS} r_i^2(\beta).$$

2.3.2 Продължение: анализ на моделни данни (2.16)

В този параграф са анализирани генерираните по модел 2.16 регресионни данни чрез LTS методологията, както и получените от тях данни, замърсени с несъгласувани наблюдения, чрез едностъпковия претеглен МНК, използвайки LTS теглата w_i^{LTS} за $i = 1, \ldots, n$. Резултата от процедурата *ltsReg* за линеен регресионен анализ по LTS от библиотека robustbase на програмната среда R е даден в Таблица 2.3. Оценката $\hat{\beta}_{\text{мнк}}^{wLTS}$ на параметрите е 0.99007 и 3.00708, съответно, докато оценката на коефициента на корелация е 0.9635. Тези оценки са статистически значими според критерия на Стюдент, тъй като съответните им $p-value$, дадени в предпоследната колона на таблицата са много по-малки от стандартните нива на съгласие $0.05, 0.01, 0.001$. Според F-теста на Фишер хипотезата H_o се отхвърля, понеже съответната $p - value :< 2.2e - 16$ е много малка в сравнение със стандартните нива на съгласие. Вижда се, че оценките по този метод съвпадат с оценките по МНК, пресметнати по генерираните по модел (2.16) данни без замърсяване. Регресионната линия в синьо, съответна на тези оценки е дадена на двата панела на фиг. 2.1. На панела вдясно е дадена правата линия, оценена по МНК по замърсените данни. Вижда се, че оценената линия е наклонена към наблюденията с малки и големи стойности в предикторната променлива, т.е., наблюденията отдалечени от мажоритарната част на предикторите.

2.3.3 FAST-LTS алгоритъм

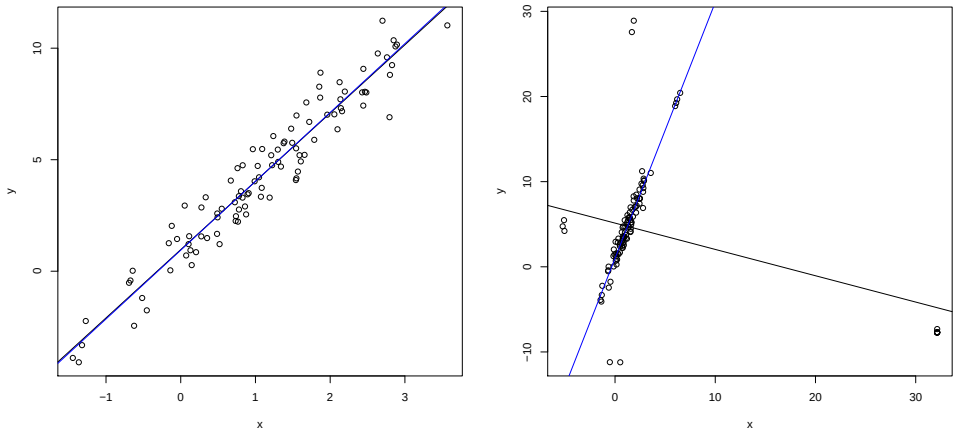
Rousseeuw and van Driessem (1999) предлагат FAST-LTS алгоритъм за приближено пресмятане на MCD оценката на многомерната средна и ковариационна матрица. Алгоритъмът се основава на следната

Процедура на "концентрация":

Таблица 2.3: Оценка на параметрите на регресионен модел $y = 1+3x+\epsilon$ по замърсени данни, чрез едностъпковия претеглен МНК с LTS теглата w_i^{LTS} за $i = 1, \ldots, n$:

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	Signif.
(Intercept)	0.99007	0.11980	8.264	9.47e-13	***
x	3.00708	0.06071	49.533	$<$ 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8772 on 93 degrees of freedom
Multiple R-squared: 0.9635, Adjusted R-squared: 0.9631
F-statistic: 2454 on 1 and 93 DF, p-value: $<$ 2.2e-16



Фигура 2.5: Плот вляво: генерирани данни по модел (2.16). Плот вдясно: генерирани данни по модел (2.16) със замърсяване. Правите линии в синьо, съответстват на $\hat{\beta}^{wLTS}_{\text{МНК}}$ оценката с процедурата *ltsReg* от библиотеката robustbase на R, а правата линия в черно на панела вдясно е по МНК.

(a) Нека $Z_{n\times p} = \{(y_1, x_1), \ldots (y_n, x_n)\},\ y_i \in R^1,\ x_i \in R^p$, а $\hat{\beta}^{(old)}$ е произволно избран вектор от стойности;

(b) Нека $r^2_{\nu(1)}(\hat{\beta}^{(old)}) \leq \ldots \leq r^2_{\nu(k)}(\hat{\beta}^{(old)}) \leq \ldots \leq r^2_{\nu(n)}(\hat{\beta}^{(old)})$ са сортираните остатъци $r_i^2(\hat{\beta}^{(old)}) = (y_i - x_i^T\hat{\beta}^{(old)})^2$ за $i = 1, \ldots, n$ във възходящ ред, $S^{(old)} = \sum_{i=1}^{k} r^2_{\nu(i)}(\hat{\beta}^{(old)})$ и $H^{new} := \{(y_{\nu(1)}, x_{\nu(1)}), \ldots, (y_{\nu(k)}, x_{\nu(k)})\}$;

(c) Нека $\hat{\beta}^{(new)}$ е оценката по МНК, изчислена по наблюденията от H^{new} и $\hat{\beta}^{(old)} := \hat{\beta}^{(new)}$;

(d) Повторение на действията (b)-(c) до достигане на сходимост;

Твърдение 2.5 $S^{(new)} \leq S^{(old)}$.

Доказателство. От дефинициите на пермутацията ν и $\hat{\beta}^{(new)}$ следва

$$S^{(new)} = \sum_{i=1}^{k} r^2_{\nu(i)}\left(\hat{\beta}^{(new)}\right) \leq \sum_{i=1}^{k} r^2_{\nu(i)}\left(\hat{\beta}^{(old)}\right) = S^{(old)}.$$

Стартирайки с $\hat{\beta}^{(old)}$ след няколко повторения на стъпките (b)-(c) в процедурата на "концентрация"се достига до сходимост. Следователно на всяка итерация имаме намаляване на целевата функция. Използваният алгоритъм е сходящ, тъй като имаме краен брой подизвадки с обем k на дадената извадка. Многократното изпълнение на тази процедура с различни стойности на $\hat{\beta}^{(old)}$ води до намирането на подизвадка от k наблюдения, за която целевата стойност на LTS е минимална.

Вместо произволен избор на вектора $\hat{\beta}^{(old)}$ е уместно използването на точното решение на системата $\widetilde{Y} = \widetilde{X}\beta$, където $\widetilde{Y}$ и $\widetilde{X}$ са векторът и матрицата от p случайно избраните наблюдения $\{(y_{i_1}, x^T_{i_1}), \ldots, (y_{i_p}, x^T_{i_p})\} \subset Z_{n\times p}$, за които $\det(\widetilde{X}) \neq 0$. Ако $\det(\widetilde{X}) = 0$ се избира нова случайна подизвадка. Ще отбележим, че вероятността за извличане на случайна подизвадка без наличие на несъгласувани наблюдения в нея е максимална, когато извличаме точно p наблюдения от n.

За ефективна програмна реализация на тази процедура е необходимо съхраняването на предисторията на извлечените подизвадки (индексите на съответните k наблюдения на подизвадките), за да не бъде подлагана повторно на процедурата на концентраци, вече използвана подизвадка.

В случаите, когато обемът на извадката n е много голям, се препоръчва разбиване на извадката по случаен начин без повторение на наблюденията в m подизвадки с (почти) равни обеми от $\widetilde{n}$ наблюдения. Върху всяка подизвадка се провежда описаната по-горе процедура на концентрация, за да бъдат определени индексите на наблюденията "без наличието на несъгласувани наблюдения"с обеми $\widetilde{k}$, такива че $k = m\widetilde{k}$. От условието за максимална прагова точка от $1/2$ следва, че обемите на подизвадките трябва да удовлетворяват изискването $\lfloor\frac{\widetilde{n}+p+1}{2}\rfloor \leq \widetilde{k} \leq \lfloor\frac{\widetilde{n}+p+2}{2}\rfloor$. Така определените m подизвадки, с обеми $\widetilde{k}$, се обединяват в извадка с обем $k = m\widetilde{k}$, върху която се прилагат стъпките (b)-(c), като за $\hat{\beta}^{(old)}$ се използва оценката по МНК на тези k наблюдения. За гарантиране на максимална прагова точка от $1/2$ следва, че $\lfloor\frac{n+p+1}{2}\rfloor \leq k \leq \lfloor\frac{n+p+2}{2}\rfloor$.

2.3.4 LTS класификация на наблюденията в линейната регресия.

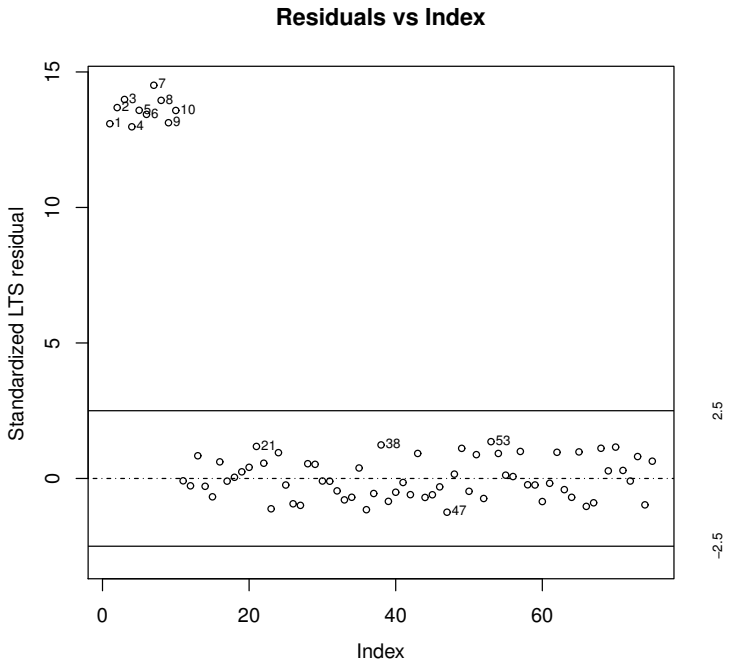
В LTS линейния регресионен анализ е приета следната класификация на наблюденията:

(1) Регулярни наблюдения са наблюденията, за които $|r_i/\hat{\sigma}^*_{LTS}| \leq 2.5$;

(2) Несъгласувани наблюдения са наблюденията, чиито стандартизирани регресионни остатъци $|r_i/\hat{\sigma}^*_{LTS}| > 2.5$;

LTS класификацията на наблюденията е демонстрирана върху данните на Hawkins, Bradu and Kass (1984). Тези данни се състоят от 75 наблюдения, следващи модел на линейна множествена регресия с 4 предикторни променливи. Наблюденията от 1 до 10 са заместени чрез небалансирани наблюдения, т.е. внесени са груби грешки в предикторните променливи x_1, x_2, x_3. Наблюденията от 11 до 14 представляват също внесени груби грешки в предикторните променливи x_1, x_2, x_3, но с тази разлика, че съответните стойности на зависимата променлива y са съгласувани с модела, основан на останалите от 15 до 75 наблюдения.

На фиг. 2.6 е даден диагностичен плот от изхода на *ltsReg* процедурата от библиотеката robustbase на програмната среда R. На оста Ox са дадени номерата на наблюденията срещу стандартизирани LTS регресионни остатъци $\frac{r_i(\hat{\beta}_{LTS})}{\hat{\sigma}^*_{LTS}}$ на оста Oy. Вижда се, че всички несъгласувани наблюдения са идентифицирани вярно. За съжаление, LTS теглата, основани на стандартизираните LTS остатъци не дават информация за типа на несъгласуваните наблюдения, дали са в зависимата променлива y или са в матрицата X на предикторните променливи. Един подход за разкриване на несъгласувани наблюдения в матрицата X ще бъде разгледан в следващата глава.



Фигура 2.6: Диагностичен плот: номер на наблюдения (Ox) срещу стандартизирани LTS регресионни остатъци $r_i(\hat{\beta}_{LTS})/\hat{\sigma}^*_{LTS}$ (Oy).

Глава 3

Робастна MCD оценка на ковариационна матрица.

Нека $\{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\}$ е извадка от независими наблюдения, която е извлечена от p-мерното нормално разпределение $N_p(\mu, \mathbf{\Sigma})$, където $\mu = (\mu_1, \ldots, \mu_p)^T$ и $\mathbf{\Sigma}_{p\times p} = (\sigma_{kl})$ за $k, l = 1, \ldots, p$. Наблюдението $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ ще разглеждаме като i-ти ред на матрицата от данни $X_{n\times p}$. Плътността на $N_p(\mu, \mathbf{\Sigma})$ е

$$\phi(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{p/2}(\det(\mathbf{\Sigma}))^{1/2}} \exp\Big(-\frac{(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}{2}\Big)$$

Дефиниция 3.1 *Максимално правдоподобната оценка* $(\widehat{\mu}, \widehat{\mathbf{\Sigma}})$ *на многомерната средна и ковариационна матрица* $(\mu, \mathbf{\Sigma})$ *се дефинира като:*

$$\max_{\mu,\Sigma} L(\mu, \mathbf{\Sigma}) \;=\; \max_{\mu,\Sigma} \prod_{i=1}^{n} \phi(\mathbf{x}_i, \mu, \mathbf{\Sigma}).$$

Понеже $\log(.)$ е монотонна функция, вместо определяне на максимум на функцията на правдоподобие $L(\mu, \mathbf{\Sigma})$ се търси максимум на $\log(L(\mu\mathbf{\Sigma}))$, което е по-добре обусловена задача от изчислителна гледна точка, или минимум на $-\log(L(\mu, \mathbf{\Sigma}))$, също така броят на необходимите пресмятания при сумиране са по-малко отколкото при произведение.

$$\begin{aligned}\min_{\mu,\Sigma}\{-\log(L(\mu, \mathbf{\Sigma}))\} &= \min_{\mu,\Sigma} \sum_{i=1}^{n} -\log\big(\phi\big(\mathbf{x}_i, \mu, \mathbf{\Sigma}\big)\big) \\ &= \min_{\mu,\Sigma}\{\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu) + \frac{n}{2}\log(\det(\mathbf{\Sigma})) + \frac{np}{2}\log(2\pi)\}.\end{aligned}$$

Решението на оптимизационната задача се достига за

$$\widehat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad \text{и} \qquad \widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\widehat{\mu})(\mathbf{x}_i-\widehat{\mu})^T,$$

където

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{ik} \qquad \text{и} \qquad \widehat{\sigma}_{kl} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_{ik}-\widehat{\mu}_k)(\mathbf{x}_{il}-\widehat{\mu}_l)$$
$$\text{за} \qquad k,l = 1,\ldots,p.$$

За да бъде $\widehat{\mathbf{\Sigma}}$ неизместена оценка на $\mathbf{\Sigma}$ се използват съответно

$$\widehat{\sigma}_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_{ik}-\widehat{\mu}_k)(\mathbf{x}_{il}-\widehat{\mu}_l) \qquad \text{и} \qquad \widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{kl}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\widehat{\mu})(\mathbf{x}_i-\widehat{\mu})^T,$$

3.1 Разкриване на многомерни несъгласувани наблюдения

Дефиниция 3.2 *Разстоянието на Махаланобис на наблюдението* $\mathbf{x}_i$ *от центъра на данните* μ *се дефинира като*

$$MD_i^2 = MD^2(\mathbf{x}_i,\mu,\mathbf{\Sigma}) := (\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu).$$

Разстоянието на Махаланобис MD_i^2 е χ_p^2 разпределена случайна величина с p степени на свобода, тъй като наблюденията x_i са извадка от p-мерното нормално разпределение $N_p(\mu,\mathbf{\Sigma})$.

Класическата процедура за разкриване на многомерни несъгласувани наблюдения в извадката се основава на сравнението на извадковото разстояние на Махаланобис $\widehat{MD}_i^2 = (\mathbf{x}_i-\widehat{\mu})^T\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{x}_i-\widehat{\mu})$ с $1-\alpha$ с квантила на χ_p^2 разпределението. Ако $\widehat{MD}_i^2 \le \chi^2_{1-\frac{\alpha}{2},p}$, то i-то наблюдение се разглежда като съгласувано, в противен случай несъгласувано. Стандартното ниво на съгласие е $\alpha = 0.05$.

От дефиницията на оценката $(\widehat{\mu},\widehat{\mathbf{\Sigma}})$ следва, че ако поне една от координатите на някое от наблюденията на извадката приеме произволно голяма стойност, тогава оценката също ще приема произволно големи стойности. Това би повлияло на вземането на неправдоподобни решение при проверката на хипотезата за съгласуваност на наблюденията. За преодоляването на този проблем са предложени различни робастни оценки. Една такава оценка на многомерната средна и ковариационна матрица, която не се влияе от несъгласуваните наблюдения в извадката е робастната MCD (minimum covariance determinant), предложена от Rousseeuw (1984). Характерното за тази оценка е, че тя е дефинирана върху подизвадка от наблюденията, удовлетворяваща определени екстремални свойства.

3.1.1 MCD оценка на многомерната средна и ковариационна матрица

Дефиниция 3.3 *Робастната MCD (minimum covariance determinant estimator, Rousseeuw, 1984) оценка на многомерната средна и ковариационна матрица се дефинира като:*

$$\widehat{\mathbf{m}}_{\mathrm{mcd}} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_{\nu(i)} \quad и \quad \widehat{\mathbf{\Sigma}}_{\mathrm{mcd}} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{x}_{\nu(i)} - \mathbf{m}_{\mathrm{mcd}})(\mathbf{x}_{\nu(i)} - \mathbf{m}_{\mathrm{mcd}})^T,$$

където пермутацията $\nu = (\nu(1), \ldots, \nu(n))$ от индекси на наблюденията $\{1, \ldots, n\}$ е такава, че $\det(\widehat{\mathbf{\Sigma}}_{\mathrm{mcd}})$ е минимална.

Нека $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, а $\mathcal{I}_k$ е множеството от всички подмножества от k индекса от n. От дефиницията следва, че $\widehat{\mathbf{\Sigma}}_{\mathrm{mcd}}$ се определя като класическа оценка на ковариационна матрица върху подизвадка с обем k от всевъзможните $\binom{n}{k}$ подизвадки, за която $\det(\mathbf{\Sigma})$ е минимална, т.е.,

$$\begin{aligned}
\det(\widehat{\mathbf{\Sigma}}_{\mathrm{mcd}}) &= \min_{I\in\mathcal{I}_k}\det\left(\frac{1}{k}\sum_{i\in I}(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})^T\right) \\
&\text{или} \\
(\widehat{\mu}_{\mathrm{mcd}}, \widehat{\mathbf{\Sigma}}_{\mathrm{mcd}}) &:= \operatorname*{argmin}_{(\mu,\mathbf{\Sigma})}\min_{I\in\mathcal{I}_k}\det\left(\frac{1}{k}\sum_{i\in I}(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})^T\right) \\
&= \min_{I\in\mathcal{I}_k}\operatorname*{argmin}_{(\mu,\mathbf{\Sigma})}\det\left(\frac{1}{k}\sum_{i\in I}(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})(\mathbf{x}_i - \mathbf{m}_{\mathrm{mcd}})^T\right)
\end{aligned}$$

Това означава, че за малки стойности на n и k е възможно да бъде намерена точната MCD оценка $\widehat{\mathbf{\Sigma}}_{\mathrm{mcd}}$ като бъдат намерени класическите оценки на $(\widehat{\mu}, \widehat{\mathbf{\Sigma}})$ по всевъзможните $\binom{n}{k}$ подизвадки от наблюдения и се вземе онази $\widehat{\mathbf{\Sigma}}$, за която $\det(\mathbf{\Sigma})$ е минимална. При големи стойности на n и k това не е възможно, поради което се търси приближена MCD оценка на $(\mu, \mathbf{\Sigma})$.

MCD оценката на многомерната средна и ковариационна матрица достига максимална прагова точка от $1/2$, когато параметърът на орязване k е в границата $\lfloor (n+p+1)/2 \rfloor \leq k \leq \lfloor (n+p+2)/2 \rfloor$. Ще отбележим, че MCD оценката ще бъде със занижена ефективност, тъй като ще бъде основана почти на половината от данните в извадката.

3.1.2 Едностъпково подобрение на MCD оценката на ковариационната матрица

Понеже MCD оценката $\widehat{\mathbf{\Sigma}}_{\mathrm{mcd}}$ на ковариационната матрица е формирана по мажоритарната част от k наблюдения, то останалите $n-k$ наблюдения би трябвало да бъдат

третирани като несъгласувани, което е малко вероятно. Извадката може да не съдържа несъгласувани наблюдения. По-вероятно е една голяма част от тези $n-k$ на брой наблюдения да са близко разположени до мажоритарната част от останалите k наблюдения, дефиниращи MCD оценката. Мярката за близост в p-мерното пространство на данните $X_{n\times p}$ се основава на разстоянието на Махаланобис, робастния аналог на който се дефинира като

$$RD_i^2 = RD^2(\mathbf{x}_i) = (\mathbf{x}_i - \widehat{\mu}_{\text{mcd}})^T \widehat{\mathbf{\Sigma}}_{\text{mcd}}^{-1}(\mathbf{x}_i - \widehat{\mu}_{\text{mcd}}) \qquad \text{за i=1,. . . ,n.}$$

С помощта на робастните разстояния на Махаланобис се дефинират бинарните тегла

$$w_i^{mcd} = \begin{cases} 1 & RD_i \leq \sqrt{\chi^2(p, 0.975)} \\ 0 & \text{в противен случай,} \end{cases}$$

Разглеждайки MCD теглата w_i^{mcd} за $i = 1, \ldots, n$ като априорни е целесъобразно и уместно провеждането на допълнителен анализ над данните от оригиналната извадка с тегла. Чрез този анализ се цели подобряване на ефективността на MCD оценката $(\widehat{\mu}, \widehat{\mathbf{\Sigma}})$. В робастната статистика е прието тази процедура да се нарича едностъпкова претеглена MCD оценка.

Дефиниция 3.4 *Едностъпковата претеглена MCD оценка* $(\widehat{\mu}_{\text{wmcd}}, \widehat{\mathbf{\Sigma}}_{\text{wmcd}})$ *на* $(\mu, \mathbf{\Sigma})$ *се дефинира като*

$$\begin{aligned} \widehat{\mu}_{\text{wmcd}} &= \frac{1}{\sum_{i=1}^n w_i^{mcd}} \sum_{i=1}^n w_i^{mcd} \mathbf{x}_i \\ \widehat{\mathbf{\Sigma}}_{\text{wmcd}} &= \frac{1}{\sum_{i=1}^n w_i^{mcd} - 1} \sum_{i=1}^n w_i^{mcd} (\mathbf{x}_i - \widehat{\mu}_{\text{wmcd}})(\mathbf{x}_i - \widehat{\mu}_{\text{wmcd}})^T. \end{aligned}$$

Понеже е малко вероятно процента на несъгласуваните наблюдения в реални данни да бъде почти 50%, то е разумно MCD оценяването да бъде проведено с максимална прагова точка от $1/2$ само с разузнавателна цел. Ако процента на несъгласуваните наблюдения α е малък, тогава е препоръчително параметъра на орязване да бъде дефиниран като $k = n(1-\alpha)$, където $0 \leq \alpha < 1/2$. Така например ако $\alpha = 0.1$ тогава $k = 0.9n$, т.е., MCD оценката ще бъде построена по 90% от данните. Ако след проверката за съгласуваност на наблюденията с робастното разстояние на Махаланобис, основано на едностъпкова претеглена MCD оценка, процента на несъгласувани наблюдения остава твърде висок, то е уместно използването на друг тип анализ на данни като клъстерни методи или моделиране на данните със смеси на p-мерни нормални разпределения.

3.1.3 FAST-MCD алгоритъм

Rousseeuw and van Drissen (2000b) предлагат FAST-MCD алгоритъма за приближено намиране на MCD оценката на $(\mu, \mathbf{\Sigma})$. В основата на този алгоритъм е залегнала

следната процедура на концентрация, която ще разгледаме по-долу за многомерното нормално разпределение.

Процедура на "концентрация":

(a) Нека $X_{n\times p} = \{x_1, \dots x_n\},\ x_i \in R^p$, а $\left(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\right)$ е произволно избрана оценка;

(b) Нека $-\log\big(\phi\big(\mathbf{x}_{\nu(1)}, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big) \le \ldots \le -\log\big(\phi\big(\mathbf{x}_{\nu(k)}, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big)$ са сортираните във възходящ ред $-\log\big(\phi\big(\mathbf{x}_i, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big),\ i = 1, \dots, n$, $Q^{(old)} = \sum_{i=1}^{k} -\log\big(\phi\big(\mathbf{x}_{\nu(1)}, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big)$ и $H^{new} := \{\mathbf{x}_{\nu(1)}, \dots, \mathbf{x}_{\nu(k)}\}$;

(c) Нека $(\widehat{\mu}^{(new)}, \widehat{\mathbf{\Sigma}}^{(new)})$ е оценката по ММП, изчислена по наблюденията от H^{new} и $\big(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big) := \big(\widehat{\mu}^{(new)}, \widehat{\mathbf{\Sigma}}^{(new)}\big)$;

(d) Повторение на (b)-(c) до достигане на сходимост;

Твърдение 3.5 $Q^{(new)} \le Q^{(old)}$.

Доказателство. От дефинициите на пермутацията $\nu(.)$ и $(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)})$ следва

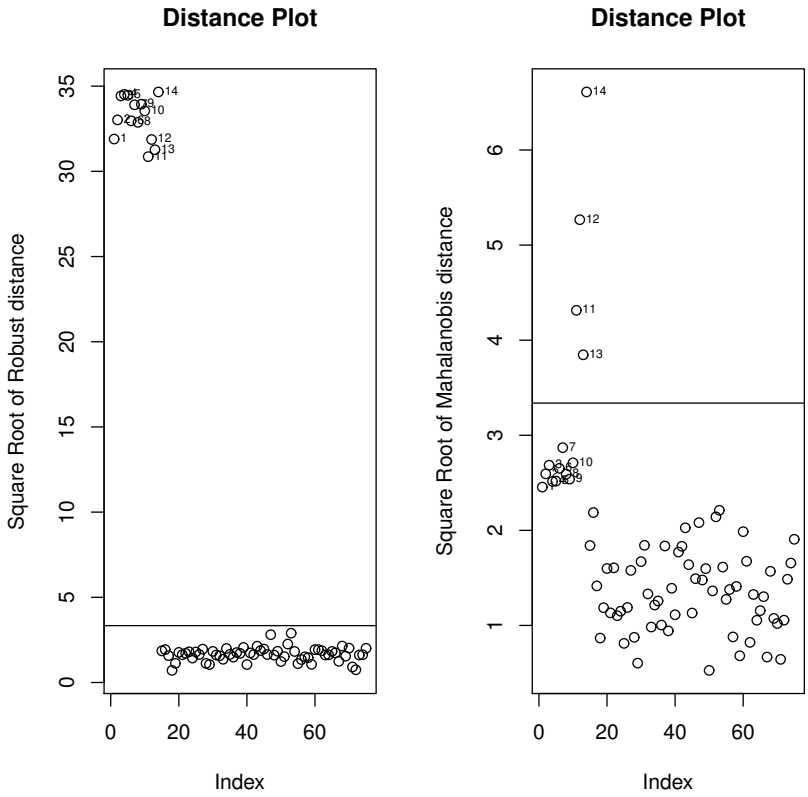
$$Q^{(new)} = \sum_{i=1}^{k} -\log\big(\phi\big(\mathbf{x}_{\nu(i)}, \widehat{\mu}^{(new)}, \widehat{\mathbf{\Sigma}}^{(new)}\big)\big) \le \sum_{i=1}^{k} -\log\big(\phi\big(\mathbf{x}_{\nu(i)}, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big) = Q^{(old)}.$$

Стартирайки с $(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)})$ след няколко повторения на стъпките (b)-(c) в процедурата на концентрация се достига до сходимост, т.е., $-\log\big(\phi\big(\mathbf{x}_{\nu(i)}, \widehat{\mu}^{(new)}, \widehat{\mathbf{\Sigma}}^{(new)}\big)\big) = -\log\big(\phi\big(\mathbf{x}_{\nu(i)}, \widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)}\big)\big)$. Ще отбележим, че $\det(\widehat{\mathbf{\Sigma}}^{(new)}) \neq 0$, понеже всеки p наблюденията са линейно независими с вероятност 1, тъй като са извадка на многомерното нормално разпределение. Многократното повторение на процедурата на "концентрация" с различни стойности на $(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)})$ води до намирането на подизвадка от k наблюдения, за която $Q(.)$ е минимална, което е еквивалентно на достигането на минимум на $\det(\widehat{\mathbf{\Sigma}})$, както показват Vandev and Neykov (2003). Следователно на всяка итерация имаме намаляване на целевата функция. По този начин чрез процедурата на "концентрация" се дефинира редицата $\det(\widehat{\mathbf{\Sigma}}_1) \ge \det(\widehat{\mathbf{\Sigma}}_2) \ge \det(\widehat{\mathbf{\Sigma}}_3) \ge \ldots \ge 0$, която е ограничена отгоре и отдолу и следователно е сходяща. Понеже извадките с обем k са краен брой, то съществува индекс m, за който $\det(\widehat{\mathbf{\Sigma}}_m) = \det(\widehat{\mathbf{\Sigma}}_{m-1})$.

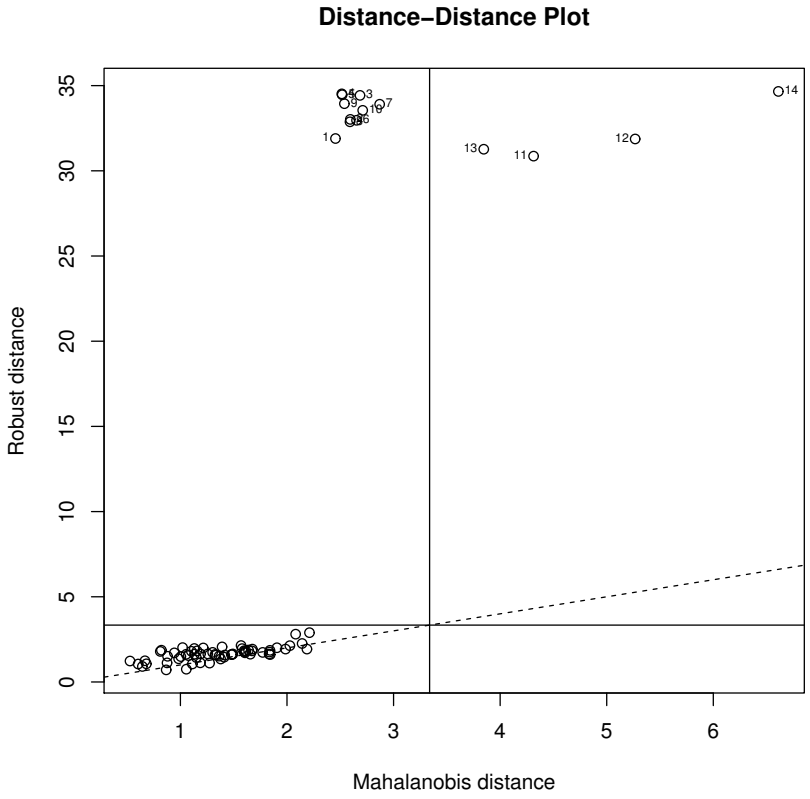
Вместо произволен избор на вектора $(\widehat{\mu}^{(old)}, \widehat{\mathbf{\Sigma}}^{(old)})$ е уместно използването на класическата оценка на μ и $\mathbf{\Sigma}$, изчислени по подизвадка с обем $p+1$. Причината за това е, че вероятността за извличане на случайна подизвадка без наличие на несъгласувани наблюдения в нея е максимална, когато извличаме точно $p+1$ наблюдения от n.

За ефективна програмна реализация на тази процедура е необходимо съхраняването на предисторията на извлечените подизвадки (индексите на съответните k наблюдения на подизвадките), за да не бъде подлагана повторно на процедурата на концентрация вече използвана подизвадка.

В случаите, когато обемът на извадката n е много голям, се препоръчва разбиване на извадката по случаен начин без повторение на наблюденията в m подизвадки с (почти) равни обеми от $\widetilde{n}$ наблюдения. Върху всяка подизвадка се провежда описаната по-горе процедура на концентрация, за да бъдат определени индексите на наблюденията "без наличието на несъгласувани наблюдения"с обеми $\widetilde{k}$, такива че $k = m\widetilde{k}$. От условието за максимална прагова точка от $1/2$ следва, че обемите на подизвадките трябва да удовлетворяват изискването $\lfloor\frac{\widetilde{n}+p+1}{2}\rfloor \leq \widetilde{k} \leq \lfloor\frac{\widetilde{n}+p+2}{2}\rfloor$. Така определените m подизвадки, с обеми $\widetilde{k}$, се обединяват в извадка с обем $k = m\tilde{k}$, върху която се прилагат стъпките (b)-(c), като за $\widehat{\mathbf{\Sigma}}^{(old)}$ се използва класическата оценка на ковариационната матрица на тези k наблюдения. За гарантиране на максимална прагова точка от $1/2$ следва, че параметъра на орязване трябва да избираме както следва $\lfloor\frac{n+p+1}{2}\rfloor \leq k \leq \lfloor\frac{n+p+2}{2}\rfloor$.



Фигура 3.1: Робастни разстояния на Махаланобис RD_i (Оу) срещу номер на наблюдение (панел вляво). Класически разстояния на Махаланобис MD_i (Оу) срещу номер на наблюдение (панел вдясно). Хоризонтални прави на двата панела $y = 3.06$ е квантила на $\sqrt{\chi^2(3, 0.975)}$.



Фигура 3.2: DD-плот: Робастни разстояния RD_i (Oy) срещу класически разстояния MD_i (Ox) на Махаланобис. Вертикалната права $y = 3.06$ и хоризонтална права $x = 3.06$ са квантилите на $\sqrt{\chi^2(3, 0.975)}$.

3.1.4 Пример - разкриване на многомерни несъгласувани наблюдения.

Предимствата на тази процедура за разкриване на несъгласувани наблюдения е демонстрирана върху добре известните в статистическата литература данни на Hawkins, Bradu and Kass (1984). Данните се състоят от 75 наблюдения, следващи модел на линейна множествена регресия с 4 предикторни променливи. Наблюденията от 1 до 10 са заместени чрез небалансирани наблюдения, т.е. внесени са груби грешки в предикторните променливи x_1, x_2, x_3. Наблюденията от 11 до 14 представляват също внесени груби грешки в предикторните променливи x_1, x_2, x_3, но с тази разлика, че съответните стойности на зависимата променлива y са съгласувани с модела, основан на останалите от 15 до 75 наблюдения. В Таблица 3.1 са дадени стойностите на класическото и робастифицирано разстояние на Махаланобис MD_i и RD_i. Наблюденията са класифицирани като несъгласувани или регулярни в пространството на предикторите x_1, x_2, x_3 в зависимост от стойността 0 или 1. на теглата, означени с $MD_i.w$ и $RD_i.w$ в

Таблица 3.1. Забелязва се, че всички несъгласувани наблюдения са идентифицирани правилно чрез RD_i. На панела вляво на фиг. 3.1 са дадени разстоянията RD_i, докато на панела вдясно са дадени MD_i разстоянията. Забелязва се, че чрез MD_i са идентифицирани наблюденията 11, 12, 13 и 14, докато чрез RD_i са идентифицират всички несъгласувани наблюдения.

Съвместното използване на разстоянията MD_i и RD_i за разкриване на несъгласувани наблюдения в матрицата на предикторите $Xn \times p$, е дадено на фиг. 3.2. Общите наблюдения, които двете разстояния идентифицират като съгласувани наблюдения се съдържат в правоъгълника на плота долу вдясно. Разстоянията RD_i и MD_i идентифицират като несъгласувани, съответно, наблюденията над хоризонталната права $y = 3.06$ и надясно от вертикалната права $y = 3.06$.

3.2 Разкриване на несъгласувани наблюдения чрез LTS и MCD оценките в линейни регресионни модели

Разкриването на несъгласуваните (грубите грешки, аутлайерите) наблюдения в данните на линейните регресионни модели представлява важна стъпка за правдоподобни статистически изводи. За целта се използват LTS стандартизираните остатъци и робастните разстояния на Махаланобис, основани на оценката на многомерната средна и ковариационна матрица на матрицата на предикторите $X_{n\times p}$ на линейния регресионен модел. Rousseeuw and Zomeren (1990) предлагат следната класификация на наблюденията в линейната множествена регресия:

(1) регулярни наблюдения (RO) са наблюденията, за които $|r_i(\hat{\beta}_{LTS})/\hat{\sigma}^*_{LTS}| \leq 2.5$ и $RD_i \leq \sqrt{\chi^2(p, 0.975)}$ (малки по абсолютна стойност стандартизирани LTS остатъци и малки RD_i);

(2) вертикални несъгласувани (VO) наблюдения (вертикални аутлайери) в зависимата променлива са наблюденията, чиито стандартизирани регресионни остатъци $|r_i(\hat{\beta}_{LTS})/\hat{\sigma}^*_{LTS}| > 2.5$ и $RD_i \leq \sqrt{\chi^2(p, 0.975)}$ (големи по абсолютна стойност стандартизирани LTS остатъци и малки RD_i);

(3) балансирани наблюдения (GLP), чиито стандартизирани регресионни остатъци $|r_i(\hat{\beta}_{LTS})/\hat{\sigma}^*_{LTS}| \leq 2.5$ и $RD_i > \sqrt{\chi^2(p, 0.975)}$(малки по абсолютна стойност стандартизирани LTS остатъци и големи RD_i);

(4) небалансирани наблюдения (BLP), чиито стандартизирани регресионни остатъци $|r_i(\hat{\beta}_{LTS})/\hat{\sigma}^*_{LTS}| > 2.5$ и $RD_i > \sqrt{\chi^2(p, 0.975)}$ (големи по абсолютна стойност стандартизирани LTS остатъци и големи RD_i); .

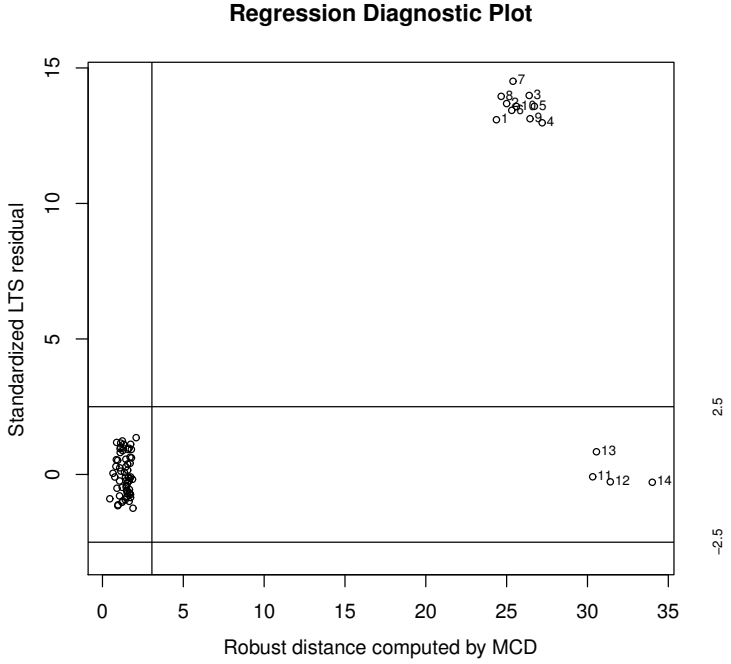
3.2.1 Пример - продължение

С предложената класификация са анализирани литературните данни на Hawkins, Bradu and Kass (1984).

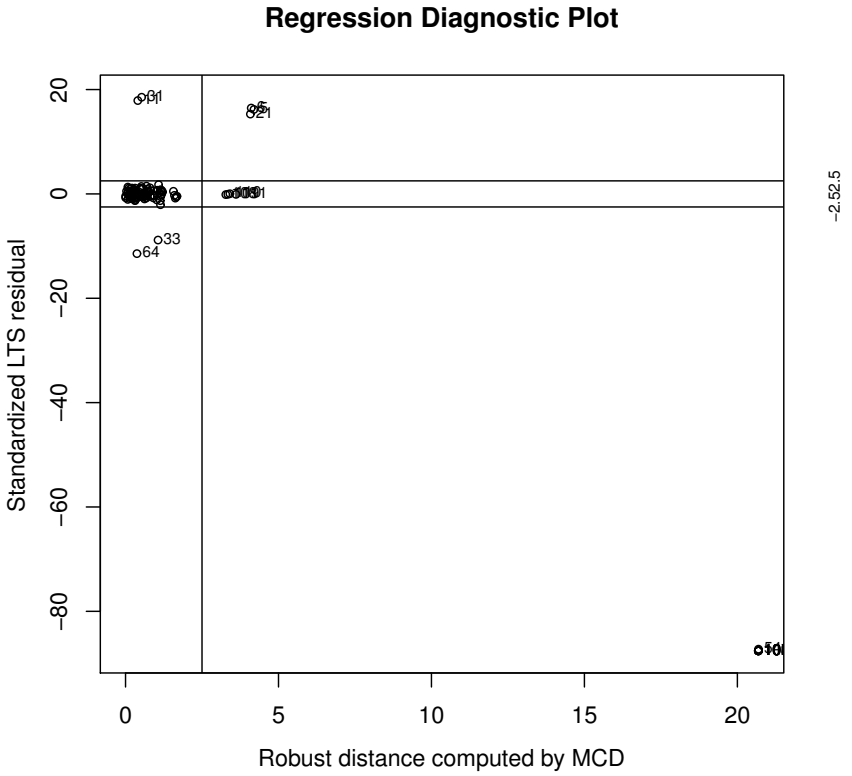
Класификацията на наблюденията е дадена в последната колона *type* на Таблица 3.1. Вижда се, че наблюденията от 1-10 са идентифицирани правилно като BLP (небалансирани), наблюденията 11-14 са идентифицирани правилно като GLP(балансирани), а останалите наблюдения са идентифицирани правилно като регулярни наблюдения (OK).

Тази класификация е дадена на плота на фиг. 3.3. Наблюденията като 15-75, които попадат в правоъгълника, образуван от правите Oy, $y = -2.5$, $y = -2.5$ и $x = 3.06$ регулярни. Наблюденията, които са в областта, наляво от $x = 3.06$ и извън полосата на правите $y = -2.5$ и над $y = 2.5$ са вертикалните несъгласувани наблюдения. Наблюденията като 11-14, които са в областта, надясно от правата $x = 3.06$ и полосата на правите $y = -2.5$ и $y = 2.5$ са балансираните наблюдения. Наблюденията като 1-10, които попадат в областта надясно от правата $x = 3.06$ и са извън полосата на правите $y = -2.5$ и $y = 2.5$ са небалансираните наблюдения.

Чрез тази класификация са анализирани генерираните данни от глава 1, въпреки че те представляват едномерен прост линеен регресионен модел. Идентификацията на наблюденията е дадена в последната колона *type* на Таблица 3.2. Вижда се, че наблюденията са идентифицирани правилно. Резултатите са представени и графично на фиг. 3.4.



Фигура 3.3: Диагностичен плот: робастни разстояния RD_i (Ox) срещу стандартизирани LTS регресионни остатъци $\frac{r_i}{\hat{\sigma}^*_{LTS}}$ (Oy); вертикалната права е в точката $3.06 = \sqrt{\chi^2(23, 0.975)}$.



Фигура 3.4: Диагностичен плот на генерираните по модел (2.16) данни: робастни разстояния RD_i (Ox) срещу стандартизирани LTS регресионни остатъци $\frac{r_i}{\hat{\sigma}^*_{LTS}}$ (Oy); вертикалната права е в точката $3.06 = \sqrt{\chi^2(1, 0.975)}$.

3.3 Приложение 1: Таблици към глави 2 и 3

Таблица 3.1: Разкриване на несъгласувани наблюдения с LTS и MCD в регресионни модели - данни на Hawkins Bradu and Kass.

case	y	x1	x2	x3	RLS.fit	RLS.resid	StdResid	MD	MD.w	RD	LTS.w	RD.w	type
1	9.7	10.1	19.6	28.3	-0.038	9.738	13.088	2.454	1	24.367	0	0	BLP
2	10.1	9.5	20.5	28.9	-0.082	10.182	13.685	2.594	1	24.998	0	0	BLP
3	10.3	10.7	20.2	31	-0.105	10.405	13.984	2.684	1	26.393	0	0	BLP
4	9.5	9.9	21.5	31.7	-0.154	9.654	12.976	2.514	1	27.197	0	0	BLP
5	10.0	10.3	21.1	31.1	-0.107	10.107	13.584	2.517	1	26.713	0	0	BLP
6	10.0	10.8	20.4	29.2	0.003	9.996	13.435	2.652	1	25.316	0	0	BLP
7	10.8	10.5	20.9	29.1	0.004	10.795	14.509	2.869	1	25.391	0	0	BLP
8	10.3	9.9	19.6	28.8	-0.080	10.380	13.951	2.591	1	24.662	0	0	BLP
9	9.6	9.7	20.7	31	-0.166	9.766	13.126	2.538	1	26.445	0	0	BLP
10	9.9	9.3	19.7	30.3	-0.203	10.103	13.578	2.710	1	25.608	0	0	BLP
11	-0.2	11	24	35	-0.135	-0.064	-0.086	4.313	0	30.322	1	0	GLP
12	-0.4	12	23	37	-0.197	-0.202	-0.271	5.266	0	31.412	1	0	GLP
13	0.7	12	26	34	0.076	0.623	0.837	3.846	0	30.553	1	0	GLP
14	0.1	11	34	34	0.314	-0.214	-0.288	6.610	0	34.007	1	0	GLP
15	-0.4	3.4	2.9	2.1	0.103	-0.503	-0.676	1.840	1	1.656	1	1	OK
16	0.6	3.1	2.2	0.3	0.144	0.455	0.612	2.187	1	1.792	1	1	OK
17	-0.2	0	1.6	0.2	-0.126	-0.073	-0.098	1.415	1	1.604	1	1	OK
18	0.0	2.3	1.6	2	-0.032	0.032	0.044	0.866	1	0.650	1	1	OK
19	0.1	0.8	2.9	1.6	-0.082	0.182	0.245	1.186	1	1.065	1	1	OK
20	0.4	3.1	3.4	2.2	0.093	0.306	0.411	1.598	1	1.710	1	1	OK
21	0.9	2.6	2.2	1.9	0.020	0.879	1.181	1.131	1	0.876	1	1	OK
22	0.3	0.4	3.2	1.9	-0.118	0.418	0.562	1.605	1	1.444	1	1	OK
23	-0.8	2	2.3	0.8	0.032	-0.832	-1.119	1.102	1	0.962	1	1	OK
24	0.7	1.3	2.3	0.5	-0.008	0.708	0.952	1.149	1	1.089	1	1	OK
25	-0.3	1	0	0.4	-0.119	-0.180	-0.242	0.811	1	1.643	1	1	OK
26	-0.8	0.9	3.3	2.5	-0.104	-0.695	-0.934	1.188	1	1.411	1	1	OK
27	-0.7	3.3	2.5	2.9	0.038	-0.738	-0.991	1.578	1	1.650	1	1	OK
28	0.3	1.8	0.8	2	-0.105	0.405	0.544	0.874	1	0.857	1	1	OK
29	0.3	1.2	0.9	0.8	-0.088	0.388	0.521	0.603	1	0.940	1	1	OK
30	-0.3	1.2	0.7	3.4	-0.230	-0.069	-0.093	1.671	1	1.747	1	1	OK
31	0.0	3.1	1.4	1	0.076	-0.076	-0.102	1.842	1	1.419	1	1	OK
32	-0.4	0.5	2.4	0.3	-0.059	-0.340	-0.457	1.331	1	1.458	1	1	OK
33	-0.6	1.5	3.1	1.5	-0.012	-0.587	-0.790	0.983	1	1.056	1	1	OK
34	-0.7	0.4	0	0.7	-0.184	-0.515	-0.693	1.214	1	1.689	1	1	OK
35	0.3	3.1	2.4	3	0.012	0.287	0.386	1.256	1	1.559	1	1	OK
36	-1.0	1.1	2.2	2.7	-0.142	-0.857	-1.152	1.003	1	0.947	1	1	OK
37	-0.6	0.1	3	2.6	-0.186	-0.413	-0.555	1.835	1	1.666	1	1	OK
38	0.9	1.5	1.2	0.2	-0.020	0.920	1.237	0.942	1	1.231	1	1	OK
39	-0.7	2.1	0	1.2	-0.071	-0.628	-0.844	1.390	1	1.485	1	1	OK
40	-0.5	0.5	2	1.2	-0.121	-0.378	-0.508	1.112	1	0.897	1	1	OK
41	-0.1	3.4	1.6	2.9	0.010	-0.110	-0.148	1.771	1	1.692	1	1	OK
42	-0.7	0.3	1	2.7	-0.255	-0.444	-0.597	1.832	1	1.493	1	1	OK
43	0.6	0.1	3.3	0.9	-0.087	0.687	0.923	2.026	1	1.786	1	1	OK

Продължава на следващата страница

Таблица 3.1 – продължение от предишната страница

case	y	x1	x2	x3	RLS.fit	RLS.resid	StdResid	MD	MD.w	RD	LTS.w	RD.w	type
44	-0.7	1.8	0.5	3.2	-0.179	-0.520	-0.699	1.639	1	1.673	1	1	OK
45	-0.5	1.9	0.1	0.6	-0.052	-0.447	-0.600	1.130	1	1.534	1	1	OK
46	-0.4	1.8	0.5	3	-0.169	-0.230	-0.310	1.491	1	1.547	1	1	OK
47	-0.9	3	0.1	0.8	0.026	-0.926	-1.245	2.081	1	1.888	1	1	OK
48	0.1	3.1	1.6	3	-0.019	0.119	0.160	1.477	1	1.558	1	1	OK
49	0.9	3.1	2.5	1.9	0.073	0.826	1.110	1.597	1	1.314	1	1	OK
50	-0.4	2.1	2.8	2.9	-0.047	-0.352	-0.473	0.526	1	1.219	1	1	OK
51	0.7	2.3	1.5	0.4	0.045	0.654	0.879	1.363	1	1.253	1	1	OK
52	-0.5	3.3	0.6	1.2	0.050	-0.550	-0.739	2.141	1	1.732	1	1	OK
53	0.7	0.3	0.4	3.3	-0.310	1.010	1.358	2.211	1	2.083	1	1	OK
54	0.7	1.1	3	0.3	0.013	0.686	0.922	1.613	1	1.576	1	1	OK
55	0.0	0.5	2.4	0.9	-0.090	0.090	0.121	1.273	1	1.137	1	1	OK
56	0.1	1.8	3.2	0.9	0.047	0.052	0.070	1.377	1	1.361	1	1	OK
57	0.7	1.8	0.7	0.7	-0.042	0.742	0.997	0.879	1	1.125	1	1	OK
58	-0.1	2.4	3.4	1.5	0.073	-0.173	-0.232	1.409	1	1.441	1	1	OK
59	-0.3	1.6	2.1	3	-0.121	-0.178	-0.239	0.680	1	1.058	1	1	OK
60	-0.9	0.3	1.5	3.3	-0.266	-0.633	-0.851	1.986	1	1.735	1	1	OK
61	-0.3	0.4	3.4	3	-0.167	-0.132	-0.178	1.674	1	1.856	1	1	OK
62	0.6	0.9	0.1	0.3	-0.118	0.718	0.965	0.821	1	1.644	1	1	OK
63	-0.3	1.1	2.7	0.2	0.006	-0.306	-0.411	1.325	1	1.491	1	1	OK
64	-0.5	2.8	3	2.9	0.017	-0.517	-0.695	1.054	1	1.535	1	1	OK
65	0.6	2	0.7	2.7	-0.129	0.729	0.980	1.155	1	1.280	1	1	OK
66	-0.9	0.2	1.8	0.8	-0.133	-0.766	-1.029	1.302	1	1.209	1	1	OK
67	-0.7	1.6	2	1.2	-0.032	-0.667	-0.897	0.667	1	0.455	1	1	OK
68	0.6	0.1	0	1.1	-0.229	0.829	1.114	1.568	1	1.743	1	1	OK
69	0.2	2	0.6	0.3	-0.009	0.209	0.281	1.074	1	1.444	1	1	OK
70	0.7	1	2.2	2.9	-0.161	0.861	1.157	1.020	1	1.124	1	1	OK
71	0.2	2.2	2.5	2.3	-0.020	0.220	0.296	0.642	1	0.830	1	1	OK
72	-0.2	0.6	2	1.5	-0.129	-0.070	-0.094	1.055	1	0.763	1	1	OK
73	0.4	0.3	1.7	2.2	-0.201	0.601	0.808	1.485	1	1.095	1	1	OK
74	-0.9	0	2.2	1.6	-0.175	-0.724	-0.973	1.655	1	1.254	1	1	OK
75	0.2	0.3	0.4	2.6	-0.274	0.474	0.637	1.905	1	1.707	1	1	OK

Y -Dependent variable
RLS.fit -Reweighed Least Squares fit based on FAST-LTS.w weights
RLS.resid -Reweighed Least Squares residuals
StdResid -Standardized Residuals of diagnostic fit
LTS.w -Diagnostic Weights based on diagnostic fit
MD -Mahalanobis Distance
RD -Robustified Mahalanobis Distance
RD.w -Leverage point Diagnostic Weights based on RD
type -Leverage type

Таблица 3.2: Разкриване на несъгласувани наблюдения с LTS и MCD в регресионни модели.

case	y	x	RLS.fit	RLS.resid	StdResid	LTS.w	RD.w	type
1	8.015	0.935	8.459	-4.447e-01	-0.369	1	1	OK
2	4.032	0.059	3.964	6.795e-02	0.056	1	1	OK
3	6.983	0.313	5.648	1.330e+00	1.109	1	1	OK
4	3.500	0.113	3.723	-2.231e-01	-0.185	1	1	OK
5	4.753	4.197	-14.715	1.949e+01	16.184	0	0	BLP
6	5.479	4.108	-14.315	1.975e+01	16.455	0	0	BLP
7	-2.232	1.565	-2.831	5.994e-01	0.498	1	1	OK
8	6.364	0.678	7.298	-9.343e-01	-0.777	1	1	OK
9	3.448	0.123	3.678	-2.300e-01	-0.191	1	1	OK
10	7.173	0.719	7.482	-3.096e-01	-0.257	1	1	OK
11	27.567	0.401	6.044	2.152e+01	17.891	0	1	VO
12	1.562	0.535	1.817	-2.552e-01	-0.211	1	1	OK
13	0.707	0.669	1.209	-5.029e-01	-0.417	1	1	OK
14	0.274	0.620	1.432	-1.151e+00	-0.962	1	1	OK
15	1.675	0.394	2.451	-7.761e-01	-0.645	1	1	OK
16	1.257	0.825	0.508	7.487e-01	0.621	1	1	OK
17	-4.083	1.627	-3.111	-9.710e-01	-0.807	1	1	OK
18	-0.415	1.166	-1.033	6.180e-01	0.514	1	1	OK
19	6.392	0.265	5.434	9.576e-01	0.795	1	1	OK
20	6.695	0.427	6.162	5.325e-01	0.442	1	1	OK
21	4.212	4.080	-14.189	1.842e+01	15.297	0	0	BLP
22	10.160	1.208	9.689	4.704e-01	0.391	1	1	OK
23	1.214	0.648	1.305	-9.182e-02	-0.076	1	1	OK
24	11.023	1.657	11.719	-6.965e-01	-0.578	1	1	OK
25	8.019	0.898	8.291	-2.727e-01	-0.226	1	1	OK
26	8.475	0.697	7.383	1.099e+00	0.907	1	1	OK
27	2.766	0.199	3.332	-5.669e-01	-0.471	1	1	OK
28	3.365	0.200	3.329	3.540e-02	0.029	1	1	OK
29	0.040	0.811	0.572	-5.328e-01	-0.442	1	1	OK
30	0.940	0.631	1.380	-4.409e-01	-0.366	1	1	OK
31	28.902	0.526	6.613	2.229e+01	18.528	0	1	VO
32	9.591	1.119	9.289	3.018e-01	0.250	1	1	OK
33	-21.200	1.062	-0.561	-1.068e+01	-8.843	0	1	VO
34	2.904	0.145	3.576	-6.726e-01	-0.558	1	1	OK
35	9.074	0.909	8.339	7.342e-01	0.610	1	1	OK
36	2.034	0.798	0.629	1.408e+00	1.167	1	1	OK
37	0.022	1.146	-0.943	9.659e-01	0.802	1	1	OK
38	2.589	0.392	2.460	1.287e-01	0.106	1	1	OK
39	3.315	0.496	1.994	1.326e+00	1.097	1	1	OK
40	4.902	0.149	4.911	-9.306e-03	-0.007	1	1	OK
41	4.621	0.212	3.275	1.345e+00	1.118	1	1	OK
42	4.087	0.307	5.623	-1.530e+00	-1.277	1	1	OK
43	7.019	0.587	6.886	1.324e-01	0.109	1	1	OK
44	7.426	0.908	8.336	-9.108e-01	-0.756	1	1	OK
45	2.859	0.534	1.823	1.030e+00	0.861	1	1	OK

Продължава на следващата страница

Таблица 3.2 – продължение от предишната страница

case	y	x	RLS.fit	RLS.resid	StdResid	LTS.w	RD.w	type
46	0.854	0.581	1.606	-7.522e-01	-0.625	1	1	OK
47	10.355	1.179	9.557	7.975e-01	0.663	1	1	OK
48	1.486	0.483	2.051	-5.657e-01	-0.470	1	1	OK
49	4.692	0.174	5.022	-3.307e-01	-0.274	1	1	OK
50	4.924	0.351	5.822	-8.985e-01	-0.746	1	1	OK
51	5.510	0.309	5.633	-1.231e-01	-0.102	1	1	OK
52	-3.888	1.680	-3.352	-5.354e-01	-0.445	1	1	OK
53	9.764	1.037	8.919	8.445e-01	0.701	1	1	OK
54	-7.312	20.689	97.646	-1.046e+02	-87.252	0	0	BLP
55	7.034	0.650	7.172	-1.383e-01	-0.115	1	1	OK
56	4.721	0.032	4.087	6.333e-01	0.526	1	1	OK
57	11.232	1.079	9.109	2.128e+00	1.764	1	1	OK
58	4.064	0.272	3.004	1.052e+00	0.880	1	1	OK
59	9.242	1.165	9.494	-2.520e-01	-0.209	1	1	OK
60	4.165	0.312	5.645	-1.480e+00	-1.230	1	1	OK
61	8.274	0.516	6.565	1.708e+00	1.420	1	1	OK
62	2.222	0.208	3.293	-1.075e+00	-0.890	1	1	OK
63	5.202	0.340	5.771	-5.693e-01	-0.473	1	1	OK
64	-21.214	0.372	2.553	-1.378e+01	-11.445	0	1	VO
65	8.804	1.146	9.409	-6.051e-01	-0.503	1	1	OK
66	6.905	1.142	9.391	-2.489e+00	-2.067	1	1	OK
67	-1.754	1.020	-0.375	-1.379e+00	-1.146	1	1	OK
68	-0.519	1.179	-1.090	5.713e-01	0.475	1	1	OK
69	10.078	1.195	9.632	4.457e-01	0.370	1	1	OK
70	5.472	0.075	3.891	1.581e+00	1.313	1	1	OK
71	5.201	0.088	4.631	5.695e-01	0.473	1	1	OK
72	2.254	0.228	3.200	-9.468e-01	-0.786	1	1	OK
73	5.758	0.275	5.476	2.816e-01	0.233	1	1	OK
74	5.892	0.471	6.363	-4.713e-01	-0.392	1	1	OK
75	3.738	0.003	4.249	-5.115e-01	-0.425	1	1	OK
76	5.217	0.385	5.972	-7.551e-01	-0.628	1	1	OK
77	2.802	0.352	2.640	1.614e-01	0.133	1	1	OK
78	8.045	0.923	8.402	-3.573e-01	-0.297	1	1	OK
79	3.301	0.074	4.568	-1.265e+00	-1.053	1	1	OK
80	-2.446	1.134	-0.886	-1.556e+00	-1.296	1	1	OK
81	1.570	0.643	1.329	2.403e-01	0.199	1	1	OK
82	5.812	0.205	5.163	6.480e-01	0.538	1	1	OK
83	4.751	0.166	3.482	1.261e+00	1.054	1	1	OK
84	8.056	0.744	7.593	4.627e-01	0.384	1	1	OK
85	3.340	0.003	4.216	-8.767e-01	-0.728	1	1	OK
86	5.740	0.198	5.130	6.098e-01	0.506	1	1	OK
87	3.587	0.181	3.413	1.732e-01	0.144	1	1	OK
88	7.785	0.524	6.601	1.187e+00	0.984	1	1	OK
89	2.548	0.135	3.621	-1.073e+00	-0.892	1	1	OK
90	4.469	0.325	5.705	-1.232e+00	-1.027	1	1	OK
91	3.300	0.167	3.476	-1.763e-01	-0.147	1	1	OK
92	5.454	0.146	4.896	5.573e-01	0.463	1	1	OK

Продължава на следващата страница

Таблица 3.2 – продължение от предишната страница

case	y	x	RLS.fit	RLS.resid	StdResid	LTS.w	RD.w	type
93	-3.315	1.598	-2.982	-3.321e-01	-0.276	1	1	OK
94	2.466	0.226	3.212	-7.460e-01	-0.620	1	1	OK
95	6.060	0.106	4.715	1.342e+00	1.117	1	1	OK
96	2.415	0.390	2.472	-5.751e-02	-0.047	1	1	OK
97	2.943	0.683	1.149	1.796e+00	1.490	1	1	OK
98	3.092	0.236	3.167	-7.518e-02	-0.062	1	1	OK
99	1.445	0.751	0.842	6.028e-01	0.500	1	1	OK
100	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
101	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
102	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
103	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
104	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
105	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
106	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
107	-7.709	20.685	97.628	-1.054e+02	-87.567	0	0	BLP
108	18.874	3.277	19.032	-1.587e-01	-0.131	1	0	GLP
109	19.247	3.344	19.333	-8.614e-02	-0.071	1	0	GLP
110	19.672	3.410	19.633	3.847e-02	0.031	1	0	GLP
111	20.427	3.610	20.536	-1.086e-01	-0.090	1	0	GLP

Y -Dependent variable
x -predictor variable
FAST-LTS.fit -Diagnostic fit based on "FAST-LTS"method
FAST-LTS.resid -Residuals of diagnostic fit
StdResid -Standardized Residuals of diagnostic fit
FAST-LTS.w -Diagnostic Weights based on diagnostic fit
RD.w -Leverage point Diagnostic Weights based on RD
type -Leverage type
RLS.fit -Reweighed Least Squares fit based on FAST-LTS.w weights
RLS.resid -Reweighed Least Squares residuals

Библиография

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental set. *Technometrics, 26*, pp. 197–208.

Neykov,N.M. and Neytchev,Pl.N.(1991). Unmasking of multivariate outliers and leverage points by means of BMDP3R. In: *Directions in Robust Statistics and Diagnostics.* Part II, W.Stahel and S. Weisberg (eds), The IMA volums in mathematics and its applications vol 34, Springer-Verlag, pp. 115–128.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.,79*, пп. 851–857.

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

Rousseeuw, P.J. and Van Driessen, K., (1999a). Computing least trimmed of squares regression for large data sets. *Estadistica, 54*, pp. 163–190.

Rousseeuw, P. J. and K. Van Driessen (1999b), A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212–223.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points by means of robust covariance matrices (with discussion). *Amer. Statist. Assoc. 85*, pp. 633–651.

R Development Core Team (2006) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Vandev, D. L. and Neykov, N. M. (1993). Robust Maximum Likelihood in the Gaussian Case, In: *N*ew Directions in Statistical Data Analysis and Robustness. S. Morgenthaler, E. Ronchetti, and W.A. Stahel (eds.). Basel: Birkhauser Verlag. 257-264.