APLICAÇÕES DE MODELOS PARA REGRESSÃO ROBUSTA

Autores: Manuel Ferreira Junior e
Marcos Antonio Bezerra da Silva Junior.
Disciplina: Regressão II
Professor: Eufrásio de Andrade Lima Neto

Sumário

Regressão Robusta

O método dos mínimos quadrados é bastante utilizado para a estimação de parâmetros em Regressão, porém com a presença de **outliers** as estimativas podem não ser confiáveis devido à influência que sofrem pelos valores extremos.

A **Regressão Robusta** torna-se uma alternativa para evitar que o modelo seja fortemente afetado por estes pontos aberrantes, apresentando meios que visam ponderar a importância de cada elemento da amostra.

Outliers

Os outliers são pontos que se diferenciam drasticamente dos demais, através de seus valores extremos, podendo causar anomalias para os resultados.

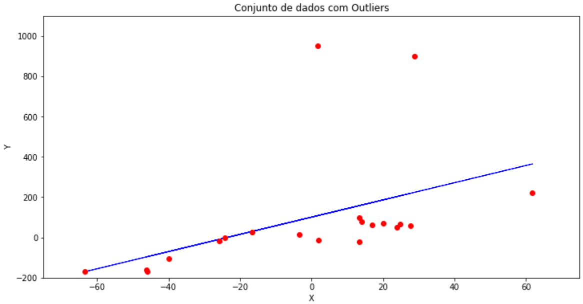


Figura: Medium, **Outlier: o ponto fora da curva.** Salles, Rodrigues. 2018.

Outliers

Na literatura existem meios para detectar estes valores. Em alguns métodos de predição, a exclusão para uma análise menos influenciada por estes valores torna-se necessária. Alguns métodos [1] para detectar outliers nas amostras:

[1] (Soetewey 2020)

Distância de Cookie

(Cook e Weisberg 1982) A distância de Cookie é uma medida baseada na exclusão de uma determinada observação, dentro de uma **análise de diagnóstico em uma regressão que utiliza-se de mínimos quadrados**. Ela é definida da seguinte forma:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot MSE}$$

onde $\hat{Y}_j$ é a previsão completa do modelo de regressão, com a observação j; $\hat{Y}_{j(i)}$ é a previsão da observação j de um modelo sem a observação i; MSE é o erro quadrático médio do modelo e p é o número de parâmetros ajustados.

Distância de Cookie

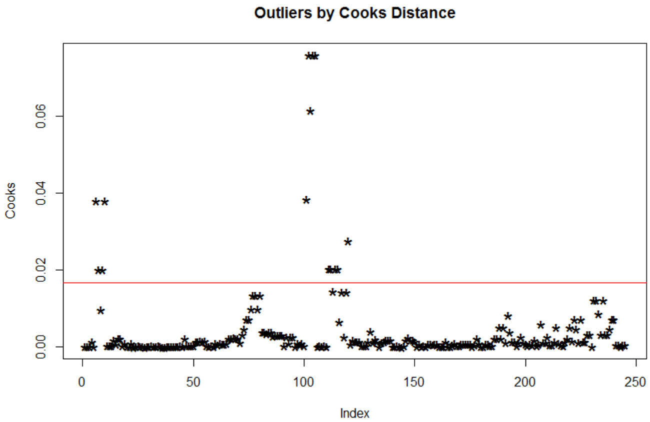


Figura: (Lappalainen e Rykova 2017)

Boxplot

O Boxplot resume um conjunto de dados, utilizando as referências de valores mínimos e máximos, primeiro e terceiro quantil, mediana e **outliers**. Quando a discrepância na variabilidade é muito elevada, o Boxplot detecta estes pontos.

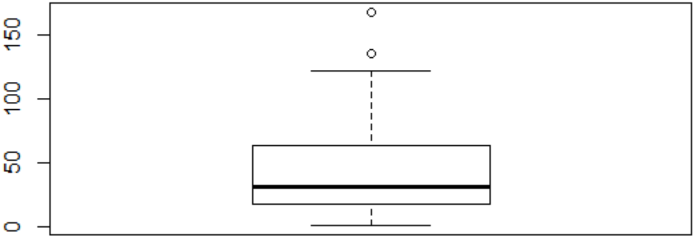


Figura: Boxplot

Percentis

Todas as observações que estiverem fora do intervalo formado pelos percentis 2,5 e 97,5 (ou 1 e 99, ou 5 e 95) serão consideradas outliers potenciais. O intervalo pode ser modificado dependendo do tipo de dados.

Percentis

- LI <- quantile(data$x, 0.01)
- LS <- quantile(data$x, 0.99)
- Se a i-ésima observação for $x_i < LI$ ou $x_i > LS$, x_i é um outlier.

Hampel filter

Considera como outliers os valores fora do intervalo formado pela mediana, mais ou menos 3 desvios absolutos medianos.

$$I = [\tilde{\mu} - 3 \cdot MAD; \tilde{\mu} + 3 \cdot MAD]$$

onde,

$$MAD = mediana(|X_i - \tilde{X}|)$$

MAD é o desvio absoluto mediano, definido como a mediana dos desvios absolutos da mediana dos dados.

Hampel filter

- `LI <- median(data$x) - 3 * mad(data$x, constant = 1)`
- `LS <- median(data$x) + 3 * mad(data$x, constant = 1)`
- Se a *i*-ésima observação for $x_i < LI$ ou $x_i > LS$, x_i é um outlier.

Teste de Grubbs

(Grubbs et al. 1950) Detecta um outlier de cada vez, podendo ser o valor mais alto ou mais baixo. As hipóteses são definidas como:

H_0 : O valor mais alto **(ou mais baixo)** não é um outlier

H_1 : O valor mais alto **(ou mais baixo)** é um outlier

- **Obs:** O teste de Grubb's não é apropriado para tamanho de amostras menor ou igual a 6.

Teste de Grubbs

- install.packages("outliers")
- library(outliers)
- grubbs.test(data$x)

Teste de Dixon

(Dixon 1950) O teste de Dixon é similar ao de Grubb's, porém é mais poderoso em amostras pequenas (menores ou iguais que 25). Assim como o anterior, detecta um outlier de cada vez, logo se houver suspeita de mais de um outlier, o teste deve ser executado individualmente.

Teste de Dixon

- install.packages("outliers")
- library(outliers)
- dixon.test(data$x)

Teste de Rosner

(Rosner 1975) Detecta vários outliers de uma vez, além de evitar um poder de mascaramento, onde um valor discrepante próximo de outro pode passar despercebido.

$$H_0 : \text{Não há outlier encontrado no conjunto de dados}$$

$$H_1 : \text{Há até r outliers no conjunto de dados}$$

O teste de Rosner é mais adequado para grandes amostras (maiores ou iguais a 20).

Teste de Rosner

- install.packages("EnvStats")
- library(EnvStats)
- rosnerTest(data$x, k = 3)

Modelos

Na literatura existem diversos modelos que consideram a robustez dos dados sem a necessidade da suposição de homocedasticidade, além de não serem afetados por estes pontos aberrantes.

Iremos apresentar a seguir alguns modelos de Regressão Robusta para a estimação dos parâmetros quando os dados se encontram nestas situações.

WLS

É uma extensão do modelo OLS, porém possui uma matriz **W** que descreve a precisão de cada observação do conjunto de dados individualmente, aplicando pesos de acordo com a importância de cada observação para o modelo. Vamos definir da seguinte forma a estimação dos $\hat{\beta}$, então:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} w_i \cdot (y_i - x_i^T \cdot \beta)^2$$

onde,

$$w_i = \frac{1}{\hat{\sigma}_{ii}^2}$$

WLS

Tal que podemos definir $\hat{\sigma}^2_{ii}$ como o i-th elemento da diagonal formada pelo resultado de $\hat{\sigma}^2 \cdot H = \hat{\sigma}^2 X(X^T X)^{-1} X^T$, sendo $\hat{\sigma}^2$ a estimativa da variância referente ao erro aleatório.

L1

Diferente do método dos mínimos quadrados que produz estimativas da média condicional, a Regressão L1 produz estimativas da mediana ou de quaisquer outros quantis de $y|x$. O método de estimação tem como objetivo minimizar os erros absolutos. Vamos definir a forma de estimar os $\hat{\beta}$ da seguinte forma:

$$\hat{\beta} = \text{arg min}_{\beta} \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i^T \beta|$$

Classe de estimadores propostas por (Huber, 1981). Utiliza-se o método de estimação de Máxima Verossimilhança e é considerada uma função-peso que penalize os outliers. Além disso, apresentam Normalidade assintótica. O método de estimação tem como objetivo minimizar os erros padronizados. Vamos definir a forma de estimação para $\hat{\beta}$ da seguinte forma:

$$\hat{\beta} = \text{arg min}_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \beta}{\hat{\sigma}}\right)$$

tal que, definimos $\rho(.)$ como uma função robusta; $\hat{\sigma}$ um estimador de escala referente ao erro.

MM

Semelhante ao estimador M, porém modificando o processo de estimação. Proposto por (Yohai 1987), realiza uma estimação em três estágios:

1. Inicialmente apresentar um estimador robusto $\hat{\beta}$, com elevado breakdown point mas de baixa eficiência;
2. Obtem um estimador robusto M para a escala ($\hat{\sigma}$);
3. Obter um estimador M baseado nos estimadores das etapas *i* e *ii*.

ETKRR

A ideia principal de um modelo de regressão robusta baseada em kernel, é minimizar a seguinte função:

$$S = \sum_{i=1}^{n} ||\phi(y_i) - \phi(\mu_i)||^2 = \sum_{i=1}^{n} [K(y_i, y_i) - 2 \cdot K(y_i, \mu_i) + K(\mu_i, \mu_i)]$$

onde, por propriedade, temos que $K(\mu_i, \mu_i) = K(y_i, y_i) = 1$, logo temos que S é dad o pela seguinte forma:

$$S = \sum_{i=1}^{n} 2 \cdot [1 - K(y_i, \mu_i)]$$

Por fim, tome que $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ é a média da variável Y_i, apresentando uma relação linear com o conjunto de variáveis X_j

Red Wine Quality

O banco utilizado refere-se a duas variantes do vinho Portugues, Vinho Verde, produto da região de Minho, localizada a Noroeste de Portugal. Esse vinho representa cerca de 15% da produção total portuguesa, sendo desse total, cerca de 10% exportado em especial o vinho branco. Os dados foram retirados de um periodo de Maio de 2004 a fevereiro de 2007. Para mais informações, consultar a referência (Cortez et al. 2009).

Red Wine Quality

Para base de dados utilizadas, temos uma composição de **13** variáveis e **1599** observações. As variáveis são:

- *fixed acidity;*
- *volatile acidity;*
- *citric acid;*
- *residual sugar;*
- *chlorides;*
- *free sulfur dioxide;*

Red Wine Quality

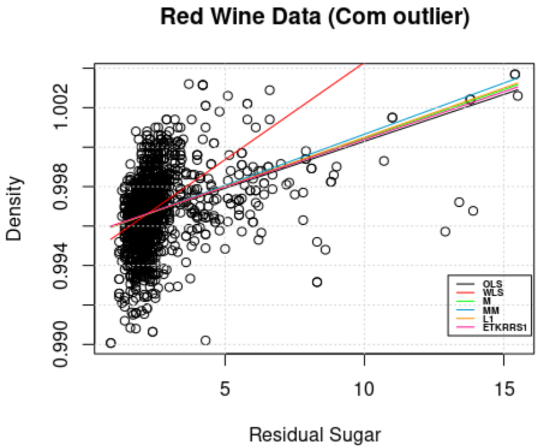
- *total sulfur dioxide;*
- *density;*
- *pH;*
- *sulphates;*
- *quality;*

Em especial, iremos trabalhar com dois pares de variáveis desse problema, sendo elas: (Y) Density vs (X) *Residual sugar* e (Y) *Free sulfur dioxide* vs (X) *Total sulfur dioxide*.

Residual sugar vs Density

Para este primeiro par de variáveis, iremos utilizar a variável *Residual sugar*, ou seja, a quantidade de açúcar que resta após a parada da fermentação, para tentar explicar a variável *Density*, que é a densidade de água.

Residual sugar vs Density



Detectando outlier pelo OLS

Detectando outlier pelo OLS

Tabela: Percentual de mudança (%) nas estimativas dos parâmetros

Método	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	2,79 x 10^{-3}	6,56
WLS	0,028 x 10^{-2}	10,79
M	0,126	0,664
MM	5,73 x 10^{-3}	3,90
L1	8,14 x 10^{-4}	0,172
ETKRR	2,53 x 10^{-3}	2,33

Detectando outlier pelo teste de Rosner

Detectando outlier pelo teste de Rosner

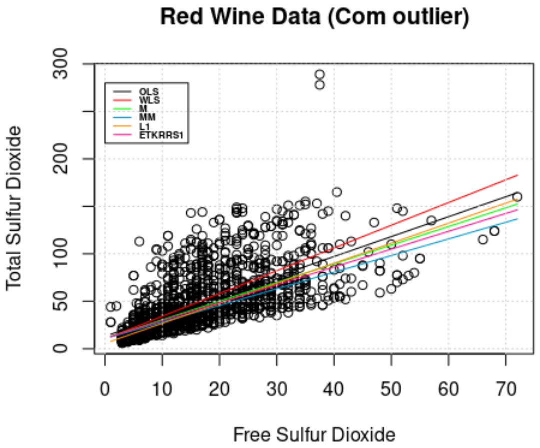
Tabela: Percentual de mudança (%) nas estimativas dos parâmetros

Método	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	0,152	143,28
WLS	0,108	49,54
M	0,141	131,42
MM	0,137	121,66
L1	0,125	118,19
ETKRR1	0,128	126,69

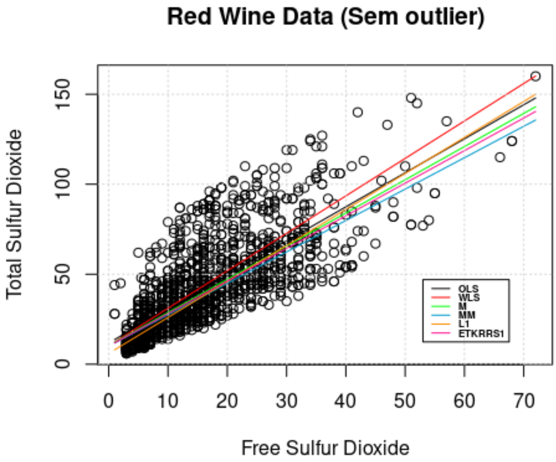
Free sulfur dioxide vs Total sulfur dioxide

Para o segundo par de variáveis, iremos utilizar a variável *Free sulfur dioxide*, sendo a forma livre de SO2 existe em equilíbrio entre o SO2 molecular (como um gás dissolvido) e o íon bissulfito, para tentar explicar a variável *Total sulfur dioxide*, que é a quantidade de formas livres e ligadas de S02.

Free sulfur dioxide vs Total sulfur dioxide



Detectando outlier pelo OLS

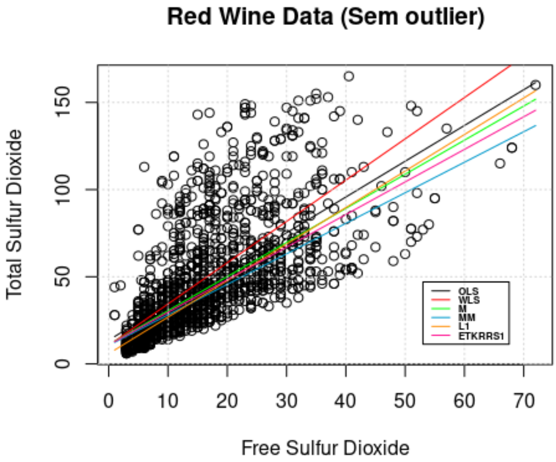


Detectando outlier pelo OLS

Tabela: Percentual de mudança (%) nas estimativas dos parâmetros

Método	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	9,910	9,962
WLS	0,636	13,11
M	5,20	6,23
MM	5,24	0,391
L1	5,88	5,26
ETKRR1	7,71	3,17

Detectando outlier pelo teste de Rosner



Detectando outlier pelo teste de Rosner

Tabela: Percentual de mudança (%) nas estimativas dos parâmetros

Método	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	3,93	2,27
WLS	3,16	1,12
M	0,819	0,498
MM	0,0395	0,0398
L1	3,36	0,752
ETKRR1	0,649	0,525

Por fim ...

Cuidado!!!

Referências

- Cortez, Paulo et al. (2009). “Modeling wine preferences by data mining from physicochemical properties”. Em: *Decision support systems* 47.4, pp. 547–553.
- Dixon, Wilfred J (1950). “Analysis of extreme values”. Em: *The Annals of Mathematical Statistics* 21.4, pp. 488–506.
- Grubbs, Frank E et al. (1950). “Sample criteria for testing outlying observations”. Em: *Annals of mathematical statistics* 21.1, pp. 27–58.
- Rosner, Bernard (1975). “On the detection of many outliers”. Em: *Technometrics* 17.2, pp. 221–227.
- Yohai, Victor J (1987). “High breakdown-point and high efficiency robust estimates for regression”. Em: *The Annals of Statistics*, pp. 642–656.

- Cook, R Dennis e Sanford Weisberg (1982). *Residuals and influence in regression*. New York: Chapman e Hall.

Referências

- Soetewey, Antoine (2020). *Outliers detection in R.* URL: https://statsandr.com/blog/outliers-detection-in-r/ (acesso em 25/06/2021).
- Lappalainen, Aaro e Evgeniia Rykova (jan. de 2017). *The Figurative Language Comprehension of Primary School Children is facilitated by Motor Ability: a multidimensional behavioural study.* DOI: 10.13140/RG.2.2.17723.72485.