Regression diagnostics

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, November 11, 2026

Outline

1. Introduction
2. Leverage points
3. Influential observations
4. Application
5. References

Outline

1. Introduction
2. Leverage points
3. Influential observations
4. Application
5. References

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \tag{1}$$

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \tag{1}$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ is known, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of unknown parameters to be estimated, $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2, also unknown, to be estimated.

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \tag{1}$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ is known, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of unknown parameters to be estimated, $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2, also unknown, to be estimated.

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \tag{2}$$

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ is the design matrix.

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ is the design matrix.

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Outline

1 Introduction

2 Leverage points

3 Influential observations

4 Application

5 References

Leverage points

Let the hat matrix be $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, with its ℓth diagonal element given by

$$h_{\ell\ell} = \boldsymbol{x}_\ell(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_\ell^\top.$$

Leverage points

Let the hat matrix be $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, with its ℓth diagonal element given by

$$h_{\ell\ell} = \boldsymbol{x}_\ell(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_\ell^\top.$$

$h_{\ell\ell}$ is a measure of distance between the ℓth observation and the $\boldsymbol{x}$-space.

Leverage points

Let the hat matrix be $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, with its ℓth diagonal element given by

$$h_{\ell\ell} = \boldsymbol{x}_\ell(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_\ell^\top.$$

$h_{\ell\ell}$ is a measure of distance between the ℓth observation and the $\boldsymbol{x}$-space.

Leverage points

Remarks:

- The average size of the diagonal elements of $\boldsymbol{H}$ is p/n;

Leverage points

Remarks:

- The average size of the diagonal elements of $\boldsymbol{H}$ is p/n;
- An observation is considered a leverage point when $h_{\ell\ell} > 2p/n$;

Leverage points

Remarks:

- The average size of the diagonal elements of $\boldsymbol{H}$ is p/n;
- An observation is considered a leverage point when $h_{\ell\ell} > 2p/n$;
- An observation with a large $h_{\ell\ell}$ and a large residual will likely be an influential point.

Leverage points

Remarks:

- The average size of the diagonal elements of $\boldsymbol{H}$ is p/n;
- An observation is considered a leverage point when $h_{\ell\ell} > 2p/n$;
- An observation with a large $h_{\ell\ell}$ and a large residual will likely be an influential point.

Outline

1 Introduction

2 Leverage points

3 Influential observations

4 Application

5 References

Influential observations

Influence diagnostics measure the effect of deleting the ℓth observation.

Influential observations

- Cook's Distance measures the effect on $\widehat{\beta}$;

Influential observations

- Cook's Distance measures the effect on $\widehat{\boldsymbol{\beta}}$;
- DFBETAS measures the effect on $\hat{\beta}_m$, $m = 1, 2, \ldots, p$;

Influential observations

- Cook's Distance measures the effect on $\widehat{\boldsymbol{\beta}}$;
- DFBETAS measures the effect on $\hat{\beta}_m$, $m = 1, 2, \ldots, p$;
- DFFITS measures the effect on $\hat{Y}_\ell$, $\ell = 1, 2, \ldots, n$;

Influential observations

- Cook's Distance measures the effect on $\widehat{\boldsymbol{\beta}}$;
- DFBETAS measures the effect on $\hat{\beta}_m$, $m = 1, 2, \ldots, p$;
- DFFITS measures the effect on $\hat{Y}_\ell$, $\ell = 1, 2, \ldots, n$;
- COVRATIO measures the effect on the covariance matrix.

Influential observations

- Cook's Distance measures the effect on $\widehat{\boldsymbol{\beta}}$;
- DFBETAS measures the effect on $\hat{\beta}_m$, $m = 1, 2, \ldots, p$;
- DFFITS measures the effect on $\hat{Y}_\ell$, $\ell = 1, 2, \ldots, n$;
- COVRATIO measures the effect on the covariance matrix.

Influential observations

Cook's distance (Cook, 1977, 1979) measures how well the model fits the observation Y_ℓ and how far the observation lies from the rest of the dataset.

Influential observations

Cook's distance (Cook, 1977, 1979) measures how well the model fits the observation Y_ℓ and how far the observation lies from the rest of the dataset.

This diagnostic is calculated as follows:

$$D_\ell = \frac{(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})}{p\,\mathrm{MSE}} = \frac{r_\ell^2}{p}\frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Influential observations

Cook's distance (Cook, 1977, 1979) measures how well the model fits the observation Y_ℓ and how far the observation lies from the rest of the dataset. This diagnostic is calculated as follows:

$$D_\ell = \frac{(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})}{p\,\mathrm{MSE}} = \frac{r_\ell^2}{p}\,\frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

An observation is considered influential if $D_\ell > 1$.

Influential observations

Cook's distance (Cook, 1977, 1979) measures how well the model fits the observation Y_ℓ and how far the observation lies from the rest of the dataset. This diagnostic is calculated as follows:

$$D_\ell = \frac{(\widehat{\beta}_{(\ell)} - \widehat{\beta})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\beta}_{(\ell)} - \widehat{\beta})}{p\,\mathrm{MSE}} = \frac{r_\ell^2}{p}\frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

An observation is considered influential if $D_\ell > 1$. Cook's distance can also be used to assess the joint influence of observations.

Influential observations

Cook's distance (Cook, 1977, 1979) measures how well the model fits the observation Y_ℓ and how far the observation lies from the rest of the dataset. This diagnostic is calculated as follows:

$$D_\ell = \frac{(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}})}{p\,\mathrm{MSE}} = \frac{r_\ell^2}{p}\frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

An observation is considered influential if $D_\ell > 1$. Cook's distance can also be used to assess the joint influence of observations.

Influential observations

DFBETAS (Belsley et al., 1980) measures how much the regression coefficient changes, in standard error units, when the ℓth observation is deleted.

This diagnostic is calculated as follows:

$$\text{DFBETAS}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S^2_{(\ell)} c_{mm}}},$$

Influential observations

DFBETAS (Belsley et al., 1980) measures how much the regression coefficient changes, in standard error units, when the ℓth observation is deleted. This diagnostic is calculated as follows:

$$\text{DFBETAS}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S^2_{(\ell)} c_{mm}}},$$

where $\hat{\beta}_{m(\ell)}$ is the mth estimated coefficient with the ℓth observation removed.

Influential observations

DFBETAS (Belsley et al., 1980) measures how much the regression coefficient changes, in standard error units, when the ℓth observation is deleted. This diagnostic is calculated as follows:

$$\text{DFBETAS}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S^2_{(\ell)} c_{mm}}},$$

where $\hat{\beta}_{m(\ell)}$ is the mth estimated coefficient with the ℓth observation removed. An observation is considered influential if $|\text{DFBETAS}_{m,\ell}| > 2/\sqrt{n}$.

Influential observations

DFBETAS (Belsley et al., 1980) measures how much the regression coefficient changes, in standard error units, when the ℓth observation is deleted. This diagnostic is calculated as follows:

$$\text{DFBETAS}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S^2_{(\ell)} c_{mm}}},$$

where $\hat{\beta}_{m(\ell)}$ is the mth estimated coefficient with the ℓth observation removed. An observation is considered influential if $|\,\text{DFBETAS}_{m,\ell}\,| > 2/\sqrt{n}$.

Influential observations

DFFITS (*difference in fit(s)*, Belsley et al., 1980) measures the influence of the ℓth observation on its predicted value, in standard error units. This diagnostic is calculated as follows:

$$\text{DFFITS}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S^2_{(\ell)} h_{\ell\ell}}}.$$

Influential observations

DFFITS (*difference in fit(s)*, Belsley et al., 1980) measures the influence of the ℓth observation on its predicted value, in standard error units. This diagnostic is calculated as follows:

$$\text{DFFITS}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S^2_{(\ell)} h_{\ell\ell}}}.$$

An observation is considered influential if $|\,\text{DFFITS}_\ell\,| > 2\sqrt{p/n}$.

Influential observations

DFFITS (*difference in fit(s)*, Belsley et al., 1980) measures the influence of the ℓth observation on its predicted value, in standard error units. This diagnostic is calculated as follows:

$$\text{DFFITS}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S^2_{(\ell)} h_{\ell\ell}}}.$$

An observation is considered influential if $|\,\text{DFFITS}_\ell\,| > 2\sqrt{p/n}$. DFBETAS and DFFITS are closely related diagnostics; see Montgomery et al. (2021).

Influential observations

DFFITS (*difference in fit(s)*, Belsley et al., 1980) measures the influence of the ℓth observation on its predicted value, in standard error units. This diagnostic is calculated as follows:

$$\text{DFFITS}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S^2_{(\ell)} h_{\ell\ell}}}.$$

An observation is considered influential if $|\,\text{DFFITS}_\ell\,| > 2\sqrt{p/n}$. DFBETAS and DFFITS are closely related diagnostics; see Montgomery et al. (2021).

Influential observations

COVRATIO provides information about the overall precision of the parameter estimates. This diagnostic is calculated as follows:

$$\text{COVRATIO}_\ell = \frac{|(\boldsymbol{X}_{(\ell)}^\top \boldsymbol{X}_{(\ell)})^{-1} S_{(\ell)}^2|}{|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\,\text{MSE}\,|} = \left[\frac{S_{(\ell)}^2}{\text{MSE}}\right]^p \frac{1}{1 - h_{\ell\ell}}.$$

Influential observations

COVRATIO provides information about the overall precision of the parameter estimates. This diagnostic is calculated as follows:

$$\text{COVRATIO}_\ell = \frac{|(\boldsymbol{X}_{(\ell)}^\top \boldsymbol{X}_{(\ell)})^{-1} S_{(\ell)}^2|}{|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\,\text{MSE}\,|} = \left[\frac{S_{(\ell)}^2}{\text{MSE}}\right]^p \frac{1}{1-h_{\ell\ell}}.$$

If $\text{COVRATIO}_\ell > 1$, deleting the ℓth observation degrades precision, whereas if $\text{COVRATIO}_\ell < 1$, deleting it improves precision.

Influential observations

COVRATIO provides information about the overall precision of the parameter estimates. This diagnostic is calculated as follows:

$$\text{COVRATIO}_\ell = \frac{|(\boldsymbol{X}_{(\ell)}^\top \boldsymbol{X}_{(\ell)})^{-1} S_{(\ell)}^2|}{|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\,\text{MSE}\,|} = \left[\frac{S_{(\ell)}^2}{\text{MSE}}\right]^p \frac{1}{1-h_{\ell\ell}}.$$

If $\text{COVRATIO}_\ell > 1$, deleting the ℓth observation degrades precision, whereas if $\text{COVRATIO}_\ell < 1$, deleting it improves precision. Cutoff points of $1+3p/n$ and $1-3p/n$ are commonly used.

Influential observations

COVRATIO provides information about the overall precision of the parameter estimates. This diagnostic is calculated as follows:

$$\text{COVRATIO}_\ell = \frac{|(\boldsymbol{X}_{(\ell)}^\top \boldsymbol{X}_{(\ell)})^{-1} S_{(\ell)}^2|}{|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\,\text{MSE}\,|} = \left[\frac{S_{(\ell)}^2}{\text{MSE}}\right]^p \frac{1}{1-h_{\ell\ell}}.$$

If $\text{COVRATIO}_\ell > 1$, deleting the ℓth observation degrades precision, whereas if $\text{COVRATIO}_\ell < 1$, deleting it improves precision. Cutoff points of $1+3p/n$ and $1-3p/n$ are commonly used.

Influential observations

Once influential observations are identified, they may be removed if they result from measurement errors or do not belong to the target population.

Otherwise, they should be retained, and the use of robust regression techniques becomes necessary (e.g., methodologies that downweight extreme observations).

Influential observations

Once influential observations are identified, they may be removed if they result from measurement errors or do not belong to the target population.

Otherwise, they should be retained, and the use of robust regression techniques becomes necessary (e.g., methodologies that downweight extreme observations).

Outline

1 Introduction

2 Leverage points

3 Influential observations

4 Application

5 References

Application

(Montgomery et al., 2021, p. 180) An electric utility is interested in developing a model relating peak-hour demand (Y, in kW) to total energy usage during the month of August (x_2, in kWh), with the scatter plot shown in Figure 1.

Application

(Montgomery et al., 2021, p. 180) An electric utility is interested in developing a model relating peak-hour demand (Y, in kW) to total energy usage during the month of August (x_2, in kWh), with the scatter plot shown in Figure 1. After fitting the model to the data, we obtain the following estimated model,

$$\hat{Y}_\ell = -0.831 + 0.004 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53.$

Application

(Montgomery et al., 2021, p. 180) An electric utility is interested in developing a model relating peak-hour demand (Y, in kW) to total energy usage during the month of August (x_2, in kWh), with the scatter plot shown in Figure 1. After fitting the model to the data, we obtain the following estimated model,

$$\hat{Y}_\ell = -0.831 + 0.004 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53.$

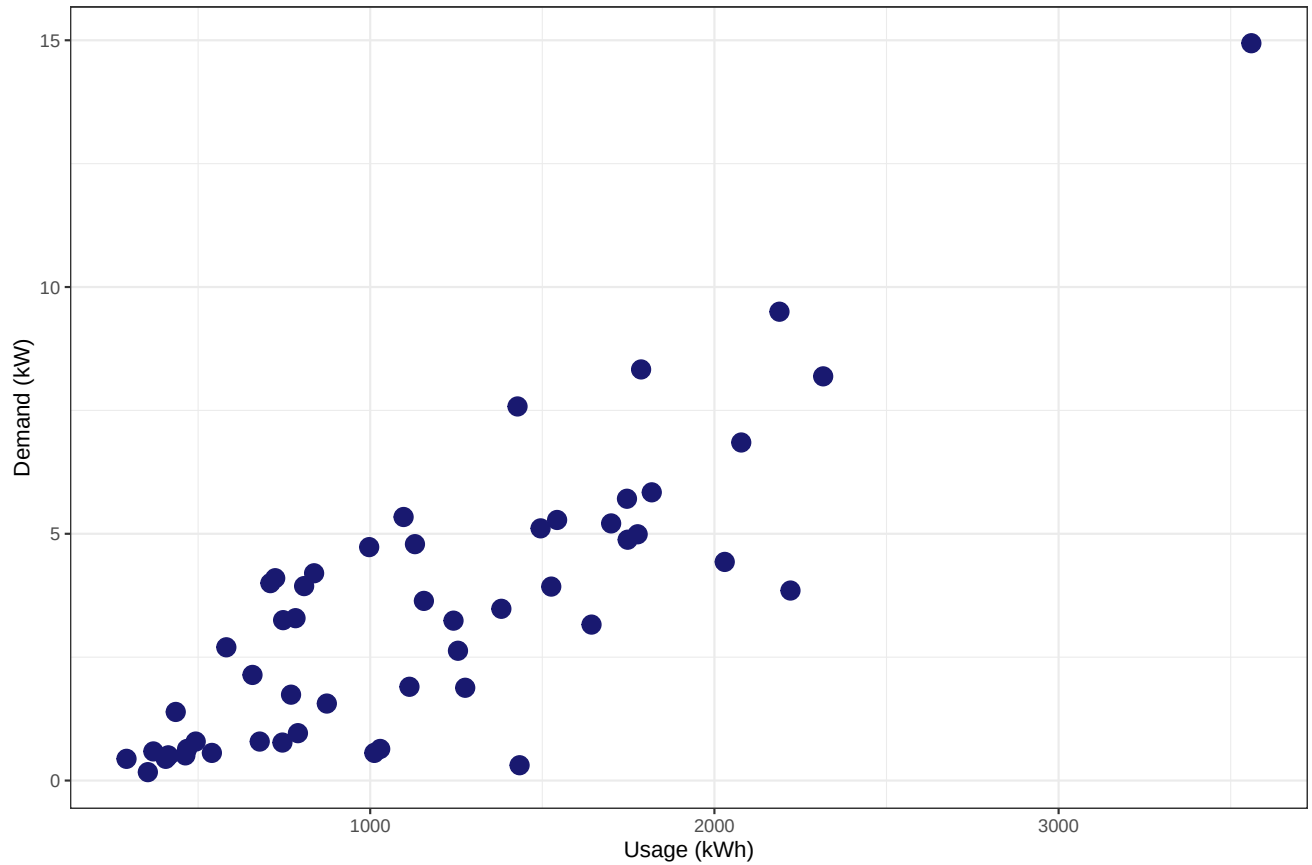


Figure 1: Scatter plot.

Example

We also have:

Table 1: Parameter estimates.

Parameter	Estimate	SE	t_0
β_1	-0.831	0.442	-1.882
β_2	0.004	0.001	11.030

Critical region, for $\alpha = 5\%$: $|t_0| > 2.008$, with MSE $= 2.488$.

Example

We also have:

Table 1: Parameter estimates.

Parameter	Estimate	SE	t_0
β_1	-0.831	0.442	-1.882
β_2	0.004	0.001	11.030

Critical region, for $\alpha = 5\%$: $|t_0| > 2.008$, with MSE $= 2.488$.

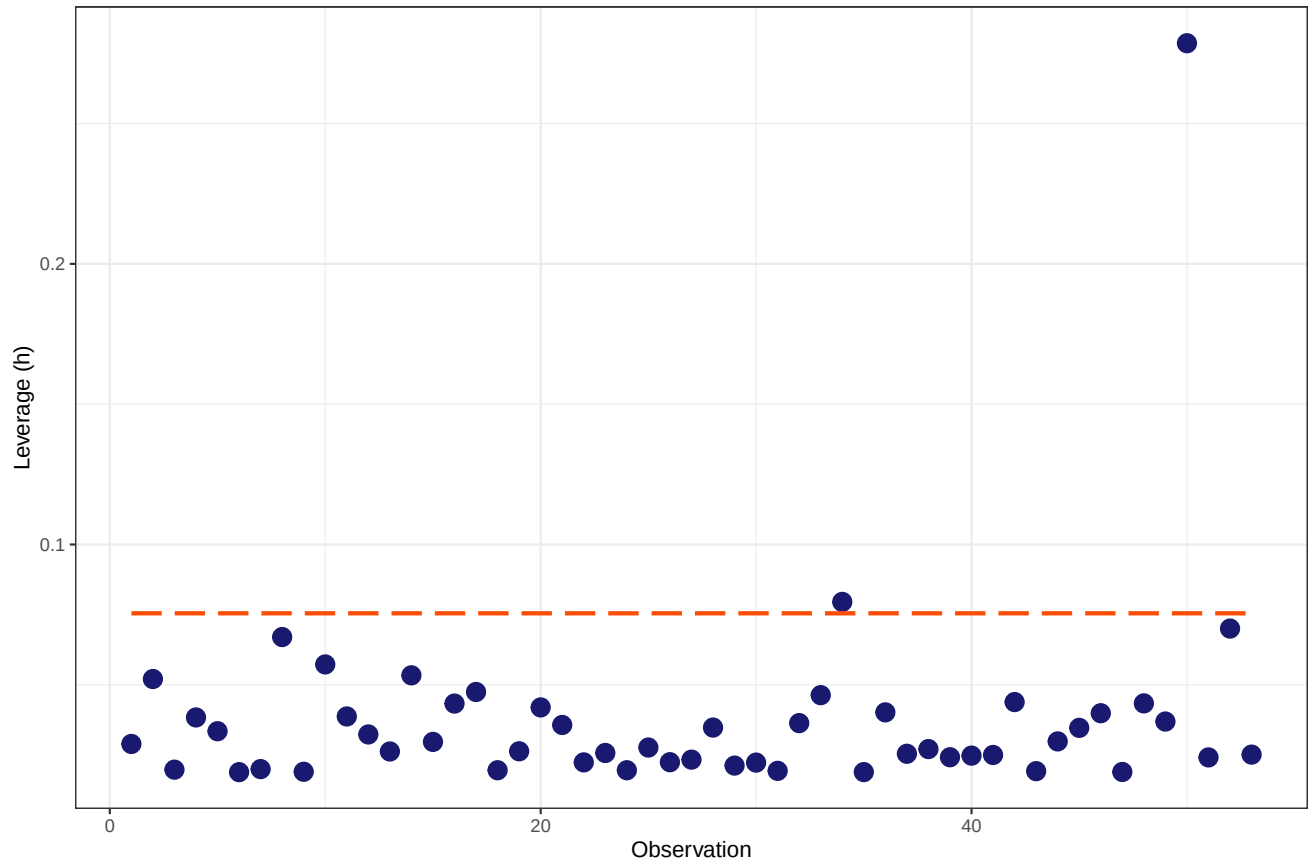


Figure 2: Leverage plot.

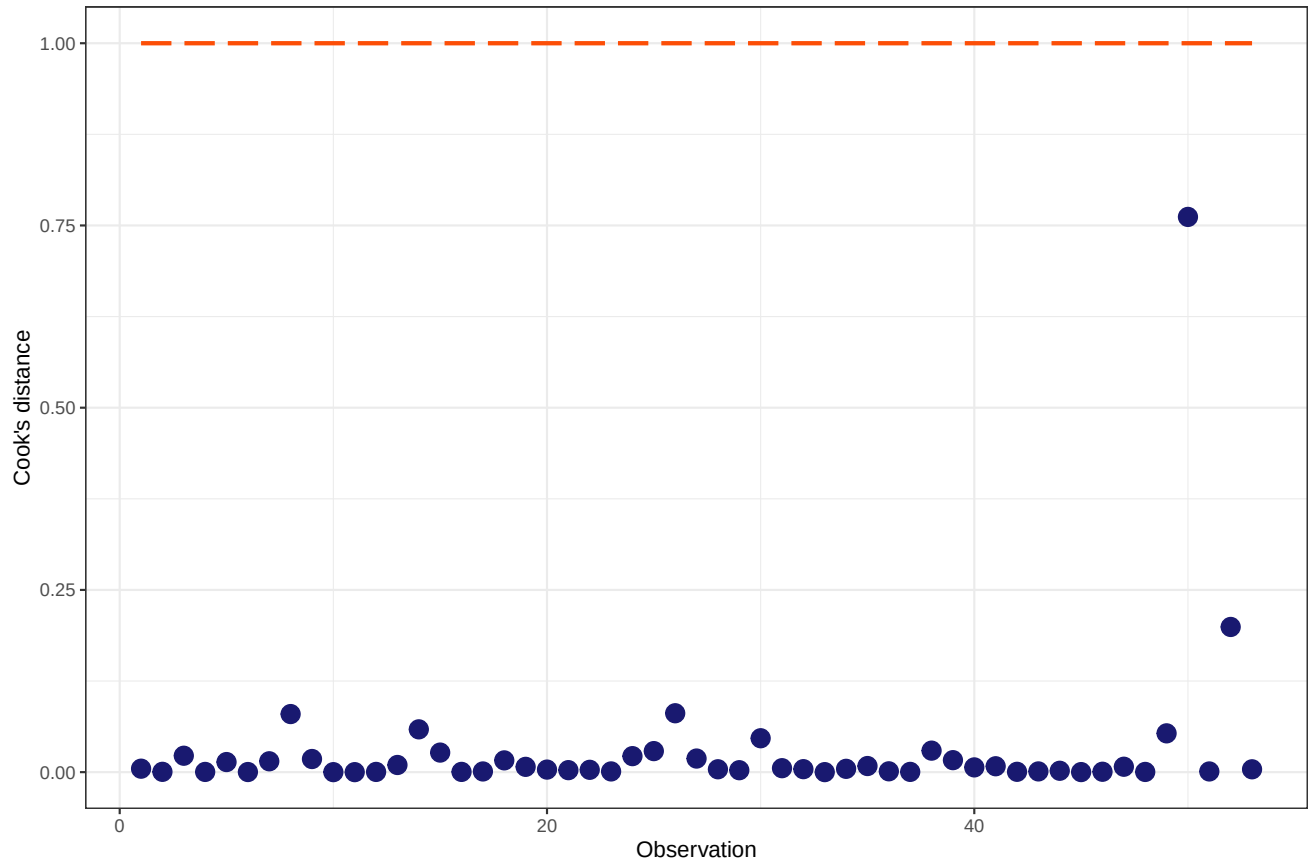
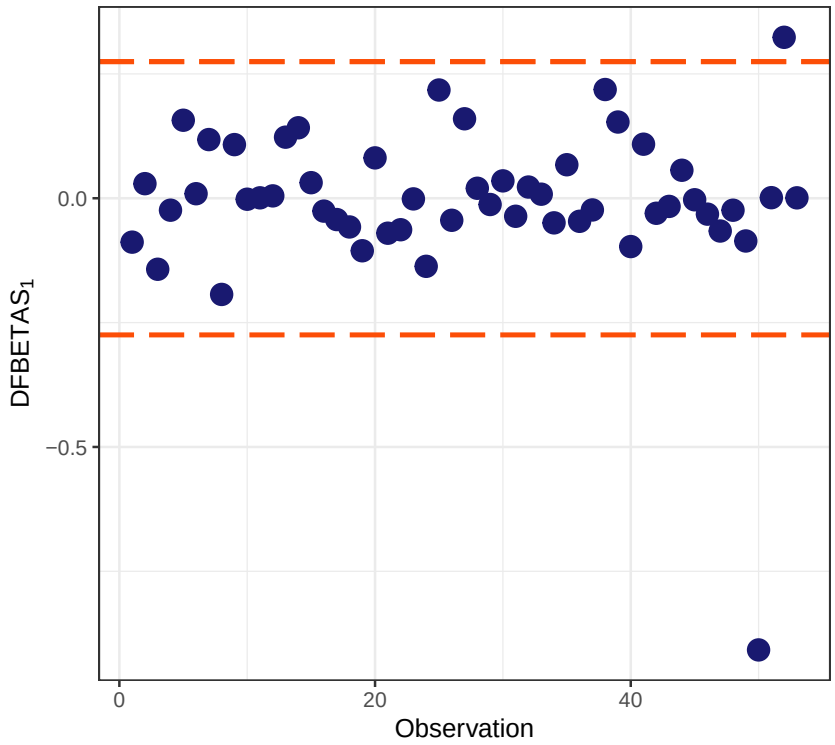
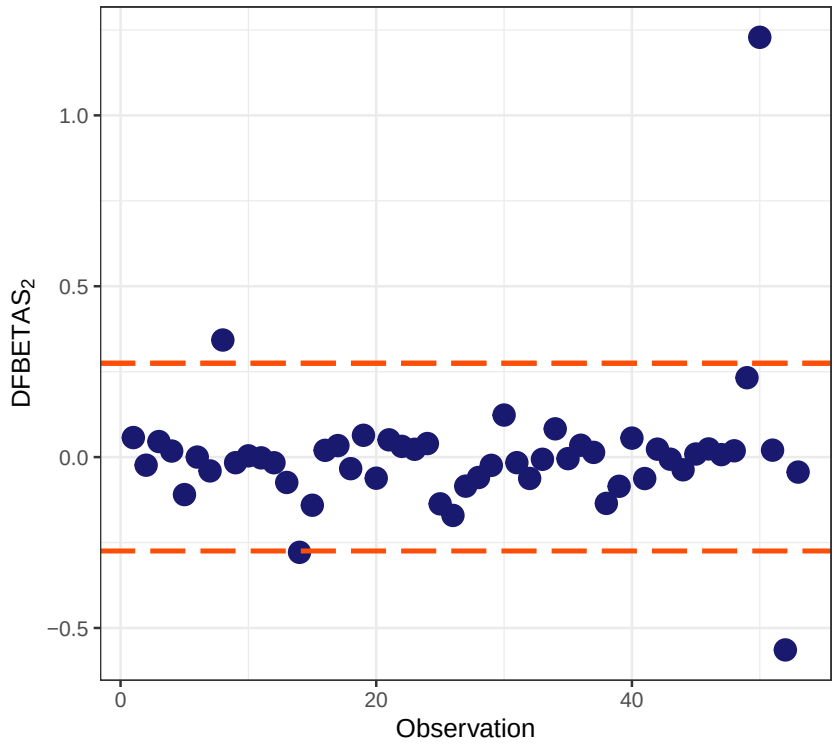


Figure 3: Cook's distance plot.



(a) Intercept (β_1).



(b) Energy usage (β_2).

Figure 4: DFBETAS plots.

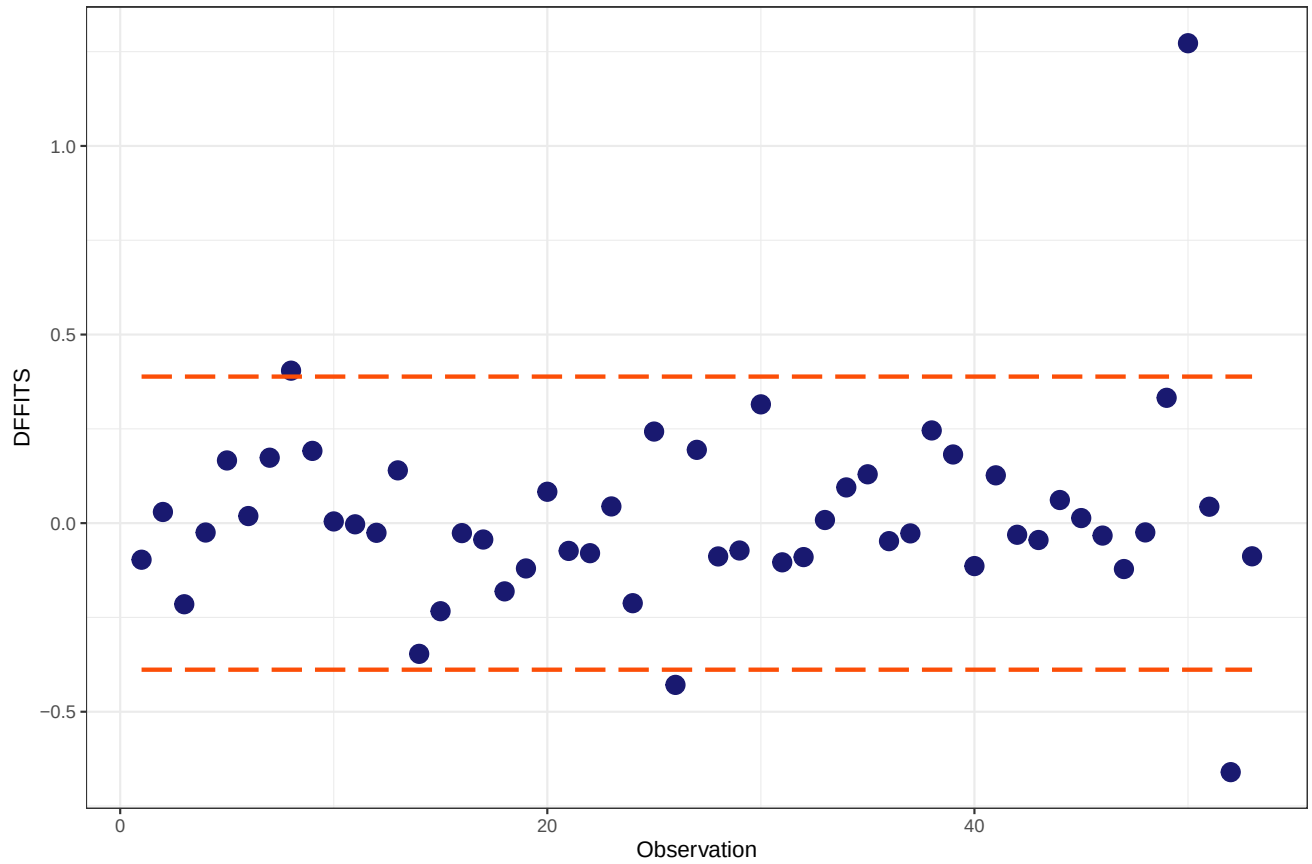


Figure 5: DFFITS plot.

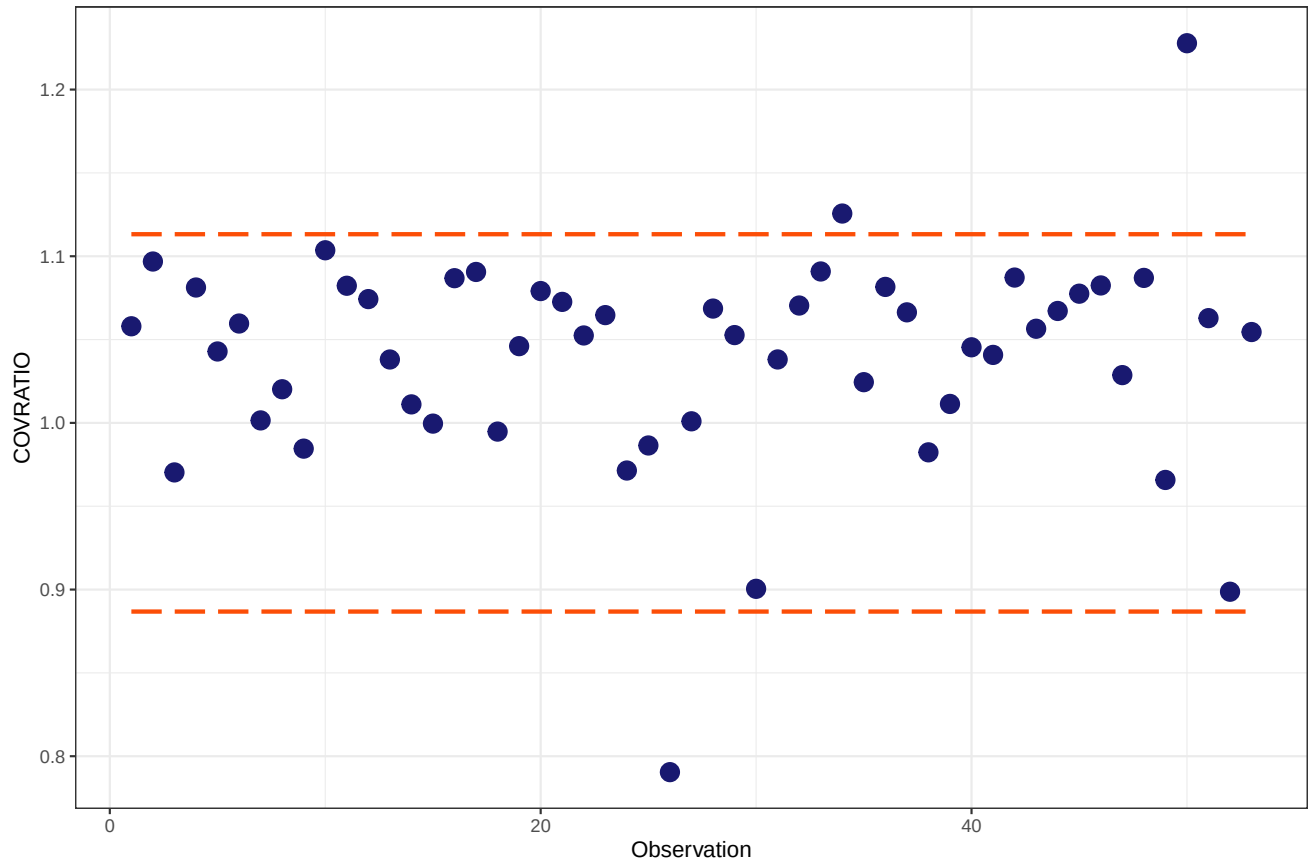


Figure 6: COVRATIO plot.

Outline

1 Introduction

2 Leverage points

3 Influential observations

4 Application

5 References

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. Wiley, New York, 6th edition.

Thank you!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Department of Statistics, Room 319

