Residual analysis

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, November 04, 2026

Outline

1. Introduction
2. Residuals
3. Residual standardization
4. Residual plots
5. PRESS statistic
6. Application
7. References

Outline

1. Introduction
2. Residuals
3. Residual standardization
4. Residual plots
5. PRESS statistic
6. Application
7. 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.

Distribution of the MLE

Under the normal linear regression model (NLM), the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right), \tag{3}$$

and, for the MLE of σ^2,

Distribution of the MLE

Under the normal linear regression model (NLM), the MLE of $\boldsymbol{\beta}$ has the exact distribution

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{\boldsymbol{p}}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right), \tag{3}$$

and, for the MLE of σ^2, we have

$$\frac{n\hat{\sigma}^2_{\text{MLE}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Distribution of the MLE

Under the normal linear regression model (NLM), the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right), \tag{3}$$

and, for the MLE of σ^2, we have

$$\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Note: $\hat{\sigma}^2_{\mathrm{MLE}}$ can also be written as a function of the mean square error (MSE).

Distribution of the MLE

Under the normal linear regression model (NLM), the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right), \tag{3}$$

and, for the MLE of σ^2, we have

$$\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Note: $\hat{\sigma}^2_{\mathrm{MLE}}$ can also be written as a function of the mean square error (MSE).

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

Residuals

We define the ℓth residual as the difference,

$$e_\ell = Y_\ell - \hat{Y}_\ell,$$

$\ell = 1, 2 \ldots, n$. As we saw previously, the residual can be expressed in matrix form:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y},$$

where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$.

Residuals

We define the ℓth residual as the difference,

$$e_\ell = Y_\ell - \hat{Y}_\ell,$$

$\ell = 1, 2 \ldots, n$. As we saw previously, the residual can be expressed in matrix form:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y},$$

where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$.

Residuals

The expected value (expectation vector) and the variance (covariance matrix) of $\boldsymbol{e}$ are given, respectively, by

$$\mathbb{E}(\boldsymbol{e}) = \mathbf{0} \text{ and } \text{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).$$

Residuals

The expected value (expectation vector) and the variance (covariance matrix) of $\boldsymbol{e}$ are given, respectively, by

$$\mathbb{E}(\boldsymbol{e}) = \mathbf{0} \text{ and } \text{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).$$

Let h_{ij} be the (i,j)th element of the matrix $\boldsymbol{H}$; then we have: $\mathbb{E}(e_i) = 0$, $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, and $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, $i, j = 1, 2, \ldots, n$.

Residuals

The expected value (expectation vector) and the variance (covariance matrix) of $\boldsymbol{e}$ are given, respectively, by

$$\mathbb{E}(\boldsymbol{e}) = \mathbf{0} \text{ and } \text{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).$$

Let h_{ij} be the (i,j)th element of the matrix $\boldsymbol{H}$; then we have: $\mathbb{E}(e_i) = 0$, $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, and $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, $i, j = 1, 2, \ldots, n$.

Residuals

Additionally, we can define an approximation for the mean variance as follows:

$$\sum_{\ell=1}^{n} \frac{(e_\ell - \bar{e})^2}{n-p} = \sum_{\ell=1}^{n} \frac{e_\ell^2}{n-p} = \text{MSE}\,.$$

Residuals

Additionally, we can define an approximation for the mean variance as follows:

$$\sum_{\ell=1}^{n} \frac{(e_\ell - \bar{e})^2}{n-p} = \sum_{\ell=1}^{n} \frac{e_\ell^2}{n-p} = \text{MSE}\,.$$

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

Residual standardization

Standardizing residuals helps us identify outliers and extreme values. We have four types of standardized residuals:

1. Standardized residuals (semi-studentized residuals);

Residual standardization

Standardizing residuals helps us identify outliers and extreme values. We have four types of standardized residuals:

1. Standardized residuals (semi-studentized residuals);
2. Studentized residuals (internally studentized);

Residual standardization

Standardizing residuals helps us identify outliers and extreme values. We have four types of standardized residuals:

1. Standardized residuals (semi-studentized residuals);
2. Studentized residuals (internally studentized);
3. PRESS residuals;

Residual standardization

Standardizing residuals helps us identify outliers and extreme values. We have four types of standardized residuals:

1. Standardized residuals (semi-studentized residuals);
2. Studentized residuals (internally studentized);
3. PRESS residuals;
4. R-student residuals (externally studentized).

Residual standardization

Standardizing residuals helps us identify outliers and extreme values. We have four types of standardized residuals:

1. Standardized residuals (semi-studentized residuals);
2. Studentized residuals (internally studentized);
3. PRESS residuals;
4. R-student residuals (externally studentized).

Residual standardization

The **semi-studentized residuals** or simply referred to as standardized residuals are expressed as follows:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **semi-studentized residuals** or simply referred to as standardized residuals are expressed as follows:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}}, \ \ell = 1, 2, \ldots, n.$$

The parameter σ^2 can be estimated by the MSE. d_ℓ has zero mean and variance approximately equal to 1. "Large" values of d_ℓ ($|d_\ell| > 3$) may indicate the presence of outliers.

Residual standardization

The **semi-studentized residuals** or simply referred to as standardized residuals are expressed as follows:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}},\ \ell = 1, 2, \ldots, n.$$

The parameter σ^2 can be estimated by the MSE. d_ℓ has zero mean and variance approximately equal to 1. "Large" values of d_ℓ ($|d_\ell| > 3$) may indicate the presence of outliers.

Residual standardization

The **studentized residuals** are expressed as follows:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **studentized residuals** are expressed as follows:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

The parameter σ^2 can be estimated by the MSE. r_ℓ has zero mean and variance equal to 1. Generally, the values of r_ℓ are larger than their corresponding d_ℓ.

Residual standardization

The **studentized residuals** are expressed as follows:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

The parameter σ^2 can be estimated by the MSE. r_ℓ has zero mean and variance equal to 1. Generally, the values of r_ℓ are larger than their corresponding d_ℓ.

Residual standardization

Let the ℓth **PRESS residual** (prediction error sum of squares) be

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)},\ \ell = 1, 2, \ldots, n.$$

Residual standardization

Let the ℓth **PRESS residual** (prediction error sum of squares) be

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)},\ \ell = 1, 2, \ldots, n.$$

where $\hat{Y}_{(\ell)}$ is the fitted response variable without observation ℓ and the ℓth row of the design matrix. This definition of residuals allows us to verify how unusual observation ℓ is.

Residual standardization

Let the ℓth **PRESS residual** (prediction error sum of squares) be

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)},\ \ell = 1, 2, \ldots, n.$$

where $\hat{Y}_{(\ell)}$ is the fitted response variable without observation ℓ and the ℓth row of the design matrix. This definition of residuals allows us to verify how unusual observation ℓ is.

Residual standardization

The ℓth PRESS residual can be written as follows (see Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The ℓth PRESS residual can be written as follows (see Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

And it is straightforward to show that:

$$\text{Var}(e_{(\ell)}) = \frac{\sigma^2}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The ℓth PRESS residual can be written as follows (see Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

And it is straightforward to show that:

$$\text{Var}(e_{(\ell)}) = \frac{\sigma^2}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **standardized PRESS residuals** are expressed as follows:

$$\frac{e_{(\ell)}}{\sqrt{\text{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1-h_{\ell\ell})}},\ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **standardized PRESS residuals** are expressed as follows:

$$\frac{e_{(\ell)}}{\sqrt{\text{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1-h_{\ell\ell})}},\ \ell = 1, 2, \ldots, n.$$

Notice that the standardized PRESS residuals coincide with the studentized residuals.

Residual standardization

The **standardized PRESS residuals** are expressed as follows:

$$\frac{e_{(\ell)}}{\sqrt{\text{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1-h_{\ell\ell})}},\ \ell = 1, 2, \ldots, n.$$

Notice that the standardized PRESS residuals coincide with the studentized residuals.

Residual standardization

Since we are taking into account the model fit without the ℓth observation, it is natural to consider an estimator for σ^2 that also omits observation ℓ.

This estimator is given by (see Montgomery et al., 2021, p. 621):

$$S^2_{(\ell)} = \frac{(n-p)\,\text{MSE} - \frac{e^2_\ell}{1-h_{\ell\ell}}}{n-p-1},\ \ell = 1, 2, \ldots, n.$$

Residual standardization

Since we are taking into account the model fit without the ℓth observation, it is natural to consider an estimator for σ^2 that also omits observation ℓ. This estimator is given by (see Montgomery et al., 2021, p. 621):

$$S^2_{(\ell)} = \frac{(n-p)\,\text{MSE} - \frac{e^2_\ell}{1-h_{\ell\ell}}}{n-p-1},\ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **R-student residual** is expressed as follows:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Residual standardization

The **R-student residual** is expressed as follows:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

This residual is also called the externally studentized residual.

Residual standardization

The **R-student residual** is expressed as follows:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

This residual is also called the externally studentized residual.

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

Residual plots

Plotting the residuals helps us observe their overall behavior and identify patterns.

Residual plots

- Normal probability plot of residuals;

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.
- Residuals vs. fitted values;

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.
- Residuals vs. fitted values;
  - Checks for non-constant variance;

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.
- Residuals vs. fitted values;
  - Checks for non-constant variance;
  - Checks for non-linearity;

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.
- Residuals vs. fitted values;
  - Checks for non-constant variance;
  - Checks for non-linearity;
  - Identifies potential outliers.

Residual plots

- Normal probability plot of residuals;
  - Checks the normality assumption.
- Residuals vs. fitted values;
  - Checks for non-constant variance;
  - Checks for non-linearity;
  - Identifies potential outliers.

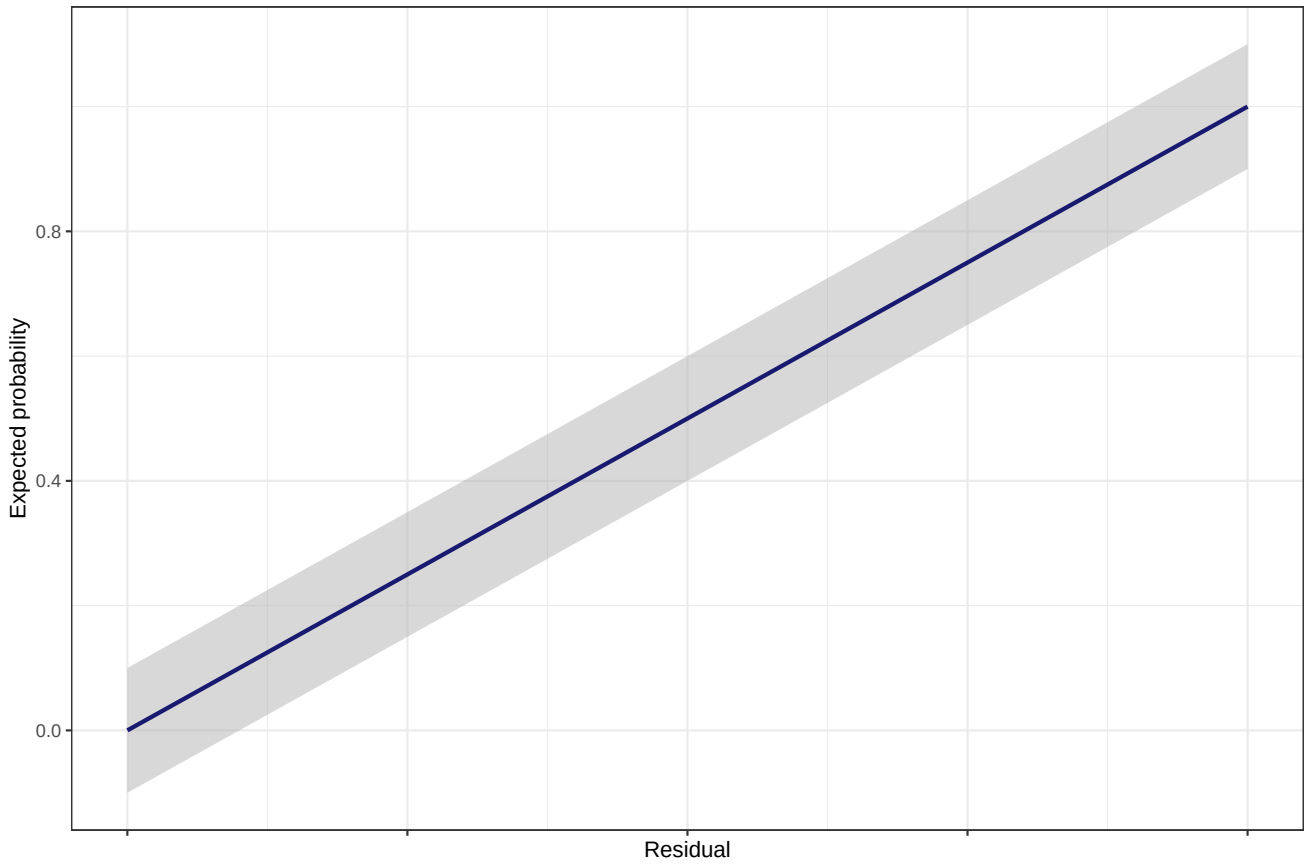
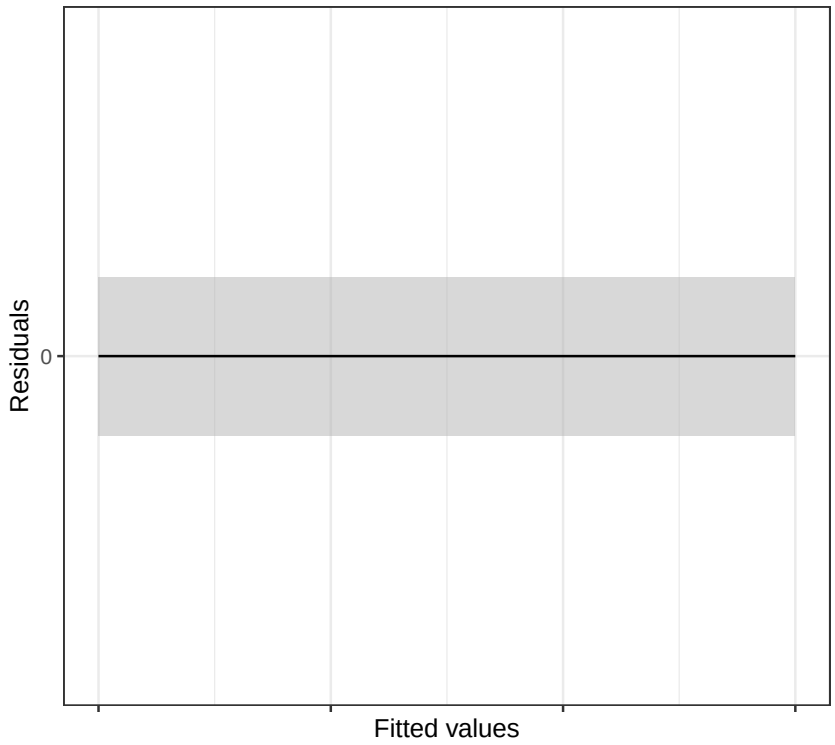
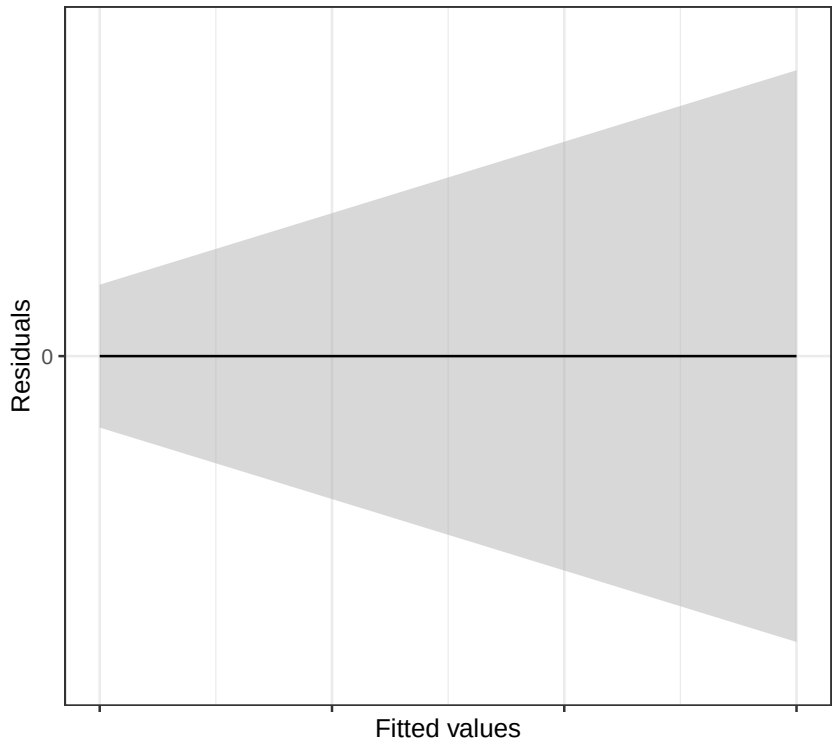


Figure 1: Normal probability plot of residuals.



(a) Constant variance.



(b) Non-constant variance.

Figure 2: Residuals versus fitted values.

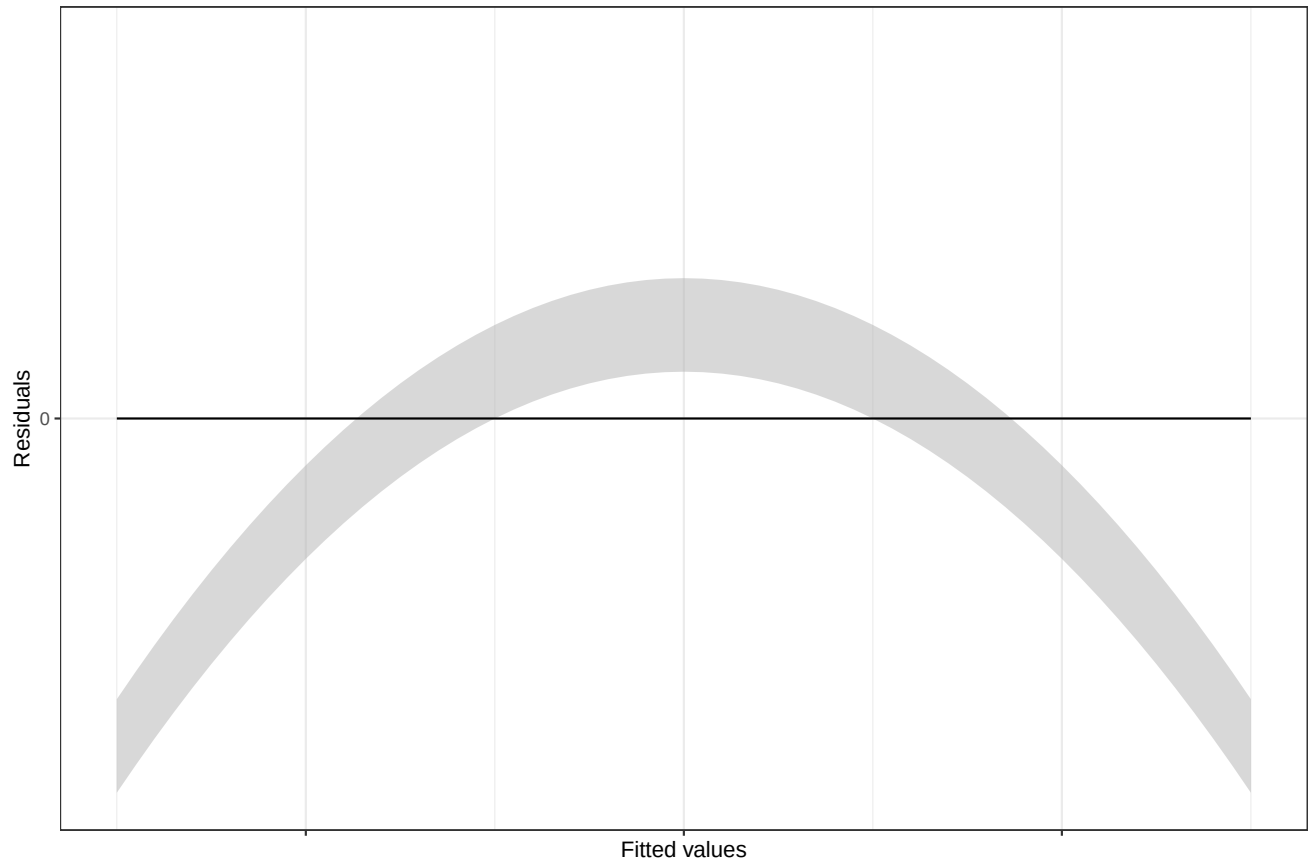


Figure 3: Non-linear pattern in residuals.

Residual plots

- Residuals vs. regressor variables in the model;

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.
- Residuals vs. regressor variables not in the model;

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.
- Residuals vs. regressor variables not in the model;
  - If a pattern emerges, it may indicate that adding this regressor can improve the model fit.

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.
- Residuals vs. regressor variables not in the model;
  - If a pattern emerges, it may indicate that adding this regressor can improve the model fit.
- Residuals vs. observation order.

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.
- Residuals vs. regressor variables not in the model;
  - If a pattern emerges, it may indicate that adding this regressor can improve the model fit.
- Residuals vs. observation order.
  - Checks for correlated errors.

Residual plots

- Residuals vs. regressor variables in the model;
  - Checks for non-constant variance;
  - Checks for non-linearity.
- Residuals vs. regressor variables not in the model;
  - If a pattern emerges, it may indicate that adding this regressor can improve the model fit.
- Residuals vs. observation order.
  - Checks for correlated errors.

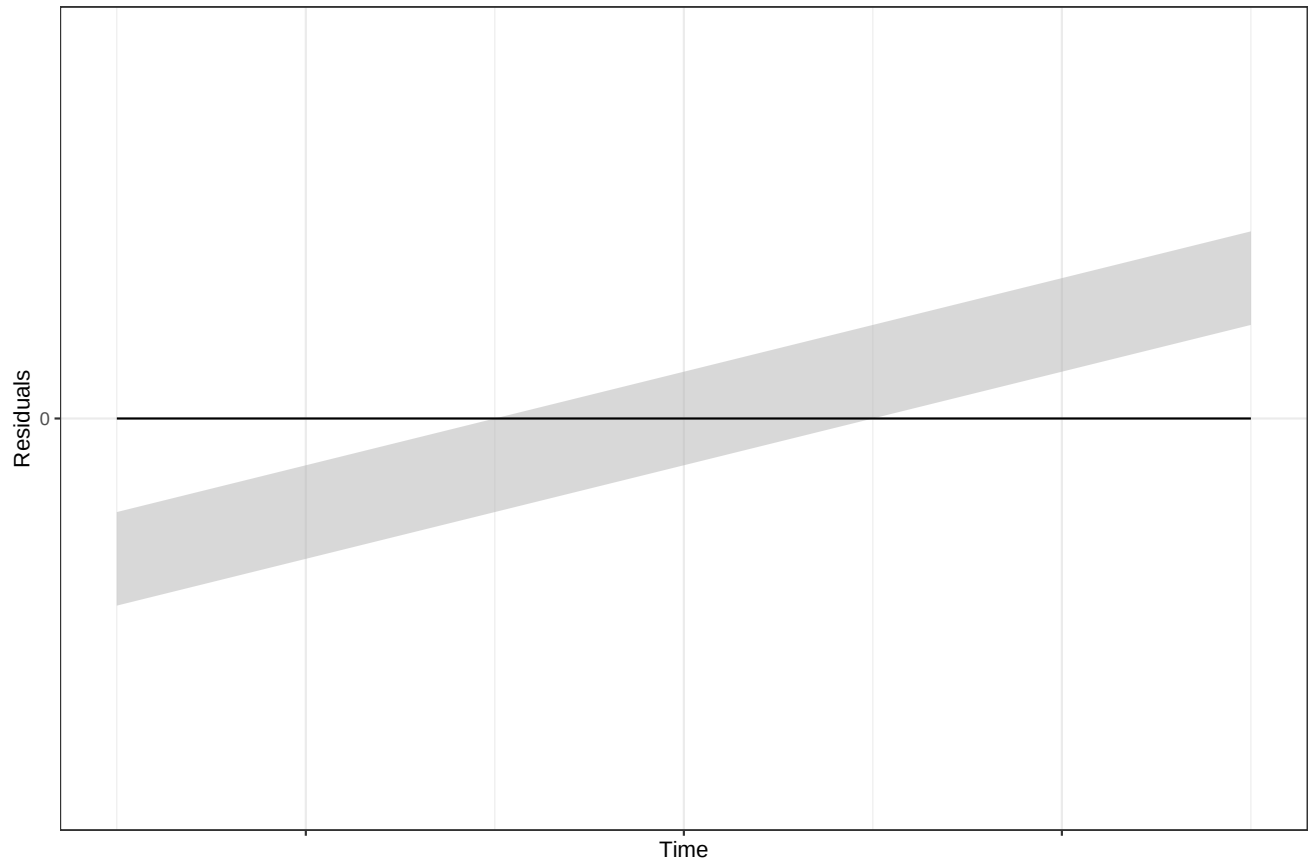


Figure 4: Non-independence of errors.

Partial regression plots

An alternative way to assess the marginal contribution of an independent variable x_m, $m = 1, 2, \ldots, p$, to the model fitted with the remaining variables is through partial regression plots. Furthermore, this method can also be used to check the correct relationship between Y and x_m.

Partial regression plots

The method consists of fitting a regression model without the covariate x_m and computing the residuals. Next, fitting a regression model with x_m as the response variable and the remaining covariates as regressors, and computing the residuals from this analysis. Finally, constructing a scatter plot of these two sets of residuals.

Partial regression plots

If

- the plot appears linear, then a linear relationship between Y and x_m seems reasonable;

Partial regression plots

If

- the plot appears linear, then a linear relationship between Y and x_m seems reasonable;
- the plot is curvilinear, it may be necessary to work with x_m^2 or $1/x_m$;

Partial regression plots

If

- the plot appears linear, then a linear relationship between Y and x_m seems reasonable;
- the plot is curvilinear, it may be necessary to work with x_m^2 or $1/x_m$;
- x_m is a candidate variable and a horizontal "band" appears, this variable adds no new information.

Partial regression plots

If

- the plot appears linear, then a linear relationship between Y and x_m seems reasonable;
- the plot is curvilinear, it may be necessary to work with x_m^2 or $1/x_m$;
- x_m is a candidate variable and a horizontal "band" appears, this variable adds no new information.

Partial regression plots

Considerations:

- Use with caution; they only suggest possible relationships;

Partial regression plots

Considerations:

- Use with caution; they only suggest possible relationships;
- They generally do not detect interaction effects;

Partial regression plots

Considerations:

- Use with caution; they only suggest possible relationships;
- They generally do not detect interaction effects;
- If multicollinearity is present, partial regression plots may yield misleading information.

Partial regression plots

Considerations:

- Use with caution; they only suggest possible relationships;
- They generally do not detect interaction effects;
- If multicollinearity is present, partial regression plots may yield misleading information.

Other plots

Constructing a scatter plot for all pairs of covariates can provide information about the relationship between them, such as:

- potential correlation;

Other plots

Constructing a scatter plot for all pairs of covariates can provide information about the relationship between them, such as:

- potential correlation;
- the presence of outliers.

Other plots

Constructing a scatter plot for all pairs of covariates can provide information about the relationship between them, such as:

- potential correlation;
- the presence of outliers.

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

PRESS statistic

A valuable method for model comparison is the PRESS statistic (Montgomery et al., 2021), defined as follows:

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^2 = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2 .$$

PRESS statistic

A valuable method for model comparison is the PRESS statistic (Montgomery et al., 2021), defined as follows:

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^2 = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2 .$$

Remark: small values of the PRESS statistic are desirable.

PRESS statistic

A valuable method for model comparison is the PRESS statistic (Montgomery et al., 2021), defined as follows:

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^2 = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2 .$$

Remark: small values of the PRESS statistic are desirable.

PRESS statistic

It is also possible to define a coefficient of determination for the PRESS statistic as follows:

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SST}}.$$

PRESS statistic

It is also possible to define a coefficient of determination for the PRESS statistic as follows:

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SST}}.$$

Interpretation: we expect the model to explain approximately $R^2\%$ of the variability in predicting a new observation.

PRESS statistic

It is also possible to define a coefficient of determination for the PRESS statistic as follows:

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SST}}.$$

Interpretation: we expect the model to explain approximately $R^2\%$ of the variability in predicting a new observation.

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

Application

(Montgomery et al., 2021, p. 76) A soft drink bottler aims to predict delivery time (Y, in minutes) to service vending machines. An engineer identified two key predictors: the number of cases stocked (x_2) and the distance walked by the driver (x_3, in feet). The least-squares fit is

$$\hat{Y}_\ell = 2.341 + 1.616x_{\ell 2} + 0.014x_{\ell 3},$$

$\ell = 1, 2, \ldots, 25$.

Example

We also have:

Table 1: Parameter estimates.

Parameter	Estimate	SE	t_c
β_1	2.341	1.097	2.135
β_2	1.616	0.171	9.464
β_3	0.014	0.004	3.981

Critical region, for $\alpha = 5\%$: $|t_c| > 2.074$ with MSE $= 10.624$.

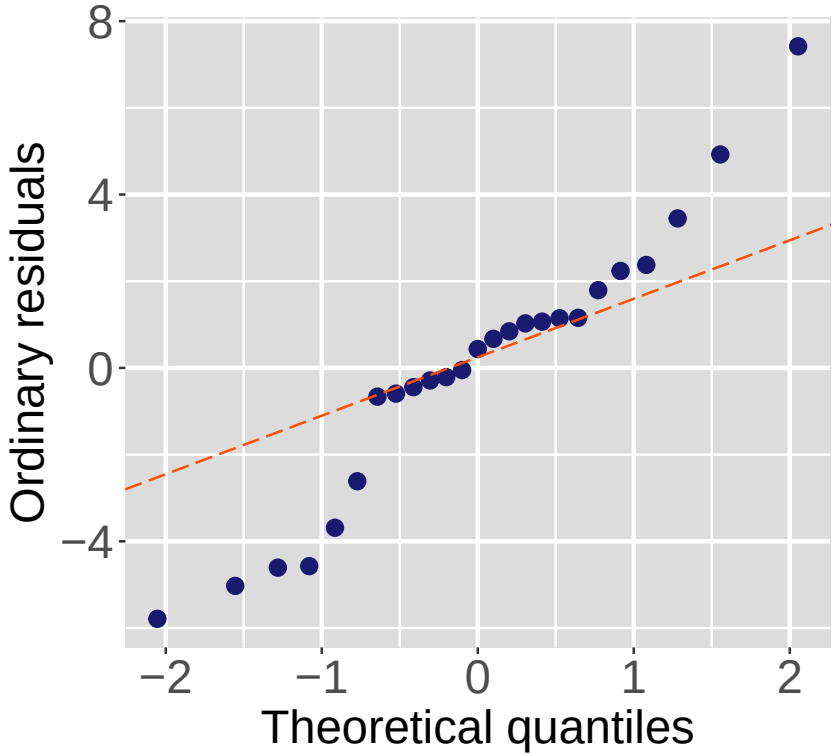
Example

We also have:

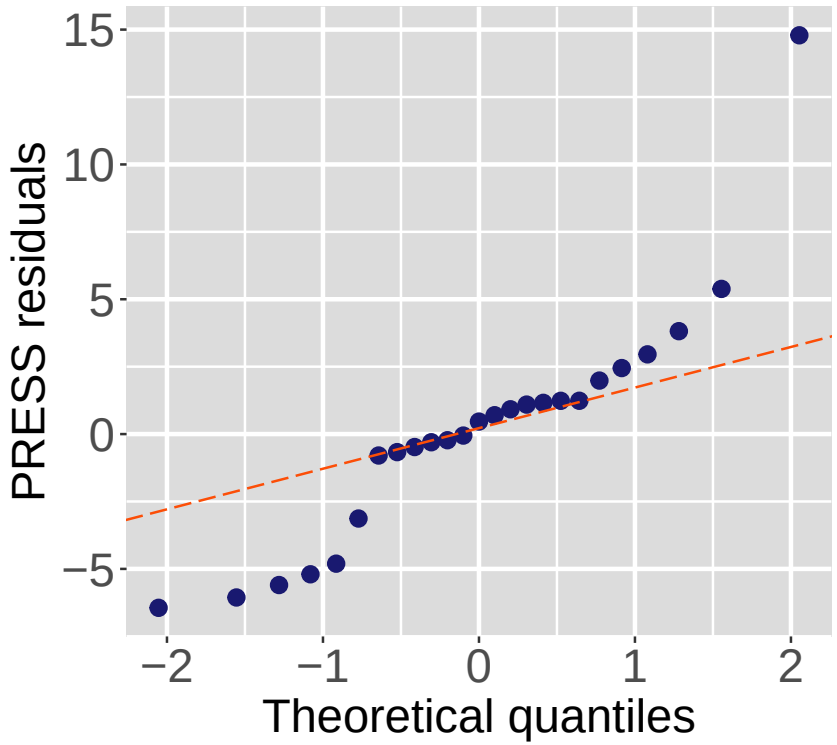
Table 1: Parameter estimates.

Parameter	Estimate	SE	t_c
β_1	2.341	1.097	2.135
β_2	1.616	0.171	9.464
β_3	0.014	0.004	3.981

Critical region, for $\alpha = 5\%$: $|t_c| > 2.074$ with MSE $= 10.624$.

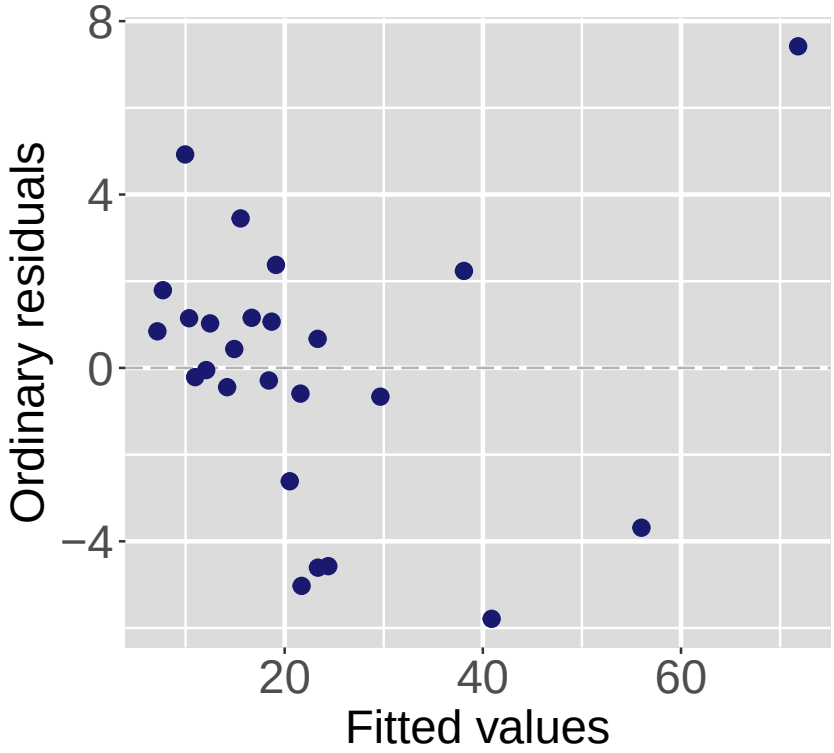


(a) Ordinary residuals.

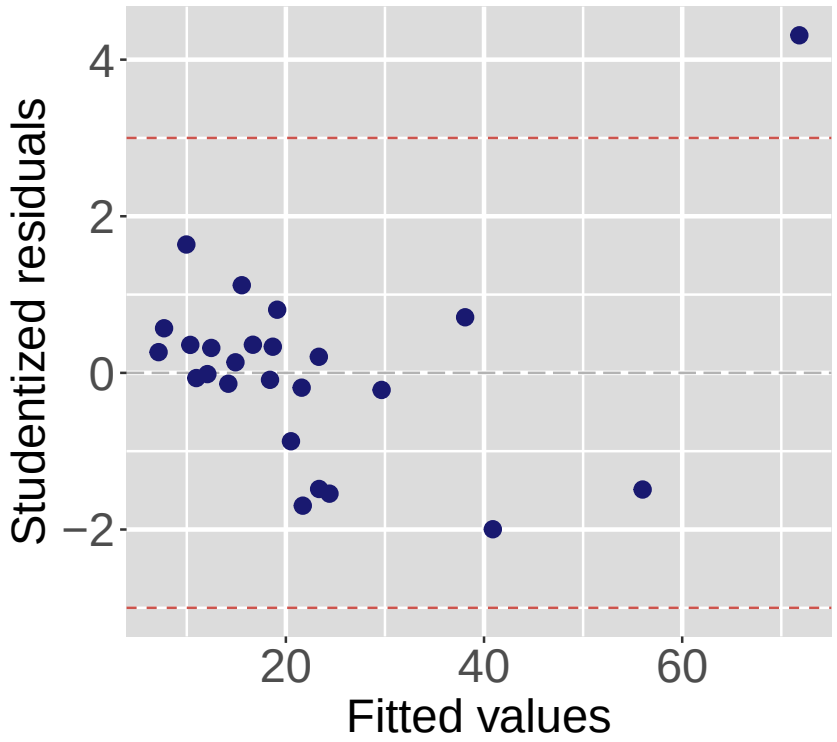


(b) PRESS residuals.

Figure 5: Normal QQ plots of residuals.



(a) Ordinary residuals.



(b) Studentized residuals.

Figure 6: Residuals versus fitted values.

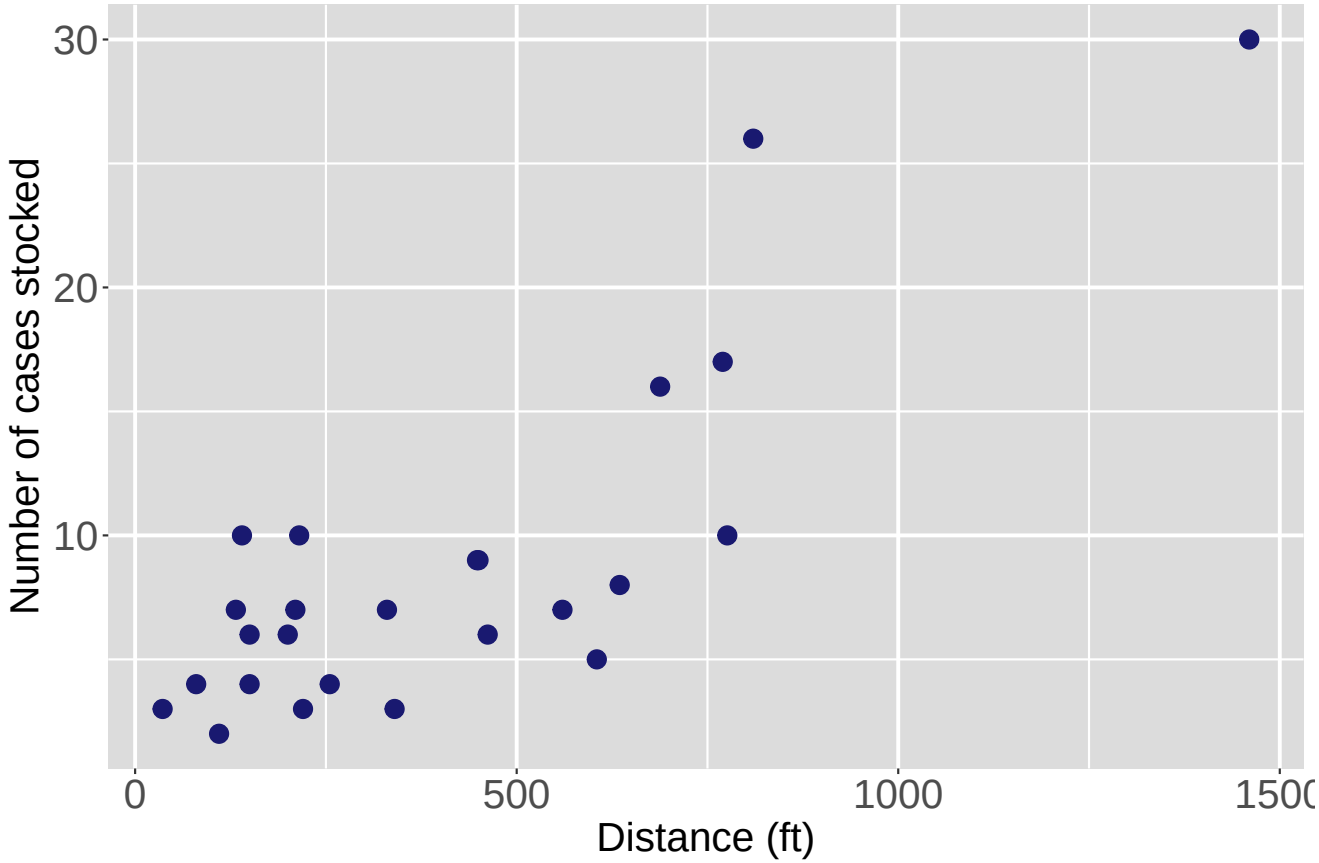


Figure 7: Scatter plot of distance versus cases stocked.

Outline

1 Introduction

2 Residuals

3 Residual standardization

4 Residual plots

5 PRESS statistic

6 Application

7 References

References

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