

Analysis of variance

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, September 23, 2026



Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study
- 7 References



Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study
- 7 References



Linear regression model

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \quad (1)$$



Linear regression model

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \quad (1)$$

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



Linear regression model

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \quad (1)$$

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



Linear regression model

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \quad (1)$$

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



Matrix form

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$



Matrix form

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

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



Matrix form

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

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



Introduction

Motivation

Analysis of variance (ANOVA) is a way of viewing the linear regression model from another perspective.



Introduction

Motivation

Analysis of variance (ANOVA) is a way of viewing the linear regression model from another perspective.



Outline

- 1 Introduction
- 2 Partition of sums of squares**
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study
- 7 References



Total sum of squares

The analysis of variance method is based on the partition of the sums of squares and the associated degrees of freedom for the response variable Y .

If all observations Y_ℓ are identical, regression analysis is not necessary.



Total sum of squares

The analysis of variance method is based on the partition of the sums of squares and the associated degrees of freedom for the response variable Y .

If all observations Y_ℓ are identical, regression analysis is not necessary. The variation of Y_ℓ would then be measured in terms of the deviations:

$$Y_\ell - \bar{Y}.$$

Figure 1 presents this deviation graphically.



Total sum of squares

The analysis of variance method is based on the partition of the sums of squares and the associated degrees of freedom for the response variable Y .

If all observations Y_ℓ are identical, regression analysis is not necessary. The variation of Y_ℓ would then be measured in terms of the deviations:

$$Y_\ell - \bar{Y}.$$

Figure 1 presents this deviation graphically.



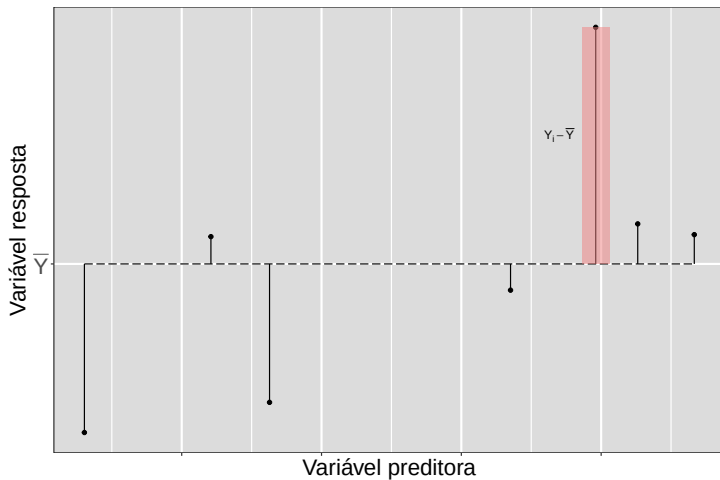


Figure 1: $Y_\ell - \bar{Y}$.

Total sum of squares

The measure of the total variation of the deviations, the **total sum of squares** (SST), is given by

$$SST = \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2.$$

Total sum of squares

The measure of the total variation of the deviations, the **total sum of squares** (SST), is given by

$$\text{SST} = \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2.$$

Note that, if $\text{SST} = 0$, all observations are equal, and the larger SST is, the greater the variability of the data.

Total sum of squares

The measure of the total variation of the deviations, the **total sum of squares** (SST), is given by

$$\text{SST} = \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2.$$

Note that, if $\text{SST} = 0$, all observations are equal, and the larger SST is, the greater the variability of the data.



Residual sum of squares

As we saw in the previous class, in a regression analysis, the variability of the data is measured relative to the regression line, that is,

$$Y_\ell - \hat{Y}_\ell.$$

Figure 2 presents this deviation graphically.



Residual sum of squares

As we saw in the previous class, in a regression analysis, the variability of the data is measured relative to the regression line, that is,

$$Y_{\ell} - \hat{Y}_{\ell}.$$

Figure 2 presents this deviation graphically.



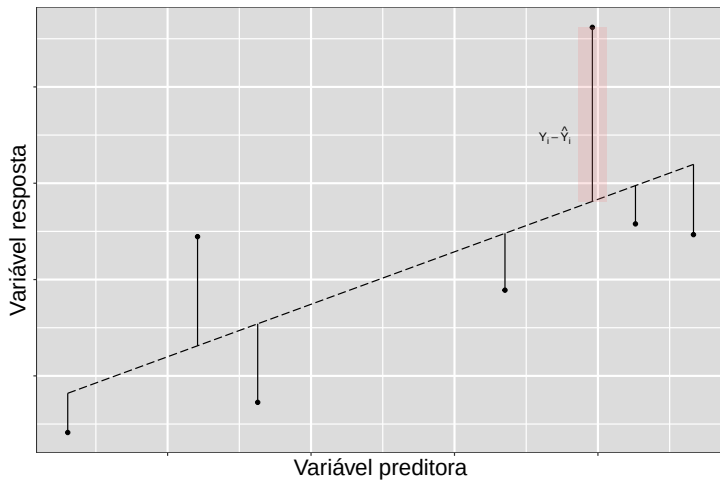


Figure 2: $Y_\ell - \hat{Y}_\ell$.

Residual sum of squares

The measure used to assess all the variation in the sample is the **residual sum of squares** (RSS), given by

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

Residual sum of squares

The measure used to assess all the variation in the sample is the **residual sum of squares** (RSS), given by

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

If $\text{RSS} = 0$, all observations lie on the regression line, and the larger RSS is, the greater the variability of the data relative to the regression line.



Residual sum of squares

The measure used to assess all the variation in the sample is the **residual sum of squares** (RSS), given by

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

If $\text{RSS} = 0$, all observations lie on the regression line, and the larger RSS is, the greater the variability of the data relative to the regression line.

Regression sum of squares

It is also possible to compare the distance between the value estimated by the regression line and the mean, that is,

$$\hat{Y}_\ell - \bar{Y}.$$

Figure 3 presents this deviation graphically.



Regression sum of squares

It is also possible to compare the distance between the value estimated by the regression line and the mean, that is,

$$\hat{Y}_\ell - \bar{Y}.$$

Figure 3 presents this deviation graphically.



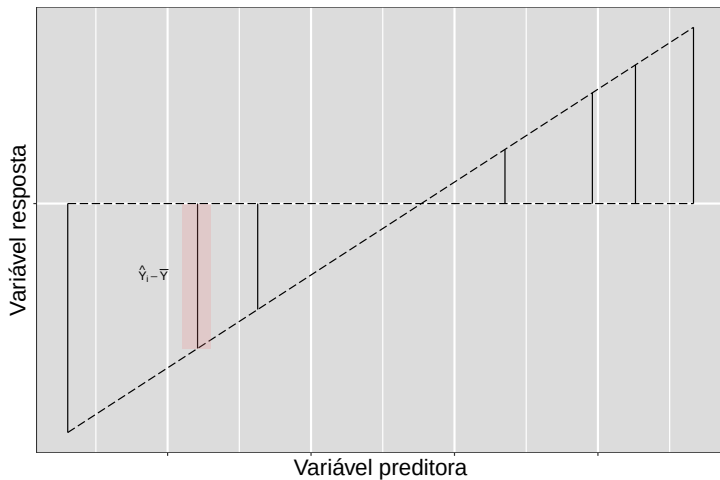


Figure 3: $\hat{Y}_\ell - \bar{Y}$.

Regression sum of squares

Consequently, we can measure the squared distance between them, called the **regression sum of squares** (SSR),

$$SSR = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

Regression sum of squares

Consequently, we can measure the squared distance between them, called the **regression sum of squares** (SSR),

$$\text{SSR} = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

If $\text{SSR} = 0$, the regression line is horizontal.



Regression sum of squares

Consequently, we can measure the squared distance between them, called the **regression sum of squares** (SSR),

$$\text{SSR} = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

If $\text{SSR} = 0$, the regression line is horizontal.



Sums of squares

Consider the deviation $Y_\ell - \bar{Y}$, the basic measure for quantifying the variation of the observations Y_ℓ . The difference between the ℓ -th value of the response variable and the sample mean can be decomposed as the sum of two terms:

$$Y_\ell - \bar{Y} = (Y_\ell - \hat{Y}_\ell) + (\hat{Y}_\ell - \bar{Y}). \quad (3)$$

Figure 4 presents this deviation graphically.



Sums of squares

Consider the deviation $Y_\ell - \bar{Y}$, the basic measure for quantifying the variation of the observations Y_ℓ . The difference between the ℓ -th value of the response variable and the sample mean can be decomposed as the sum of two terms:

$$Y_\ell - \bar{Y} = (Y_\ell - \hat{Y}_\ell) + (\hat{Y}_\ell - \bar{Y}). \quad (3)$$

Figure 4 presents this deviation graphically.



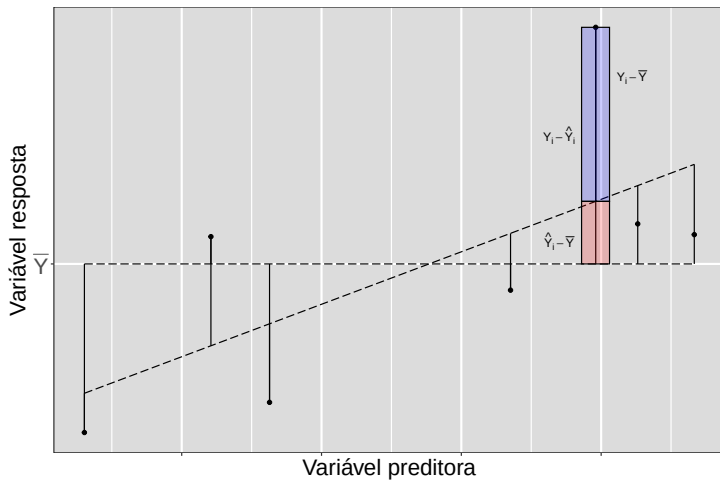


Figure 4: $Y_\ell - \hat{Y}_\ell$ and $\hat{Y}_\ell - \bar{Y}$.

Sums of squares

From (3) and rearranging conveniently, we have

$$\begin{aligned} \text{SST} &= \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2 = \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y}) + (Y_{\ell} - \hat{Y}_{\ell})]^2 \\ &= \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y})^2 + 2(\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell}) + (Y_{\ell} - \hat{Y}_{\ell})^2] \\ &= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \underbrace{2 \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell})}_0 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\ &= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\ &= \text{SSR} + \text{RSS}. \end{aligned}$$

Sums of squares

From (3) and rearranging conveniently, we have

$$\begin{aligned} \text{SST} &= \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2 = \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y}) + (Y_{\ell} - \hat{Y}_{\ell})]^2 \\ &= \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y})^2 + 2(\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell}) + (Y_{\ell} - \hat{Y}_{\ell})^2] \\ &= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \underbrace{2 \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell})}_0 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\ &= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\ &= \text{SSR} + \text{RSS}. \end{aligned}$$



Sums of squares

Therefore, we have $SST = SSR + RSS$, that is, the total variability of the data is divided into two parts:

- the variability explained by the model (SSR);



Sums of squares

Therefore, we have $SST = SSR + RSS$, that is, the total variability of the data is divided into two parts:

- the variability explained by the model (SSR);
- the variability that cannot be explained by the regression (RSS).



Sums of squares

Therefore, we have $SST = SSR + RSS$, that is, the total variability of the data is divided into two parts:

- the variability explained by the model (SSR);
- the variability that cannot be explained by the regression (RSS).



Sums of squares

In matrix notation, SST, SSR, and RSS are given, respectively, by

$$\begin{aligned} \text{SST} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SSR} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{RSS} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

Sums of squares

In matrix notation, SST, SSR, and RSS are given, respectively, by

$$\begin{aligned} \text{SST} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SSR} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{RSS} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

where $\mathbf{J}$ is an $n \times n$ matrix with all its elements equal to 1. For demonstrations, it is convenient to write $\frac{1}{n} \mathbf{J} = \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} \mathbf{1}_n^\top$, where $\mathbf{1}_n$ is a vector of length n consisting of 1's.

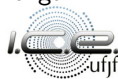


Sums of squares

In matrix notation, SST, SSR, and RSS are given, respectively, by

$$\begin{aligned} \text{SST} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SSR} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{RSS} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

where $\mathbf{J}$ is an $n \times n$ matrix with all its elements equal to 1. For demonstrations, it is convenient to write $\frac{1}{n} \mathbf{J} = \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} \mathbf{1}_n^\top$, where $\mathbf{1}_n$ is a vector of length n consisting of 1's.



Degrees of freedom

The degrees of freedom (DF) are values associated with the sums of squares. They are the number of independent pieces of information, involving the observations $Y_1, Y_2, \dots, Y_n$, that are necessary to calculate the sums of squares.



Degrees of freedom

Basically, it is the sample size minus the number of restrictions. For each sum of squares, we have

- SST has $n - 1$ DF;



Degrees of freedom

Basically, it is the sample size minus the number of restrictions. For each sum of squares, we have

- SST has $n - 1$ DF;
- SSR has $p - 1$ DF;



Degrees of freedom

Basically, it is the sample size minus the number of restrictions. For each sum of squares, we have

- SST has $n - 1$ DF;
- SSR has $p - 1$ DF;
- RSS has $n - p$ DF.



Degrees of freedom

Basically, it is the sample size minus the number of restrictions. For each sum of squares, we have

- SST has $n - 1$ DF;
- SSR has $p - 1$ DF;
- RSS has $n - p$ DF.



Mean square

The regression mean square (MSR) and the residual mean square (MSE) are defined by, respectively, dividing the sums of squares by the degrees of freedom, that is,

$$\text{MSR} = \frac{\text{SSR}}{p - 1} \text{ and } \text{MSE} = \frac{\text{RSS}}{n - p}. \quad (5)$$



Mean square

The regression mean square (MSR) and the residual mean square (MSE) are defined by, respectively, dividing the sums of squares by the degrees of freedom, that is,

$$\text{MSR} = \frac{\text{SSR}}{p - 1} \text{ and } \text{MSE} = \frac{\text{RSS}}{n - p}. \quad (5)$$

Mean square

It can be shown that (Montgomery et al., 2012, p. 581),

$$\mathbb{E}(\text{MSR}) = \sigma^2 + f(\beta_2, \beta_3, \dots, \beta_p),$$

$$\mathbb{E}(\text{MSE}) = \sigma^2,$$

where $f(\cdot)$ is a positive function of $\beta_2, \beta_3, \dots, \beta_p$, equal to 0 when $\beta_2 = \beta_3 = \dots = \beta_p = 0$.



Mean square

It can be shown that (Montgomery et al., 2012, p. 581),

$$\mathbb{E}(\text{MSR}) = \sigma^2 + f(\beta_2, \beta_3, \dots, \beta_p),$$

$$\mathbb{E}(\text{MSE}) = \sigma^2,$$

where $f(\cdot)$ is a positive function of $\beta_2, \beta_3, \dots, \beta_p$, equal to 0 when $\beta_2 = \beta_3 = \dots = \beta_p = 0$.



Mean square

Additionally, when $\beta_2 = \beta_3 = \dots = \beta_p = 0$, we have that,

$$F_c = \frac{\text{MSR}}{\text{MSE}} \sim F(p-1, n-p).$$

Mean square

Additionally, when $\beta_2 = \beta_3 = \dots = \beta_p = 0$, we have that,

$$F_c = \frac{\text{MSR}}{\text{MSE}} \sim F(p - 1, n - p).$$

ANOVA

Table 1: Analysis of variance.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_c
Regression	SSR	$p - 1$	MSR	$\frac{MSR}{MSE}$
Residuals	RSS	$n - p$	MSE	
Total	SST	$n - 1$		

Hypothesis testing

Let $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$. The hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H} : \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A} : \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

Hypothesis testing

Let $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$. The hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H}: \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A}: \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}}{\text{MSE}}.$$



Hypothesis testing

Let $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$. The hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H}: \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A}: \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}}{\text{MSE}}.$$



Hypothesis testing

Under $\mathcal{H}$, $F_c \sim F(p-1, n-p)$. Thus, we reject $\mathcal{H}$ if $F_c > F(1-\alpha; p-1, n-p)$, that is, the null hypothesis is rejected, for a given nominal level α , if the test statistic is greater than the $100(1-\alpha)\%$ quantile of an $F(p-1, n-p)$ distribution.



Hypothesis testing

Interpretation

In this case, we assume that there is a linear relationship between the response variable and at least one predictor variable, at the α significance level.

Hypothesis testing

Interpretation

In this case, we assume that there is a linear relationship between the response variable and at least one predictor variable, at the α significance level.

Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination**
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study
- 7 References



Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{RSS}{SST}. \quad (6)$$

Note that $0 \leq R^2 \leq 1$.

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{RSS}{SST}. \quad (6)$$

Note that $0 \leq R^2 \leq 1$. It tends to increase as new covariates are added to the model (even nonsignificant ones).



Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{RSS}{SST}. \quad (6)$$

Note that $0 \leq R^2 \leq 1$. It tends to increase as new covariates are added to the model (even nonsignificant ones).



Coefficient of determination

Interpretation 1

It measures the reduction in uncertainty when predicting the response variable by using the considered regression model instead of predicting it by its sample mean.

Coefficient of determination

Interpretation 1

It measures the reduction in uncertainty when predicting the response variable by using the considered regression model instead of predicting it by its sample mean.



Coefficient of determination

Interpretation 2

It measures the proportion of the variability of the response variable that is explained by the considered regression model.

Coefficient of determination

Interpretation 2

It measures the proportion of the variability of the response variable that is explained by the considered regression model.

Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \cdots = \beta_p = 0$.

Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \cdots = \beta_p = 0$.
- $R^2 = 1$ if $Y_\ell = \hat{Y}_\ell$ for all observations.

Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ if $Y_\ell = \hat{Y}_\ell$ for all observations.
- R^2 is the simple coefficient of determination of a regression model of Y_ℓ on $\hat{Y}_\ell$.

Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ if $Y_\ell = \hat{Y}_\ell$ for all observations.
- R^2 is the simple coefficient of determination of a regression model of Y_ℓ on $\hat{Y}_\ell$.
- A high R^2 does not necessarily indicate that accurate predictions can be made.

Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ if $Y_\ell = \hat{Y}_\ell$ for all observations.
- R^2 is the simple coefficient of determination of a regression model of Y_ℓ on $\hat{Y}_\ell$.
- A high R^2 does not necessarily indicate that accurate predictions can be made.
- It necessarily increases or remains constant with the inclusion of an additional predictor variable.



Remarks

- $R^2 = 0$ if $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ if $Y_\ell = \hat{Y}_\ell$ for all observations.
- R^2 is the simple coefficient of determination of a regression model of Y_ℓ on $\hat{Y}_\ell$.
- A high R^2 does not necessarily indicate that accurate predictions can be made.
- It necessarily increases or remains constant with the inclusion of an additional predictor variable.



Adjusted coefficient of determination

$$R_a^2 = 1 - \frac{RSS/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p} \frac{RSS}{SST}. \quad (7)$$

It penalizes models that add covariates with little significance.



Adjusted coefficient of determination

$$R_a^2 = 1 - \frac{RSS/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p} \frac{RSS}{SST}. \quad (7)$$

It penalizes models that add covariates with little significance.

Correlation coefficient

$$r = \sqrt{R^2}.$$

Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares**
- 5 Application
- 6 Simulation study
- 7 References



Extra sum of squares

Extra sums of squares measure the marginal reduction in the residual sum of squares when one or more predictor variables are added to the model, given that the other predictor variables are already present in the model.



Extra sum of squares

Example 1. Consider the following LRM,

$$y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell.$$

Extra sum of squares

Example 1. Consider the following LRM,

$$y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell.$$

Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;

Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;
- $RSS(\mathbf{x}_2, \mathbf{x}_3)$: RSS in a model with $\mathbf{x}_2$ and $\mathbf{x}_3$;

Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;
- $RSS(\mathbf{x}_2, \mathbf{x}_3)$: RSS in a model with $\mathbf{x}_2$ and $\mathbf{x}_3$;
- $SSR(\mathbf{x}_3|\mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3)$: Extra sum of squares of $\mathbf{x}_3$ given that $\mathbf{x}_2$ is already present in the model.

Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;
- $RSS(\mathbf{x}_2, \mathbf{x}_3)$: RSS in a model with $\mathbf{x}_2$ and $\mathbf{x}_3$;
- $SSR(\mathbf{x}_3|\mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3)$: Extra sum of squares of $\mathbf{x}_3$ given that $\mathbf{x}_2$ is already present in the model.

It can be shown that,



Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;
- $RSS(\mathbf{x}_2, \mathbf{x}_3)$: RSS in a model with $\mathbf{x}_2$ and $\mathbf{x}_3$;
- $SSR(\mathbf{x}_3|\mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3)$: Extra sum of squares of $\mathbf{x}_3$ given that $\mathbf{x}_2$ is already present in the model.

It can be shown that, $SSR(\mathbf{x}_3|\mathbf{x}_2) = SSR(\mathbf{x}_2, \mathbf{x}_3) - SSR(\mathbf{x}_2)$.

Extra sum of squares

- $RSS(\mathbf{x}_2)$: RSS in a model with only $\mathbf{x}_2$;
- $RSS(\mathbf{x}_2, \mathbf{x}_3)$: RSS in a model with $\mathbf{x}_2$ and $\mathbf{x}_3$;
- $SSR(\mathbf{x}_3|\mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3)$: Extra sum of squares of $\mathbf{x}_3$ given that $\mathbf{x}_2$ is already present in the model.

It can be shown that, $SSR(\mathbf{x}_3|\mathbf{x}_2) = SSR(\mathbf{x}_2, \mathbf{x}_3) - SSR(\mathbf{x}_2)$.

Extra sum of squares

Example 2. Consider the following LRM,

$$y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \beta_4 x_{\ell 4} + \varepsilon_\ell.$$

Extra sum of squares

Example 2. Consider the following LRM,

$$y_{\ell} = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \beta_4 x_{\ell 4} + \varepsilon_{\ell}.$$

Extra sum of squares

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: Extra sum of squares of $\mathbf{x}_4$ given that $\mathbf{x}_2$ and $\mathbf{x}_3$ are already present in the model;

Extra sum of squares

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: Extra sum of squares of $\mathbf{x}_4$ given that $\mathbf{x}_2$ and $\mathbf{x}_3$ are already present in the model;
- $SSR(\mathbf{x}_3, \mathbf{x}_4 | \mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: Extra sum of squares of $\mathbf{x}_3$ and $\mathbf{x}_4$ given that $\mathbf{x}_2$ is already present in the model.

Extra sum of squares

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: Extra sum of squares of $\mathbf{x}_4$ given that $\mathbf{x}_2$ and $\mathbf{x}_3$ are already present in the model;
- $SSR(\mathbf{x}_3, \mathbf{x}_4 | \mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: Extra sum of squares of $\mathbf{x}_3$ and $\mathbf{x}_4$ given that $\mathbf{x}_2$ is already present in the model.

Extra sum of squares

The decomposition of the regression sum of squares is given as follows,



Extra sum of squares

The decomposition of the regression sum of squares is given as follows,

- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3|\mathbf{x}_2) + SSR(\mathbf{x}_4|\mathbf{x}_2, \mathbf{x}_3);$

Extra sum of squares

The decomposition of the regression sum of squares is given as follows,

- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3|\mathbf{x}_2) + SSR(\mathbf{x}_4|\mathbf{x}_2, \mathbf{x}_3);$
- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_3) + SSR(\mathbf{x}_4|\mathbf{x}_3) + SSR(\mathbf{x}_2|\mathbf{x}_3, \mathbf{x}_4);$



Extra sum of squares

The decomposition of the regression sum of squares is given as follows,

- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3|\mathbf{x}_2) + SSR(\mathbf{x}_4|\mathbf{x}_2, \mathbf{x}_3);$
- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_3) + SSR(\mathbf{x}_4|\mathbf{x}_3) + SSR(\mathbf{x}_2|\mathbf{x}_3, \mathbf{x}_4);$
- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3, \mathbf{x}_4|\mathbf{x}_2).$



Extra sum of squares

The decomposition of the regression sum of squares is given as follows,

- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3|\mathbf{x}_2) + SSR(\mathbf{x}_4|\mathbf{x}_2, \mathbf{x}_3);$
- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_3) + SSR(\mathbf{x}_4|\mathbf{x}_3) + SSR(\mathbf{x}_2|\mathbf{x}_3, \mathbf{x}_4);$
- $SSR(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_2) + SSR(\mathbf{x}_3, \mathbf{x}_4|\mathbf{x}_2).$



Degrees of freedom of the extra sum of squares

The degrees of freedom of the extra sum of squares are equal to the number of variables associated with it, that is, the number of variables before the “|” in its notation.

Degrees of freedom of the extra sum of squares

Examples:

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: 1 degree of freedom;

Degrees of freedom of the extra sum of squares

Examples:

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: 1 degree of freedom;
- $SSR(\mathbf{x}_3, \mathbf{x}_4 | \mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_3 | \mathbf{x}_2) + SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3)$
2 degrees of freedom.



Degrees of freedom of the extra sum of squares

Examples:

- $SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3) = RSS(\mathbf{x}_2, \mathbf{x}_3) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$: 1 degree of freedom;
- $SSR(\mathbf{x}_3, \mathbf{x}_4 | \mathbf{x}_2) = RSS(\mathbf{x}_2) - RSS(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = SSR(\mathbf{x}_3 | \mathbf{x}_2) + SSR(\mathbf{x}_4 | \mathbf{x}_2, \mathbf{x}_3)$
2 degrees of freedom.

Extra sum of squares

Remark

The ANOVA table can be constructed by decomposing the residual sum of squares into several extra sums of squares.

Extra sum of squares

Remark

The ANOVA table can be constructed by decomposing the residual sum of squares into several extra sums of squares.

Extra sum of squares

The extra sum of squares can be used to assess whether the regression coefficients are significant.



Extra sum of squares

For example, if the objective is to test a single parameter β_m , $m = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = 0 \\ \mathcal{A}: \beta_m \neq 0 \end{cases}.$$

Extra sum of squares

For example, if the objective is to test a single parameter β_m , $m = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = 0 \\ \mathcal{A}: \beta_m \neq 0 \end{cases}.$$

The test statistic is given by:



Extra sum of squares

For example, if the objective is to test a single parameter β_m , $m = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = 0 \\ \mathcal{A}: \beta_m \neq 0 \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}(\mathbf{x}_m | \mathbf{x}_2, \dots, \mathbf{x}_{m-1}, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p)}{\text{MSE}}.$$



Extra sum of squares

For example, if the objective is to test a single parameter β_m , $m = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = 0 \\ \mathcal{A}: \beta_m \neq 0 \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}(\mathbf{x}_m | \mathbf{x}_2, \dots, \mathbf{x}_{m-1}, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p)}{\text{MSE}}.$$



Hypothesis testing

Under $\mathcal{H}$, $F_c \sim F(1, n - p)$. Thus, we reject $\mathcal{H}$ if $F_c > F(1 - \alpha; 1, n - p)$, that is, the null hypothesis is rejected, for a given nominal level α , if the test statistic is greater than the $100(1 - \alpha)\%$ quantile of an $F(1, n - p)$ distribution.



Extra sum of squares

Now, if the objective is to test two parameters β_m and $\beta_{m'}$, $m, m' = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H} : \beta_m = \beta_{m'} = 0 \\ \mathcal{A} : \text{At least one equality is not satisfied} \end{cases} .$$

Extra sum of squares

Now, if the objective is to test two parameters β_m and $\beta_{m'}$, $m, m' = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = \beta_{m'} = 0 \\ \mathcal{A}: \text{At least one equality is not satisfied} \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}(\mathbf{x}_m, \mathbf{x}_{m'} | -\mathbf{x}_m, -\mathbf{x}_{m'})}{\text{MSE}},$$

where “ $-\mathbf{x}_m, -\mathbf{x}_{m'}$ ” denotes all predictor variables except $\mathbf{x}_m$ and $\mathbf{x}_{m'}$.



Extra sum of squares

Now, if the objective is to test two parameters β_m and $\beta_{m'}$, $m, m' = 2, 3, \dots, p$, the hypothesis test using the extra sum of squares is as follows:

$$\begin{cases} \mathcal{H}: \beta_m = \beta_{m'} = 0 \\ \mathcal{A}: \text{At least one equality is not satisfied} \end{cases}.$$

The test statistic is given by:

$$F_c = \frac{\text{MSR}(\mathbf{x}_m, \mathbf{x}_{m'} | -\mathbf{x}_m, -\mathbf{x}_{m'})}{\text{MSE}},$$

where “ $-\mathbf{x}_m, -\mathbf{x}_{m'}$ ” denotes all predictor variables except $\mathbf{x}_m$ and $\mathbf{x}_{m'}$.



Hypothesis testing

Under $\mathcal{H}$, $F_c \sim F(2, n - p)$. Thus, we reject $\mathcal{H}$ if $F_c > F(1 - \alpha; 2, n - p)$, that is, the null hypothesis is rejected, for a given nominal level α , if the test statistic is greater than the $100(1 - \alpha)\%$ quantile of an $F(2, n - p)$ distribution.



Partial coefficient of determination

It measures the marginal relative reduction in the variability of Y when we include $\mathbf{x}_m$ in a model that already contains $\mathbf{x}_2, \dots, \mathbf{x}_{m-1}, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p$. It is calculated as follows:

$$R^2_{Y|\mathbf{x}_m | - \mathbf{x}_m} = \frac{\text{SSR}(\mathbf{x}_m | - \mathbf{x}_m)}{\text{RSS}}.$$

Partial coefficient of determination

It measures the marginal relative reduction in the variability of Y when we include $\mathbf{x}_m$ in a model that already contains $\mathbf{x}_2, \dots, \mathbf{x}_{m-1}, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p$. It is calculated as follows:

$$R^2_{Y|\mathbf{x}_m|\mathbf{x}_m} = \frac{\text{SSR}(\mathbf{x}_m | - \mathbf{x}_m)}{\text{RSS}}.$$

Thus, like R^2 , the partial coefficient of determination is also greater than or equal to 0 and less than or equal to 1.



Partial coefficient of determination

It measures the marginal relative reduction in the variability of Y when we include $\mathbf{x}_m$ in a model that already contains $\mathbf{x}_2, \dots, \mathbf{x}_{m-1}, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p$. It is calculated as follows:

$$R^2_{Y|\mathbf{x}_m|\mathbf{x}_m} = \frac{\text{SSR}(\mathbf{x}_m|\mathbf{x}_m)}{\text{RSS}}.$$

Thus, like R^2 , the partial coefficient of determination is also greater than or equal to 0 and less than or equal to 1.



Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application**
- 6 Simulation study
- 7 References



Application

Westwood Company data set (Neter et al., 1983, p. 36). After fitting the model,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

where Y_ℓ : man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \dots, 10$.

Application

Westwood Company data set (Neter et al., 1983, p. 36). After fitting the model,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

where Y_ℓ : man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \dots, 10$.

ANOVA

Using the expressions given in (4) and (5) and substituting the results into Table 1, the ANOVA table is as follows:

Table 2: Analysis of variance.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_c
Regression	13,600	1	13,600	1,813.3
Residuals	60	8	7.5	
Total	13,660	9		



ANOVA

Using the expressions given in (4) and (5) and substituting the results into Table 1, the ANOVA table is as follows:

Table 2: Analysis of variance.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_c
Regression	13,600	1	13,600	1,813.3
Residuals	60	8	7.5	
Total	13,660	9		



ANOVA

For this data set, the hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H} : \beta_2 = 0 \\ \mathcal{A} : \beta_2 \neq 0 \end{cases}.$$

Since $F_c = 1,1813.3 > 5.3 = F(0.95; 1, 8)$,



ANOVA

For this data set, the hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H} : \beta_2 = 0 \\ \mathcal{A} : \beta_2 \neq 0 \end{cases} .$$

Since $F_c = 1,1813.3 > 5.3 = F(0.95; 1, 8)$, we reject $\mathcal{H}$, that is, at a 5% significance level, the regression model is statistically significant.



ANOVA

For this data set, the hypothesis test using the analysis of variance approach is as follows:

$$\begin{cases} \mathcal{H} : \beta_2 = 0 \\ \mathcal{A} : \beta_2 \neq 0 \end{cases} .$$

Since $F_c = 1,1813.3 > 5.3 = F(0.95; 1, 8)$, we reject $\mathcal{H}$, that is, at a 5% significance level, the regression model is statistically significant.



ANOVA

From (6) and (7), the coefficients of determination are given by

$$R^2 = 0.996 \text{ and } R_a^2 = 0.995.$$

ANOVA

From (6) and (7), the coefficients of determination are given by

$$R^2 = 0.996 \text{ and } R_a^2 = 0.995.$$

The proportion of the variability of the response variable that is explained by the regression model is approximately 99%.



ANOVA

From (6) and (7), the coefficients of determination are given by

$$R^2 = 0.996 \text{ and } R_a^2 = 0.995.$$

The proportion of the variability of the response variable that is explained by the regression model is approximately 99%.



Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study**
- 7 References



Simulation study

Suppose the following LRM,

$$Y_{\ell} = 1 + 0x_{\ell 2} + \varepsilon_{\ell},$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_{\ell} \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \dots, n$. We will assume that $x_{\ell 2} \sim$ uniform on the interval $(0, 1)$. The covariate $x_{\ell 2}$, for each n ($n = 10, 30, \dots, 110$), is kept fixed across the Monte Carlo replicates. The simulations were performed using R (R Core Team, 2021).



Simulation study

Suppose the following LRM,

$$Y_{\ell} = 1 + 0x_{\ell 2} + \varepsilon_{\ell},$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_{\ell} \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \dots, n$. We will assume that $x_{\ell 2} \sim$ uniform on the interval $(0, 1)$. The covariate $x_{\ell 2}$, for each n ($n = 10, 30, \dots, 110$), is kept fixed across the Monte Carlo replicates. The simulations were performed using R (R Core Team, 2021).



Simulation study

We will conduct a Monte Carlo study with 10,000 replicates, in which we will evaluate

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

Simulation study

We will conduct a Monte Carlo study with 10,000 replicates, in which we will evaluate

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

through the test size,



Simulation study

We will conduct a Monte Carlo study with 10,000 replicates, in which we will evaluate

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

through the test size, i.e.,

$$\alpha = \mathbb{P}(\text{Reject } \mathcal{H} | \mathcal{H} \text{ is true}).$$



Simulation study

We will conduct a Monte Carlo study with 10,000 replicates, in which we will evaluate

$$\begin{cases} \mathcal{H} : \beta_2 = 0 \\ \mathcal{A} : \beta_2 \neq 0 \end{cases},$$

through the test size, i.e.,

$$\alpha = \mathbb{P}(\text{Reject } \mathcal{H} | \mathcal{H} \text{ is true}).$$



Simulation study

Assuming that $\mathcal{H}$ is true, the empirical size of the ANOVA test is calculated as the proportion of times that the test statistic (F_c) was greater than the value F_0 in the 10,000 Monte Carlo replicates, for $\alpha = 10\%$, 5% , and 1% .



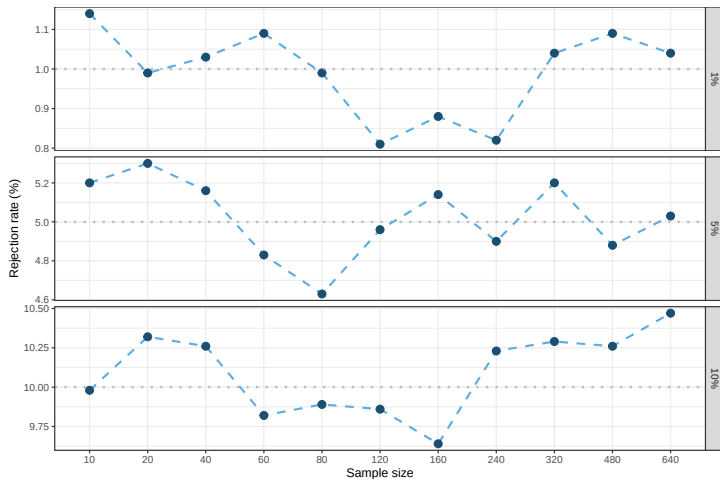


Figure 5: Test size.

Outline

- 1 Introduction
- 2 Partition of sums of squares
- 3 Coefficient of determination
- 4 Extra sum of squares
- 5 Application
- 6 Simulation study
- 7 References



References

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012), *Introduction to linear regression analysis*, 5th edn, Wiley, New York.

Neter, J., Wasserman, W. and Kutner, M. H. (1983), *Applied linear regression models*, Richard D. Irwin Inc, Homewood, Illinois.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: <https://www.R-project.org/>



Thank you!

✉ tiago.magalhaes@ufjf.br

📄 ufjf.br/tiago_magalhaes

🌐 Department of Statistics, Room 319

