Statistics II
Lesson 5. Regression analysis (second part)

Year 2010/11

Lesson 5. Regression analysis (second part)

Contents

- Diagnosis: Residual analysis
- The ANOVA (ANalysis Of VAriance) decomposition
- Nonlinear relationships and linearizing transformations
- A matrix treatment of the linear regression model
- Introduction to multiple linear regression

Lesson 5. Regression analysis (second part)

Bibliography references

- Newbold, P. “Estadística para los negocios y la economía” (1997)
  - Chapters 12, 13 and 14
- Peña, D. “Regresión y Análisis de Experimentos” (2002)
  - Chapters 5 and 6

Regression diagnostics

Diagnostics

- The theoretical assumptions for the simple linear regression model with one response variable Y and one explicative variable x are:
  - Linearity: $y_i = \beta_0 + \beta_1 x_i + u_i$, for $i = 1, \ldots, n$
  - Homogeneity: $E[u_i] = 0$, for $i = 1, \ldots, n$
  - Homoscedasticity: $\text{Var}[u_i] = \sigma^2$, for $i = 1, \ldots, n$
  - Independence: u_i and u_j are independent for $i \neq j$
  - Normality: $u_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, n$
- We study how to apply diagnostic procedures to test if these assumptions are appropriate for the available data (x_i, y_i)
  - Based on the analysis of the residuals $e_i = y_i - \hat{y}_i$

Diagnostics: scatterplots

Scatterplots

- The simplest diagnostic procedure is based on the visual examination of the scatterplot for (x_i, y_i)
- Often this simple but powerful method reveals patterns suggesting whether the theoretical model might be appropriate or not
- We illustrate its application on a classical example. Consider the four following datasets

Diagnostics: scatterplots

The Anscombe datasets

TABLE 3-10
Four Data Sets

Data Set 1 X	Data Set 1 Y	Data Set 2 X	Data Set 2 Y
10.0	8.04	10.0	9.14
8.0	6.95	8.0	8.14
13.0	7.58	13.0	8.74
9.0	8.81	9.0	8.77
11.0	8.33	11.0	9.26
14.0	9.96	14.0	8.10
6.0	7.24	6.0	6.13
4.0	4.26	4.0	3.10
12.0	10.84	12.0	9.13
7.0	4.82	7.0	7.26
5.0	5.68	5.0	4.74

Data Set 3 X	Data Set 3 Y	Data Set 4 X	Data Set 4 Y
10.0	7.46	8.0	6.58
8.0	6.77	8.0	5.76
13.0	12.74	8.0	7.71
9.0	7.11	8.0	8.84
11.0	7.81	8.0	8.47
14.0	8.84	8.0	7.04
6.0	6.08	8.0	5.25
4.0	5.39	19.0	12.50
12.0	8.15	8.0	5.56
7.0	6.42	8.0	7.91
5.0	5.73	8.0	6.89

SOURCE: F. J. Anscombe, *op. cit.*

Diagnostics: scatterplots

The Anscombe example

- The estimated regression model for each of the four previous datasets is the same
  - $y_i = 3{,}0 + 0{,}5x_i$
  - $n = 11, \quad \bar{x} = 9{,}0, \quad \bar{y} = 7{,}5, \quad r_{xy} = 0{,}817$
- The estimated standard error of the estimator $\hat{\beta}_1$,

$$\sqrt{\frac{s_R^2}{(n-1)s_x^2}}$$

  takes the value 0,118 in all four cases. The corresponding T statistic takes the value $T = 0{,}5/0{,}118 = 4{,}237$
- But the corresponding scatterplots show that the four datasets are quite different. Which conclusions could we reach from these diagrams?

Diagnostics: scatterplots

Anscombe data scatterplots

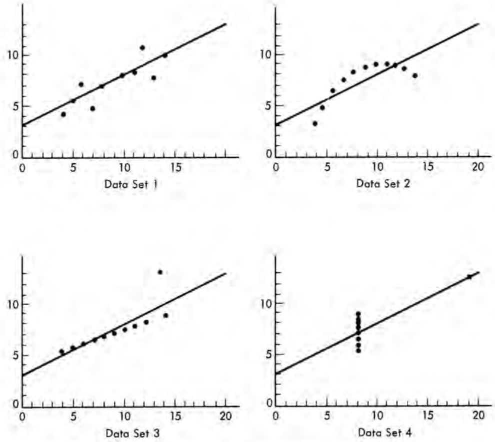


FIGURE 3-29 Scatterplots for the four data sets of Table 3-10
SOURCE: F. J. Anscombe, *op cit.*

Residual analysis

Further analysis of the residuals

- If the observation of the scatterplot is not sufficient to reject the model, a further step would be to use diagnosis methods based on the analysis of the residuals $e_i = y_i - \hat{y}_i$
- This analysis starts by standarizing the residuals, that is, dividing them by the residuals (quasi-)standard deviation s_R. The resulting quantities are known as standarized residuals:

$$\frac{e_i}{s_R}$$

- Under the assumptions of the linear regression model, the standarized residuals are approximately independent standard normal random variables
- A plot of these standarized residuals should show no clear pattern

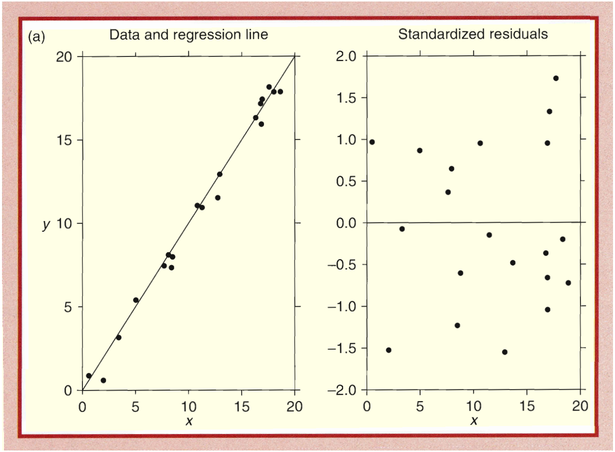
Residual analysis

Residual plots

- Several types of residual plots can be constructed. The most common ones are:
  - Plot of standardized residuals vs. x
  - Plot of standardized residuals vs. $\hat{y}$ (the predicted responses)
- Deviations from the model hypotheses result in patterns on these plots, which should be visually recognizable

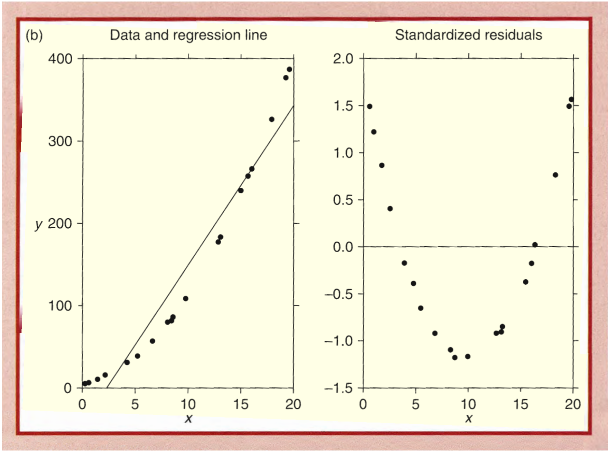
Residual plots examples

Consistency of the theoretical model



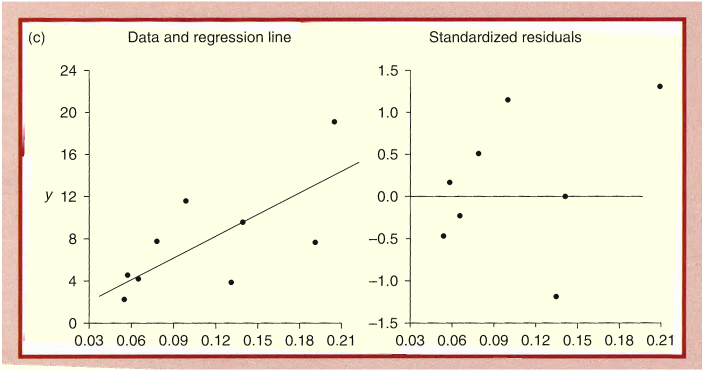
Residual plots examples

Nonlinearity



Residual plots examples

Heteroscedasticity



Residual analysis

Outliers

- In a plot of the regression line we may observe outlier data, that is, data that show significant deviations from the regression line (or from the remaining data)
- The parameter estimators for the regression model, $\hat{\beta}_0$ and $\hat{\beta}_1$, are very sensitive to these outliers
- It is important to identify the outliers and make sure that they really are valid data
- Statgraphics is able to show for example the data that generate "Unusual residuals" or "Influential points"

Residual analysis

Normality of the errors

- One of the theoretical assumptions of the linear regression model is that the errors follow a normal distribution
- This assumption can be checked visually from the analysis of the residuals e_i, using different approaches:
  - By inspection of the frequency histogram for the residuals
  - By inspection of the "Normal Probability Plot" of the residuals (significant deviations of the data from the straight line in the plot correspond to significant departures from the normality assumption)

The ANOVA decomposition

Introduction

- ANOVA: *ANalysis Of VAriance*
- When fitting the simple linear regression model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ to a data set (x_i, y_i) for $i = 1, \ldots, n$, we may identify three sources of variability in the responses
  - variability associated to the model:

    $$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

    where the initials *SS* denote "sum of squares" and *M* refers to the model
  - variability of the residuals:

    $$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  - total variability: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- The ANOVA decomposition states that: $SST = SSM + SSR$

The ANOVA decomposition

The coefficient of determination R^2

- The ANOVA decomposition states that $SST = SSM + SSR$
- Note that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$
- $SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ measures the variation in the responses due to the regression model (explained by the predicted values $\hat{y}_i$)
- Thus, the ratio SSR/SST is the proportion of the variation in the responses that is not explained by the regression model
- The ratio $R^2 = SSM/SST = 1 - SSR/SST$ is the proportion of the variation in the responses that is explained by the regression model. It is known as the coefficient of determination
- The value of the coefficient of determination satisfies $R^2 = r_{xy}^2$ (the squared correlation coefficient)
- For example, if $R^2 = 0{,}85$ the variable x explains 85 % of the variation in the response variable y

The ANOVA decomposition

ANOVA table

Source of variability	SS	DF	Mean	F ratio
Model	SSM	1	$SSM/1$	SSM/s_R^2
Residuals/errors	SSR	$n-2$	$SSR/(n-2) = s_R^2$	
Total	SST	$n-1$		

The ANOVA decomposition

ANOVA hypothesis testing

- Hypothesis test, $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$
- Consider the ratio

$$F = \frac{SSM/1}{SSR/(n-2)} = \frac{SSM}{s_R^2}$$

- Under H_0, F follows an $F_{1,n-2}$ distribution
- Test at a significance level α: reject H_0 if $F > F_{1,n-2;\alpha}$
- How does this result relate to the test based on the Student-t we saw in Lesson 4? They are equivalent

The ANOVA decomposition

Statgraphics output

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Precio en ptas.
Independent variable: Produccion en kg.
-----------------------------------------------------------------------------
                                Standard          T
Parameter        Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept         74,1151        8,73577        8,4841          0,0000
Slope            -1,35368         0,3002      -4,50924          0,0020
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     528,475      1      528,475      20,33       0,0020
Residual                  207,925      8      25,9906
-----------------------------------------------------------------------------
Total (Corr.)               736,4      9

Correlation Coefficient = -0,84714
R-squared = 71,7647 percent
Standard Error of Est. = 5,0981
```

25,9906 → s_R^2

Nonlinear relationships and linearizing transformations

Introduction

- Consider the case when the deterministic part $f(x_i; a, b)$ of the response in the model

$$y_i = f(x_i; a, b) + u_i, \quad i = 1, \ldots, n$$

  is a nonlinear function of x that depends on two parameters a and b (for example, $f(x; a, b) = ab^x$)
- In some cases we may apply transformations to the data to linearize them. We are then able to apply the linear regression procedure
- From the original data (x_i, y_i) we obtain the transformed data (x_i', y_i')
- The parameters β_0 and β_1 corresponding to the linear relation between x_i' and y_i' are transformations of the parameters a and b

Nonlinear relationships and linearizing transformations

Linearizing transformations

- Examples of linearizing transformations:
  - If $y = f(x; a, b) = ax^b$ then $\log y = \log a + b \log x$. We have
    - $y' = \log y$, $x' = \log x$, $\beta_0 = \log a$, $\beta_1 = b$
  - If $y = f(x; a, b) = ab^x$ then $\log y = \log a + (\log b)x$. We have
    - $y' = \log y$, $x' = x$, $\beta_0 = \log a$, $\beta_1 = \log b$
  - If $y = f(x; a, b) = 1/(a + bx)$ then $1/y = a + bx$. We have
    - $y' = 1/y$, $x' = x$, $\beta_0 = a$, $\beta_1 = b$
  - If $y = f(x; a, b) = \log(ax^b)$ then $y = \log a + b(\log x)$. We have
    - $y' = y$, $x' = \log x$, $\beta_0 = \log a$, $\beta_1 = b$

A matrix treatment of linear regression

Introduction

- Remember the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n$$

- If we write one equation for each one of the observations, we have

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + u_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + u_2 \\ \vdots & \quad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + u_n \end{aligned}$$

A matrix treatment of linear regression

The model in matrix form

- We can write the preceding equations in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

- And splitting the parameters β from the variables x_i,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

A matrix treatment of linear regression

The regression model in matrix form

- We can write the preceding matrix relationship

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- $\mathbf{y}$: response vector; $\mathbf{X}$: explanatory variables matrix (or experimental design matrix); $\boldsymbol{\beta}$: vector of parameters; $\mathbf{u}$: error vector

The regression model in matrix form

Covariance matrix for the errors

- We denote as $\text{Cov}(\mathbf{u})$ the $n \times n$ matrix of covariances for the errors. Its (i, j)-th element is given by $\text{cov}(u_i, u_j) = 0$ if $i \neq j$ and $\text{cov}(u_i, u_i) = \text{Var}(u_i) = \sigma^2$
- Thus, $\text{Cov}(\mathbf{u})$ is the identity matrix $\mathbf{I}_{n \times n}$ multiplied by σ^2:

$$\text{Cov}(\mathbf{u}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

The regression model in matrix form

Least-squares estimation

- The least-squares vector parameter estimate $\hat{\beta}$ is the unique solution of the 2×2 matrix equation (check the dimensions)

$$(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y},$$

that is,

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

- The vector $\hat{\mathbf{y}} = (\hat{y}_i)$ of response estimates is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

and the residual vector is defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

The multiple linear regression model

Introduction

- Use of the simple linear regression model: predict the value of a response y from the value of an explanatory variable x
- In many applications we wish to predict the response y from the values of several explanatory variables $x_1, \ldots, x_k$
- For example:
  - forecast the value of a house as a function of its size, location, layout and number of bathrooms
  - forecast the size of a parliament as a function of the population, its rate of growth, the number of political parties with parliamentary representation, etc.

The multiple linear regression model

The model

- Use of the multiple linear regression model: predict a response y from several explanatory variables $x_1, \ldots, x_k$
- If we have n observations, for $i = 1, \ldots, n$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

- We assume that the error variables u_i are independent random variables following a $N(0, \sigma^2)$ distribution

The multiple linear regression model

The least-squares fit

- We have n observations, and for $i = 1, \ldots, n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

- We wish to fit to the data $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$ a hyperplane of the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

- The residual for observation i is defined as $e_i = y_i - \hat{y}_i$
- We obtain the parameter estimates $\hat{\beta}_j$ as the values that minimize the sum of the squares of the residuals

The multiple linear regression model

The model in matrix form

- We can write the model as a matrix relationship,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

  and in compact form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- $\mathbf{y}$: response vector; $\mathbf{X}$: explanatory variables matrix (or experimental design matrix); $\boldsymbol{\beta}$: vector of parameters; $\mathbf{u}$: error vector

The multiple linear regression model

Least-squares estimation

- The least-squares vector parameter estimate $\hat{\beta}$ is the unique solution of the $(k+1) \times (k+1)$ matrix equation (check the dimensions)

$$(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y},$$

and as in the $k = 1$ case (simple linear regression) we have

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

- The vector $\hat{\mathbf{y}} = (\hat{y}_i)$ of response estimates is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

and the residual vector is defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

The multiple linear regression model

Variance estimation

- For the multiple linear regression model, an estimator for the error variance σ^2 is the residual (quasi-)variance,

$$s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1},$$

  and this estimator is unbiased
  - Note that for the simple linear regression case we had $n - 2$ in the denominator

The multiple linear regression model

The sampling distribution of $\hat{\boldsymbol{\beta}}$

- Under the model assumptions, the least-squares estimator $\hat{\boldsymbol{\beta}}$ for the parameter vector $\boldsymbol{\beta}$ follows a multivariate normal distribution
- $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (it is an unbiased estimator)
- The covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- We estimate $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ using $s_R^2(\mathbf{X}^T\mathbf{X})^{-1}$
- The estimate of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ provides estimates $s^2(\hat{\beta}_j)$ for the variance $\mathrm{Var}(\hat{\beta}_j)$. $s^2(\hat{\beta}_j)$ is the standard error of the estimator $\hat{\beta}_j$
- If we standardize $\hat{\beta}_j$ we have

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k-1} \quad \text{(the Student-t distribution)}$$

The multiple linear regression model

Inference on the parameters $\hat{\beta}_j$

- Confidence interval for β_j at a confidence level $1-\alpha$

$$\hat{\beta}_j \pm t_{n-k-1;\alpha/2}\, s(\hat{\beta}_j)$$

- Hypothesis testing for $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ at a confidence level α
  - Reject H_0 if $|T| > t_{n-k-1;\alpha/2}$, where

$$T = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

    is the test statistic

The ANOVA decomposition

The multivariate case

- ANOVA: *ANalysis Of VAriance*
- When fitting the multiple linear regression model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ to a data set $(x_{i1}, \ldots, x_{ik}, y_i)$ for $i = 1, \ldots, n$, we may identify three sources of variability in the responses
  - variability associated to the model:

    $$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

    where the initials SS denote "sum of squares" and M refers to the model
  - variability of the residuals:

    $$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  - total variability: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- The ANOVA decomposition states that: $SST = SSM + SSR$

The ANOVA decomposition

The coefficient of determination R^2

- The ANOVA decomposition states that $SST = SSM + SSR$
- Note that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$
- $SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ measures the variation in the responses due to the regression model (explained by the predicted values $\hat{y}_i$)
- Thus, the ratio SSR/SST is the proportion of the variation in the responses that is not explained by the regression model
- The ratio $R^2 = SSM/SST = 1 - SSR/SST$ is the proportion of the variation in the responses that is explained by the explanatory variables. It is known as the coefficient of multiple determination
- The value of this coefficient satisfies $R^2 = r_{\hat{y}y}^2$ (the squared correlation coefficient)
- For example, if $R^2 = 0{,}85$ the variables $x_1, \ldots, x_k$ explain 85 % of the variation in the response variable y

The ANOVA decomposition

ANOVA table

Source of variability	SS	DF	Mean	F ratio
Model	SSM	k	SSM/k	$(SSM/k)/s_R^2$
Residuals/errors	SSR	$n-k-1$	$SSR/(n-k-1) = s_R^2$	
Total	SST	$n-1$		

The ANOVA decomposition

ANOVA hypothesis testing

- Hypothesis test, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. $H_1 : \beta_j \neq 0$ for some $j = 1, \ldots, k$
- H_0: the response does not depend on any x_j
- Consider the ratio

$$F = \frac{SSM/k}{SSR/(n-k-1)} = \frac{SSM}{s_R^2}$$

- Under H_0, F follows an $F_{k,n-k-1}$ distribution
- Test at a significance level α: reject H_0 if $F > F_{k,n-k-1;\alpha}$