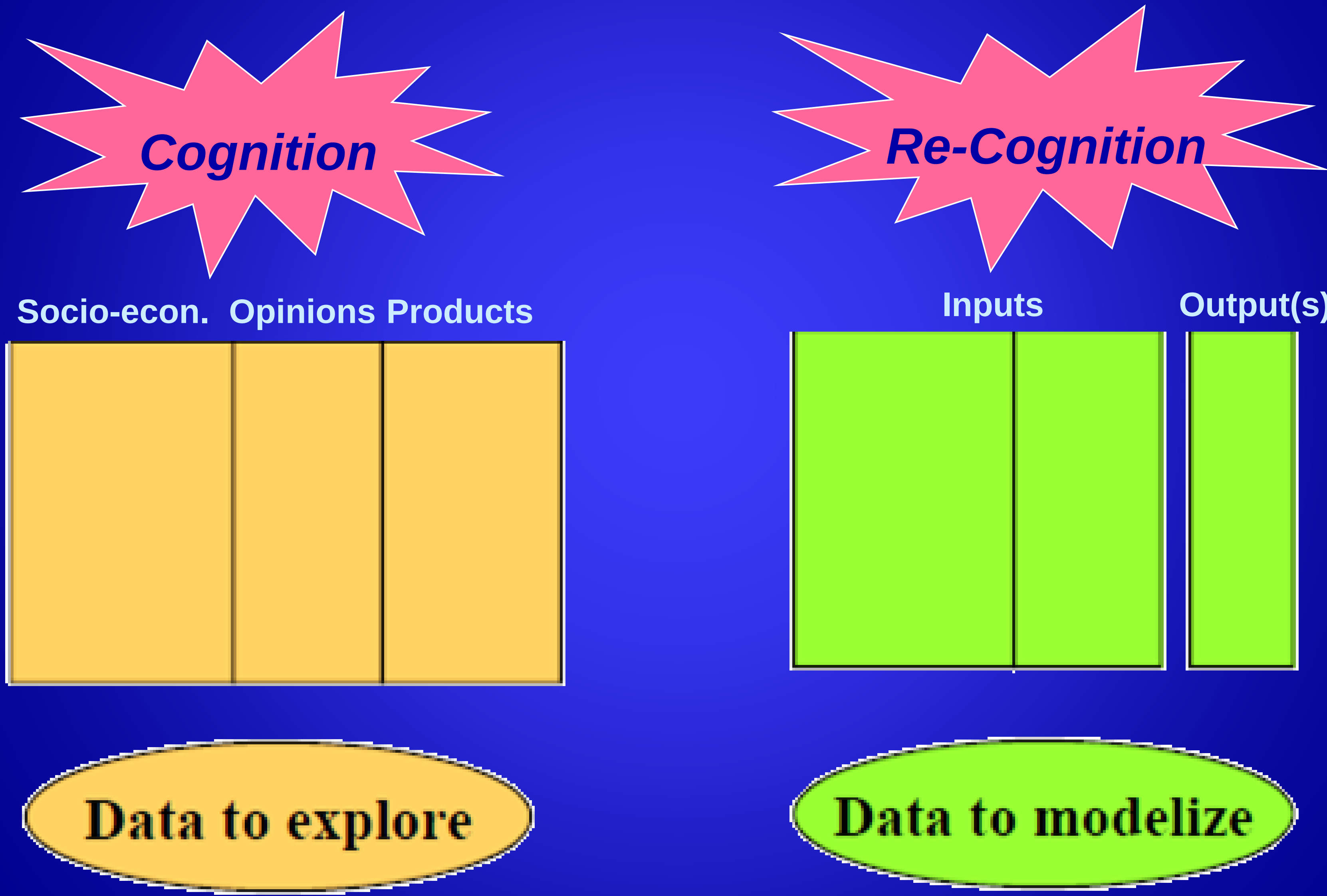
Predictive Methods

K. Gibert[1]

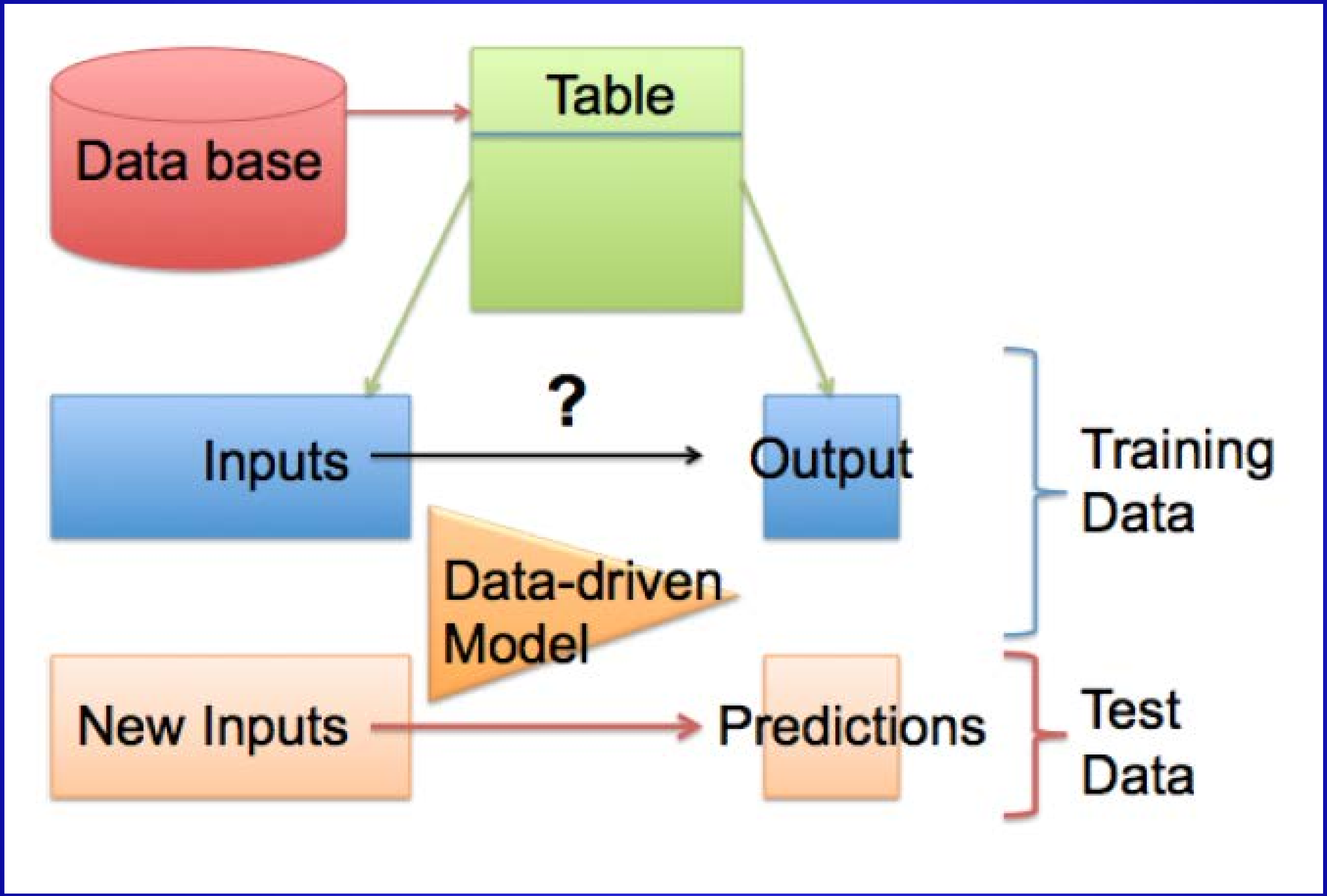
[1]*Department of Statistics and Operation Research*

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

Modelling


Cognition
Re-Cognition
Socio-econ.
Opinions
Products
Inputs
Output(s)
Data to explore
Data to modelize

Supervised Learning



Supervised learning tasks

DM goals [Fayyad et al., 1996]

- **Classification** – labeling a data item into one of several predefined classes (e.g. classify the type of credit client, "good" or "bad", given the status of her/his bank account, credit purpose and amount);
- **Regression** – estimate a real-value (the *dependent variable*) from several (*independent*) attributes (e.g. predict the price of a house based on its number of rooms, age and other characteristics);

- **Classification**: Decision Tree, Random Forest, Classification Rules, Linear Discriminant Analysis, Naive Bayes, Logistic Regression, Neural Networks (MLP, RBF), SVM, ...
- **Regression**: Regression Tree, Random Forest, Multiple Regression, Neural Networks (MLP, RBF), SVM, ...

Statistical Modelling

Data= Fit+Error

- Fit:
  - Structural
  - Law governing the phenomenon
  - Analytic Function

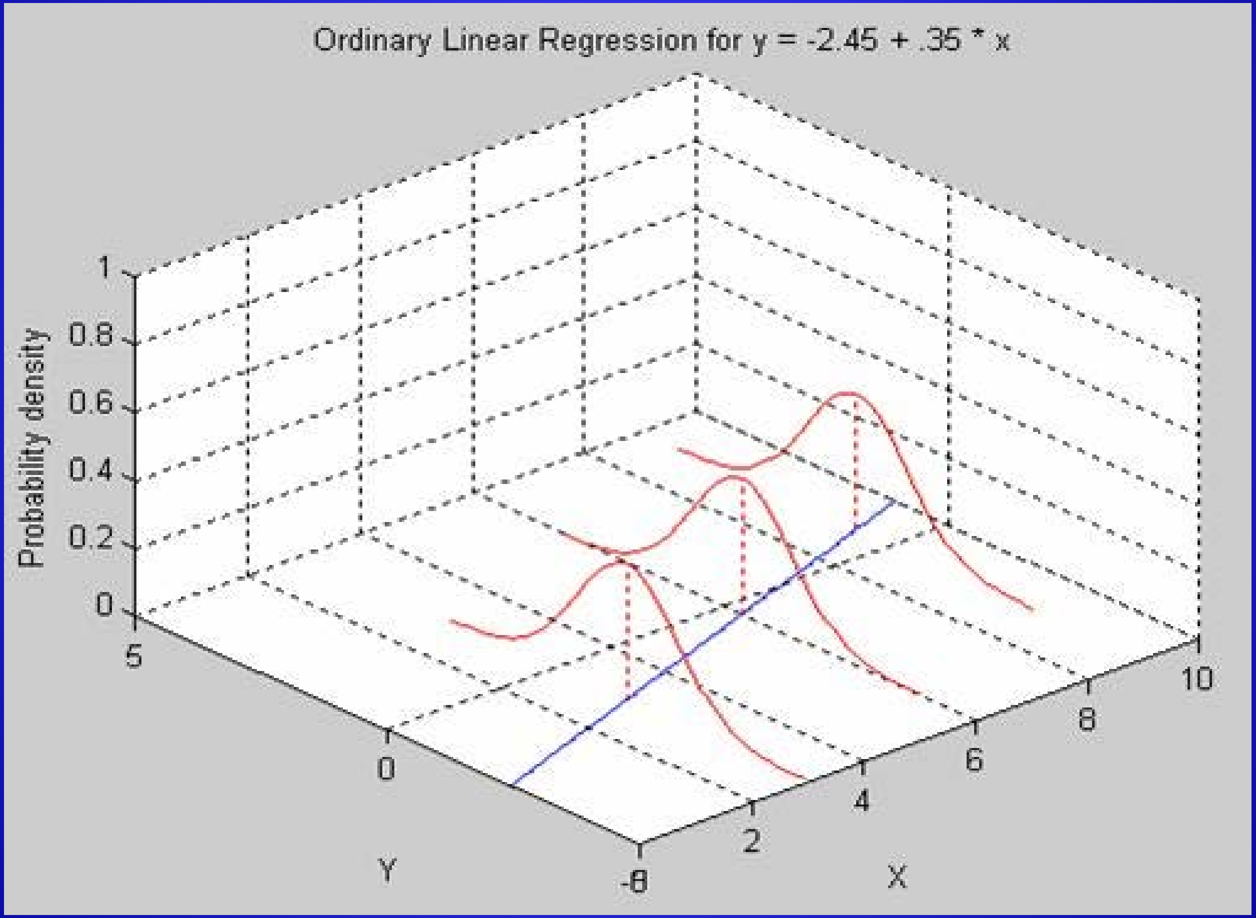
- Error:
  - Random
  - Variability arround Fit (null expectation)
  - Probabilistic model

Statistical models

- Determine the family of fits:
  - Linear
  - Quadratic
  - Exponential
  - .....

- Determine the law of error:
  - Normal
  - Poisson
  - Binomial....

E1 Normally Distributed Error



Linear Multiple regression

- *Fit= linear; Error=Normal and centered*
- *Formalization: I=i:n observations*

  Y: Response variable

  $X_1 \ldots\ldots X_K$ *: ExplanatoryVariables*

  Find $\beta_{0\ldots}\ \beta_K$ such that

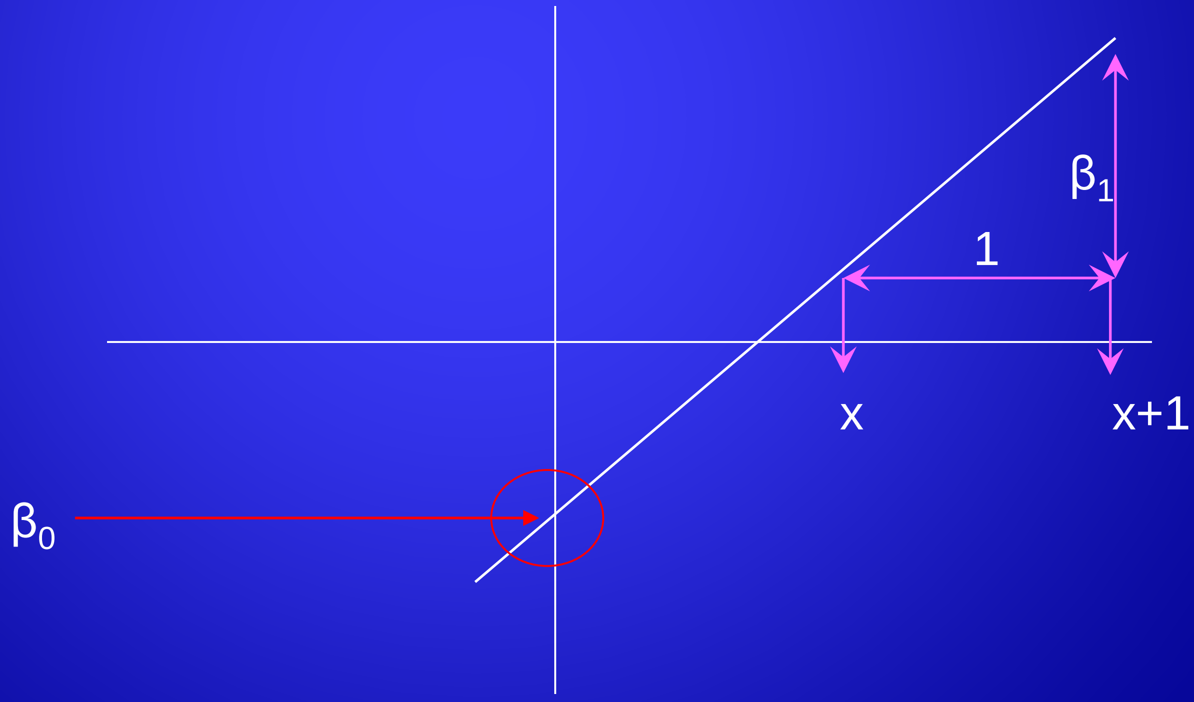
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon$$

- *Assumptions:*
  - *Linearity:* $E(Y \mid X = x) = \mu_{y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K$; $E[\varepsilon] = 0$

    Population regression line
  - *Normality:* $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma_i)$, i=1:n
  - *Homokedasticity:* $Var[\varepsilon_i] = \sigma^2$ forall i
  - *Independence:* $Cov(\varepsilon_i, \varepsilon_j) = 0$ forall i,,j

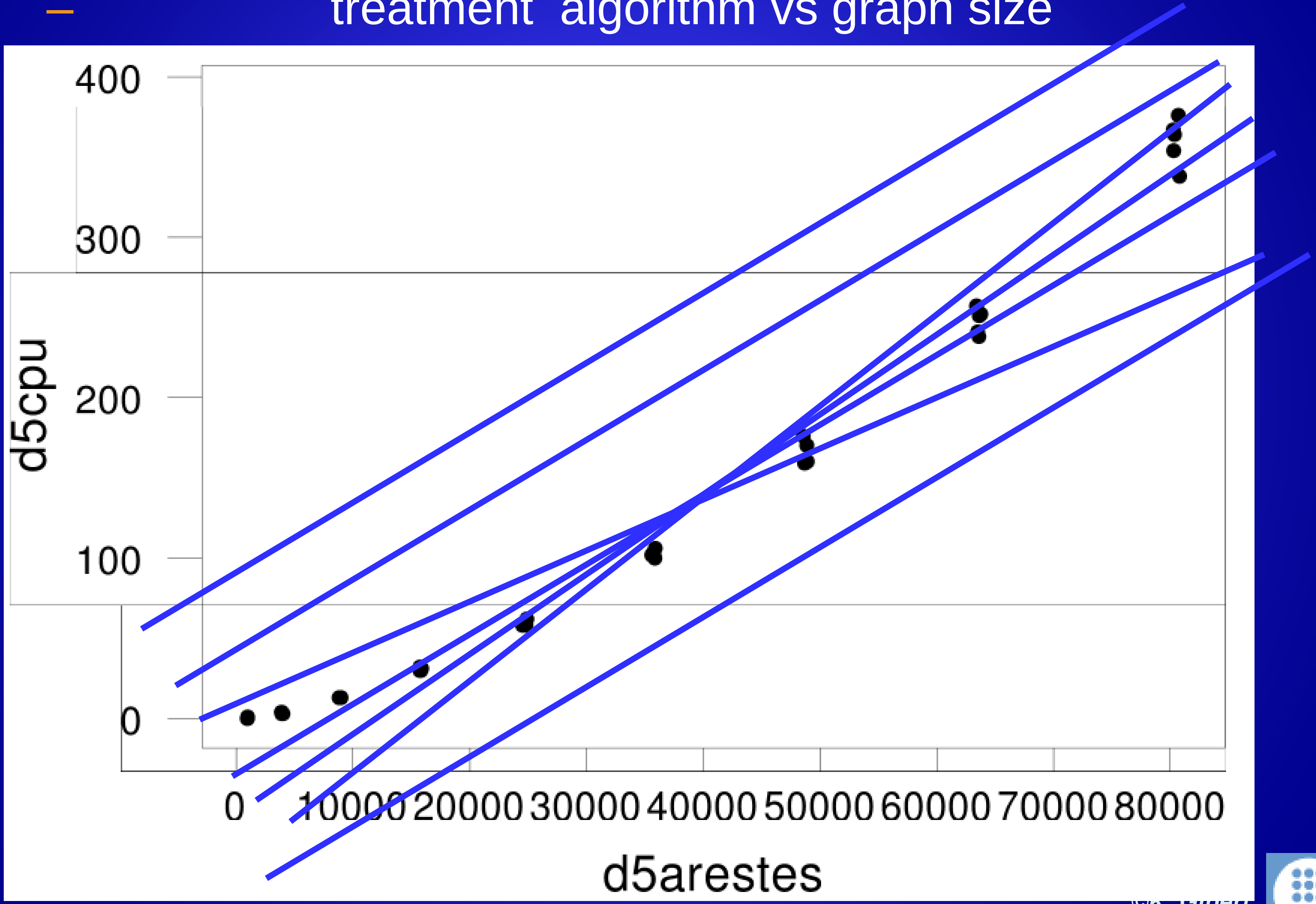
What is "Linear"?

- Remember this:
- $Y = \beta_0 + \beta_1 X$ *(simple linear regression, one X)*

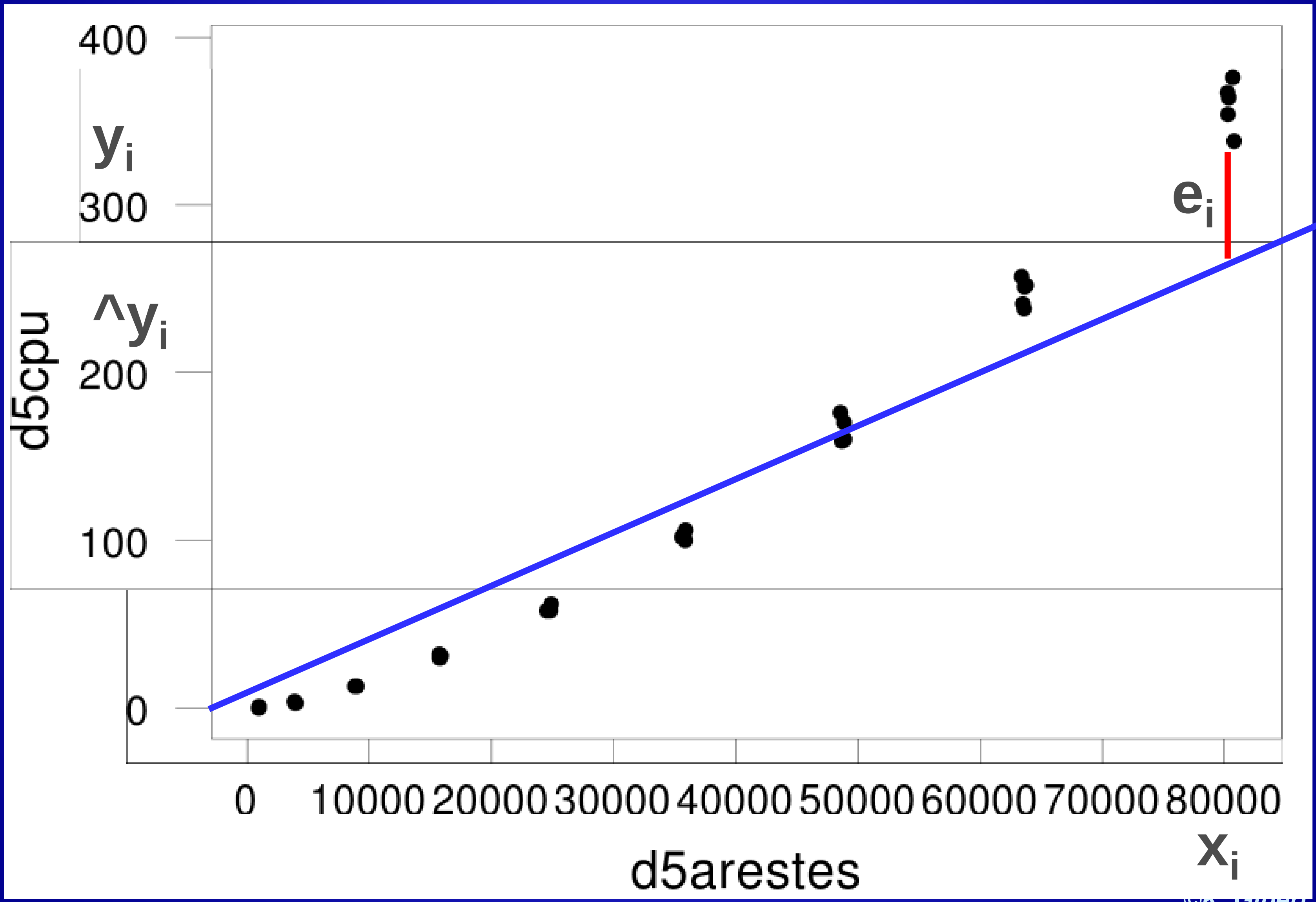


– Real case: Experimental CPU time of a graph

– treatment algorithm vs graph size



– Real case: Experimental CPU time of a graph
– treatment algorithm vs graph size



Minimum Least Squares solution

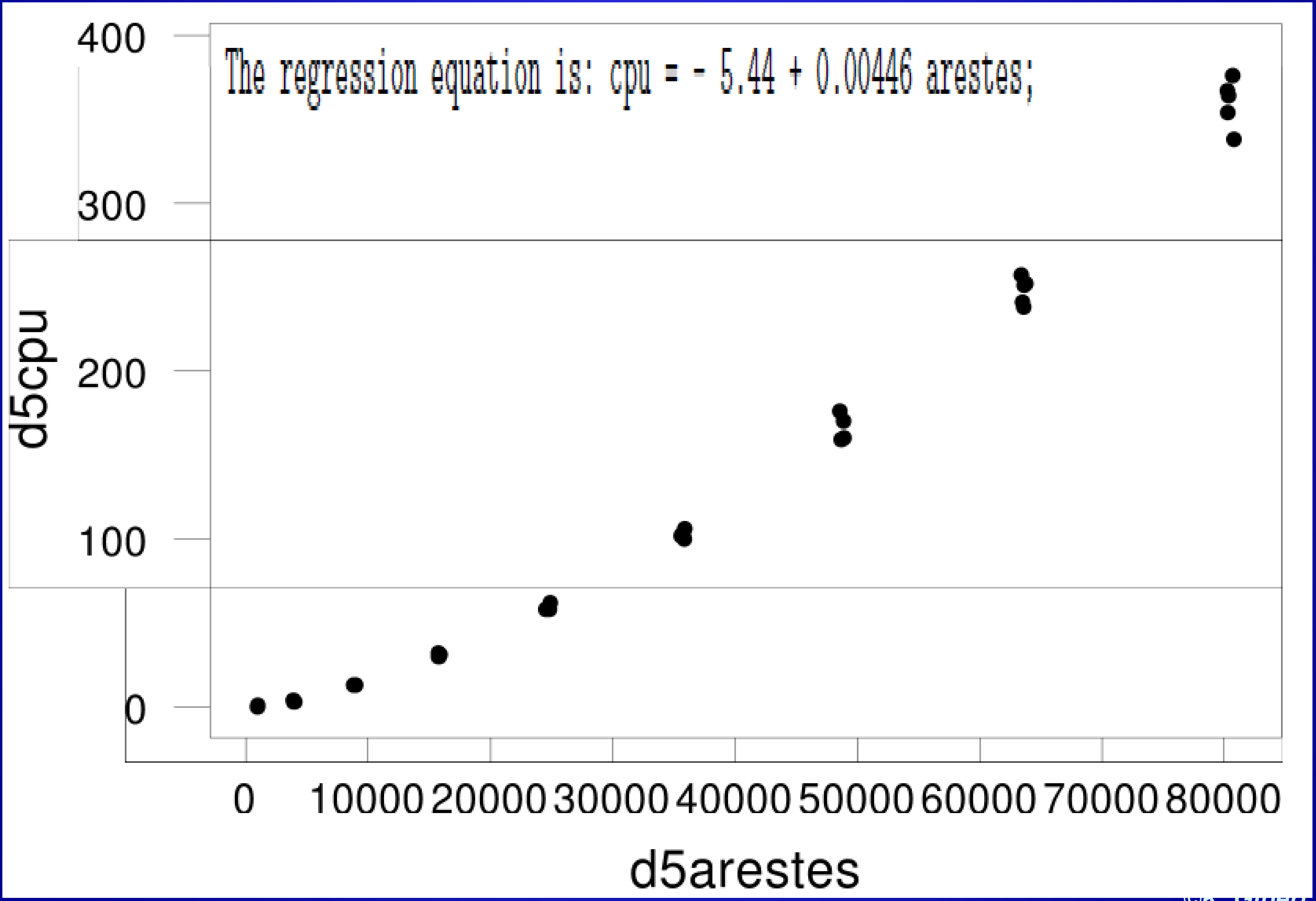
Find $\widehat{\beta}_0, \widehat{\beta}_1$ such that $\min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{\forall i} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

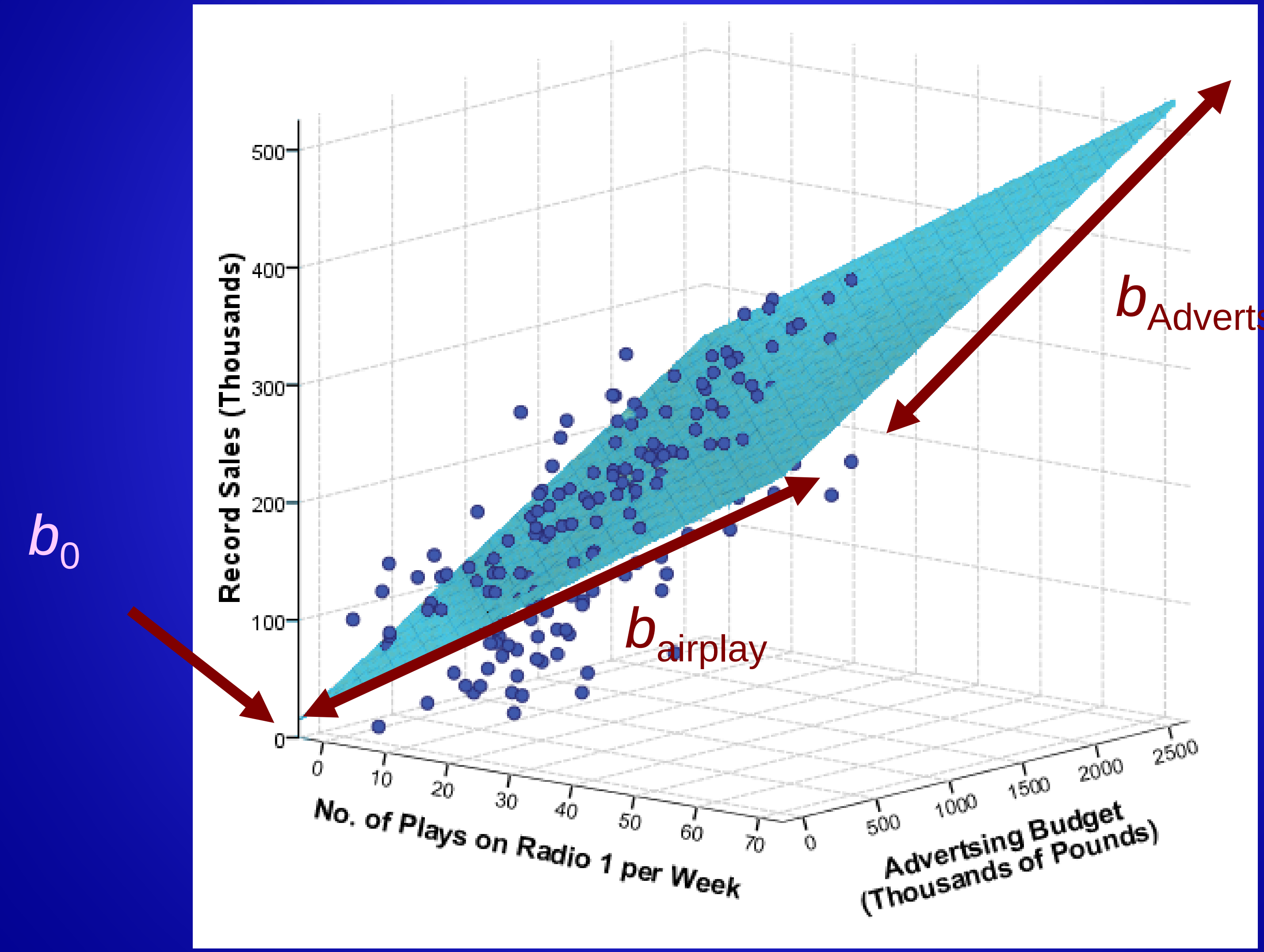
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

– Real case: Experimental CPU time of a graph

– treatment algorithm vs graph size



The Model with Two Predictors



Matricial formulation

Regression fit criterion: $\min_r E\left[\left(y_i - r(x_{i1},\cdots,x_{ip})\right)^2\right]$

$$r(x_{i1},\cdots,x_{ip}) = E\left[y_i \middle| x_{i1},\cdots,x_{ip}\right]$$

$$E\left[y_i \middle| x_{i1},\cdots,x_{ip}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Estimation of coefficients

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i$$

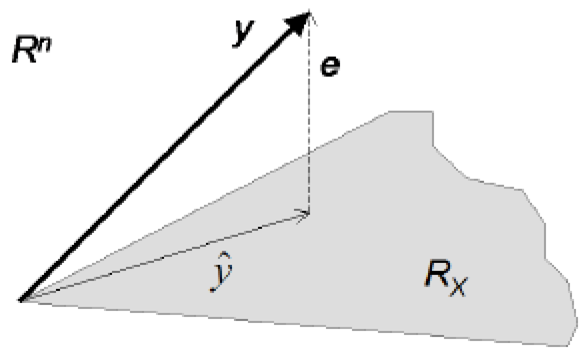
In matrix notation

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1\, x_{11} & \cdots & x_{1p} \\ \vdots \quad \vdots & \ddots & \vdots \\ 1\, x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \equiv \quad y = Xb + e = \hat{y} + e$$

Geometric interpretation

$$y_i = \hat{y}_i + e$$

$$\begin{array}{ccccc} y & & \hat{y} & & e \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} & + & \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \end{array}$$



$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$$

Criterion: $$\min_{b_0,\dots,b_p} \sum_{i=1}^{n} (e_i)^2 = \|e\|^2$$

$$\langle \hat{y}, e \rangle = \langle \hat{y}, y - \hat{y} \rangle = 0$$

$$\hat{y} = Xb, \;\; b'X'y - b'X'Xb = 0$$

$$\boxed{b = (X'X)^{-1} X'y}$$

Validation

- Technical Assumptions
  - normality, linearity, independence, homokedasticity
  - Tools
    - Graphical residuals analysis
    - Influence-point indicators (hi)
- Quality:
  - R2 (determination coeficient): goodness, reliability
  - s-2: noise, precision
  - Both guarantee generalizability (only interpolation)

Quantify Goodnes of model

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (e_i)^2}{n-2}$$

Estimates the variance of residuals

The biggest, the worst the model, more impresice predictions

Non-standardized

Quantify Goodnes of fit

R^2 : proportion of explained variance

SStotal = V(Y) variance of response variable

Decomposition: SStotal = SSexplainedByModel + SSerrror

Dividing all sides by SStotal:

$$R^2 = \frac{SSexplainedByModel}{SSTotal} = 1 - \frac{SSError}{SSTotal}$$

Quantify Goodnes of model

SSTotal= V(Y) = $\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

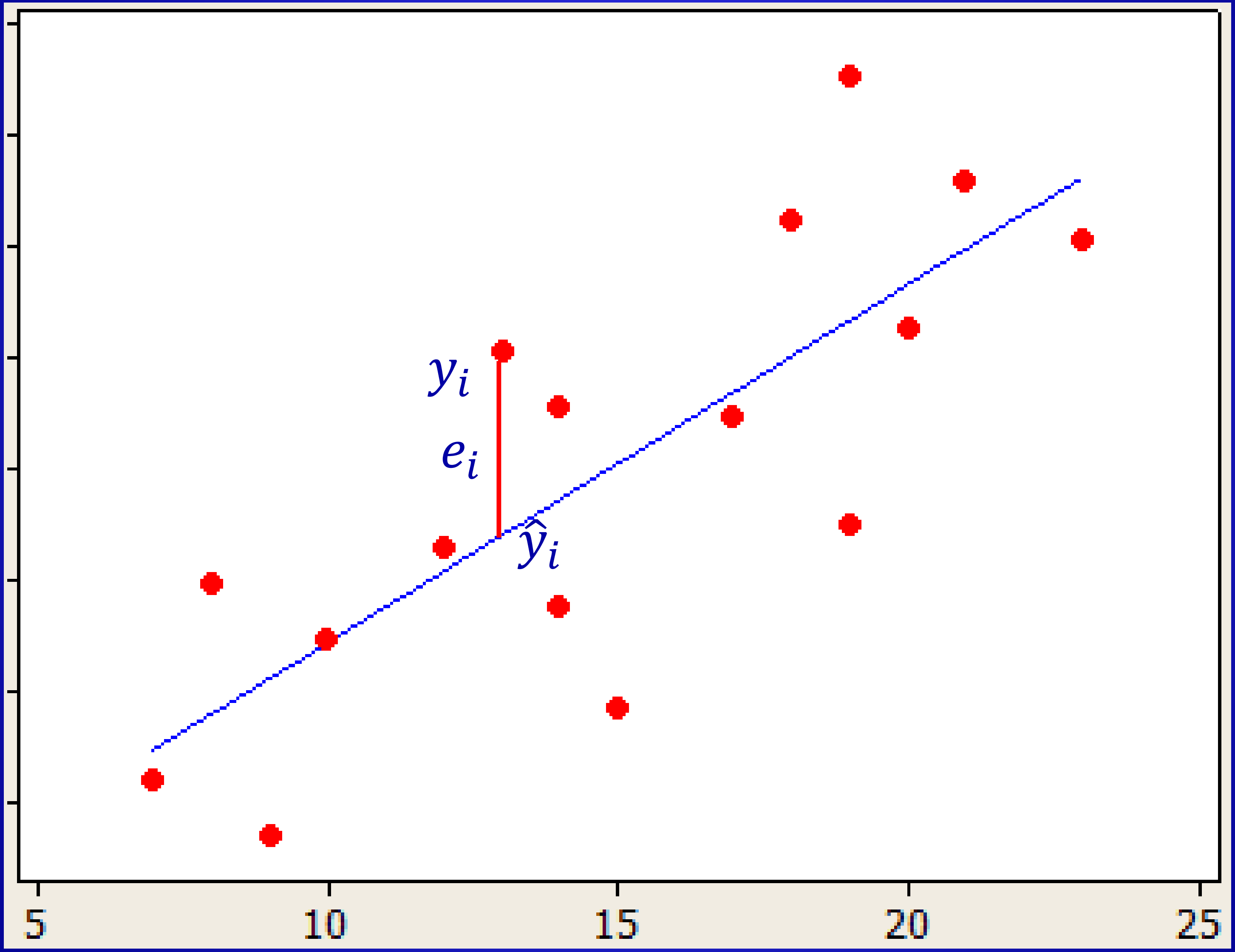
SSExplainedByModel= $\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k-1}$

SSError= $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ x_{1i} + \hat{\beta}_2 \ x_{2i} + .. + \hat{\beta}_k \ x_{Ki}$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

The residuals



Quantify Goodnes of model

SSTotal= V(Y) = $\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

SSExplainedByModel= $\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k-1}$

SSError= $\frac{\sum_{i=1}^{n}(e_i)^2}{n-k}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \, x_{1i} + \hat{\beta}_2 \, x_{2i} + .. + \hat{\beta}_k \, x_{Ki}$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Quantify Goodnes of model

assume linearity

R^2 = proportion of explained variance

$$R^2 = 1 - \frac{SSError}{SSTotal}$$

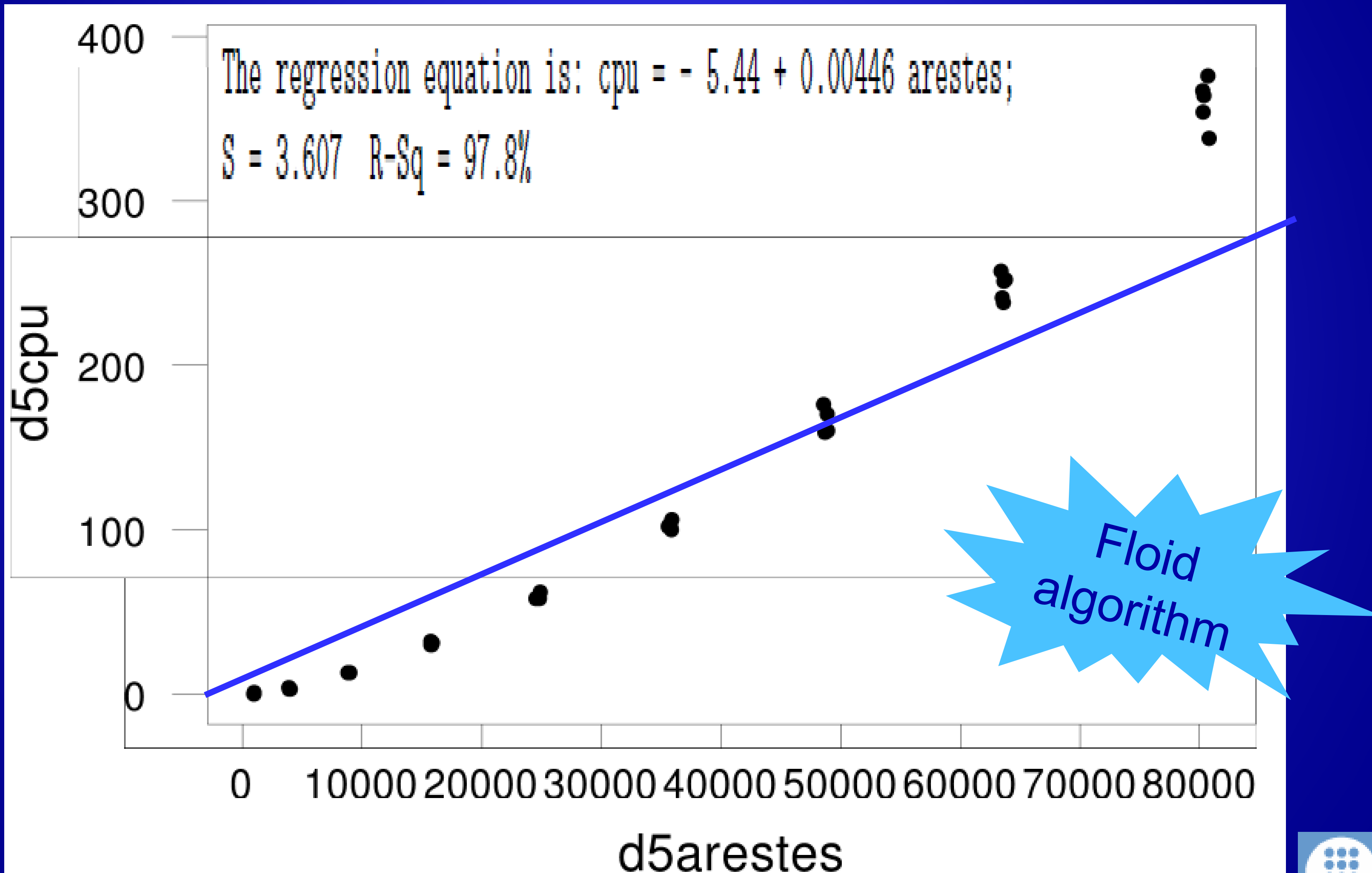
$0 < R^2 < 1$

The biggest R^2, the better the model explains Y

For simple linear regression R2=Corr(Y, X)^2

– Real case: Experimental CPU time of a graph

– treatment algorithm vs graph size



Model inference

To test significance of the model

$$F = \frac{SSExplainedByModel}{SSError} \sim F_{(k-1,n-k)}$$

To test significance of a model term $\widehat{\beta}_k$

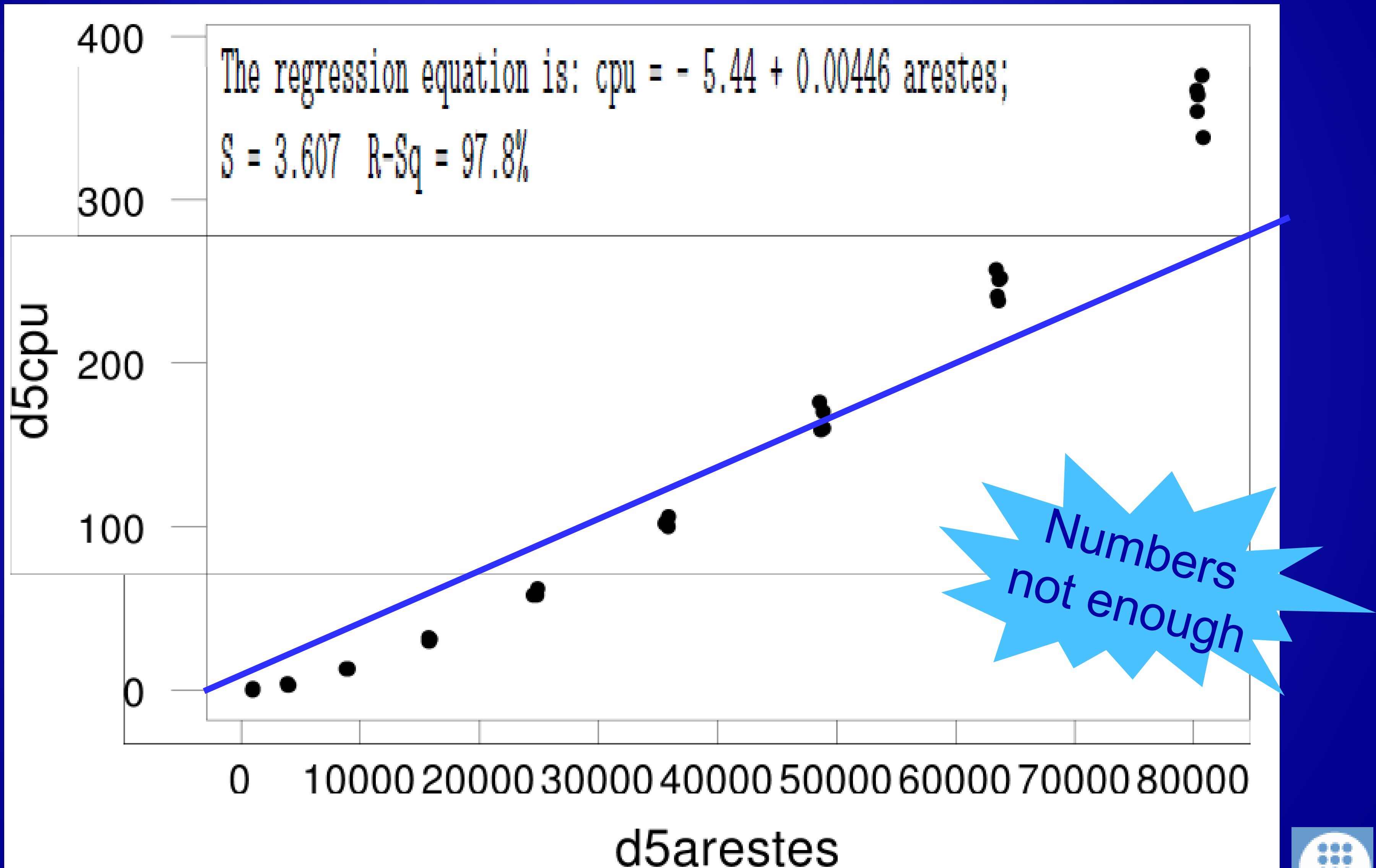
$$t_k = \frac{\widehat{\beta}_k}{S_{\widehat{\beta}_k}} \sim t_{n-K}$$

To test significance of a model term

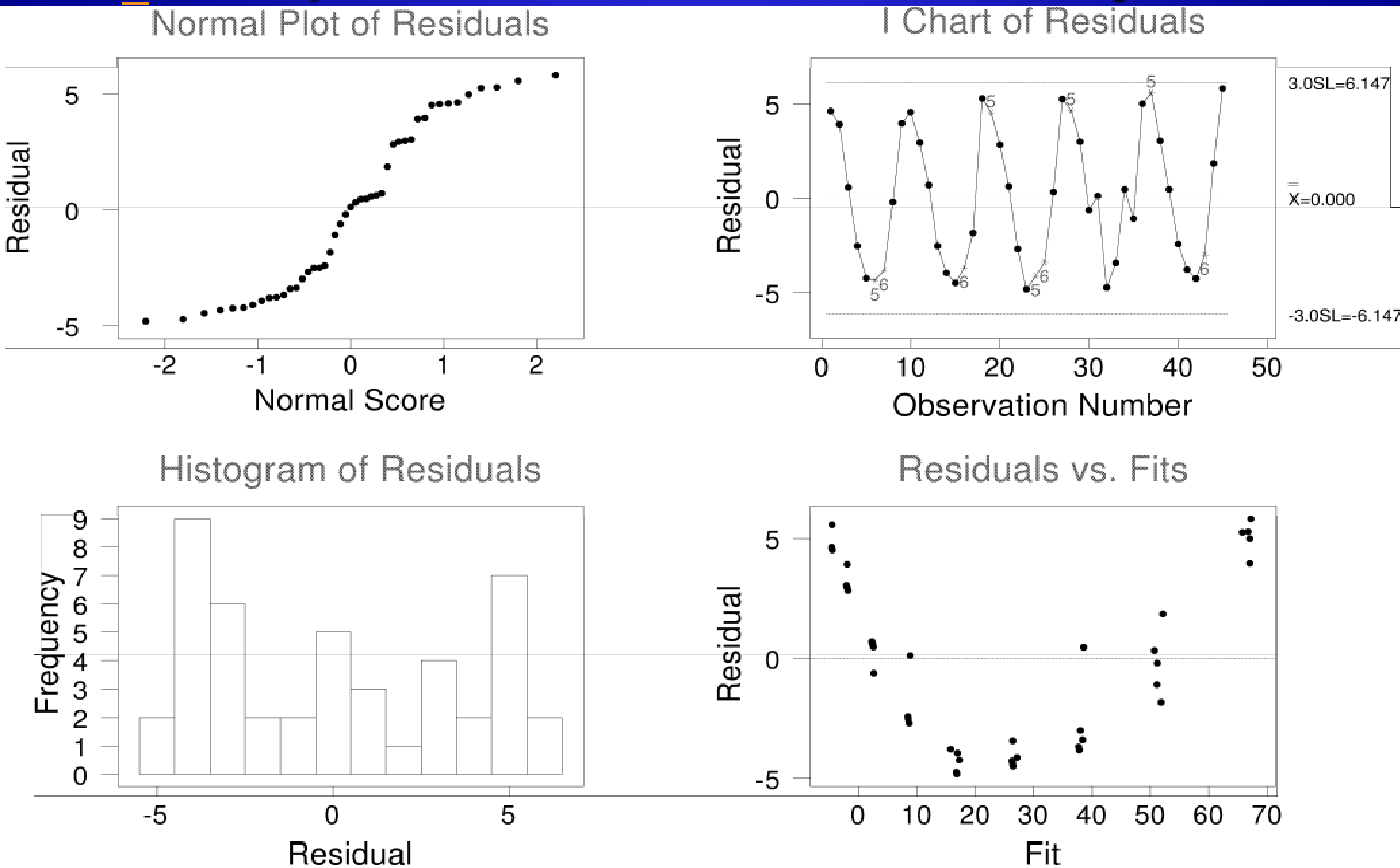
Both assume normality

– Real case: Experimental CPU time of a graph

– treatment algorithm vs graph size

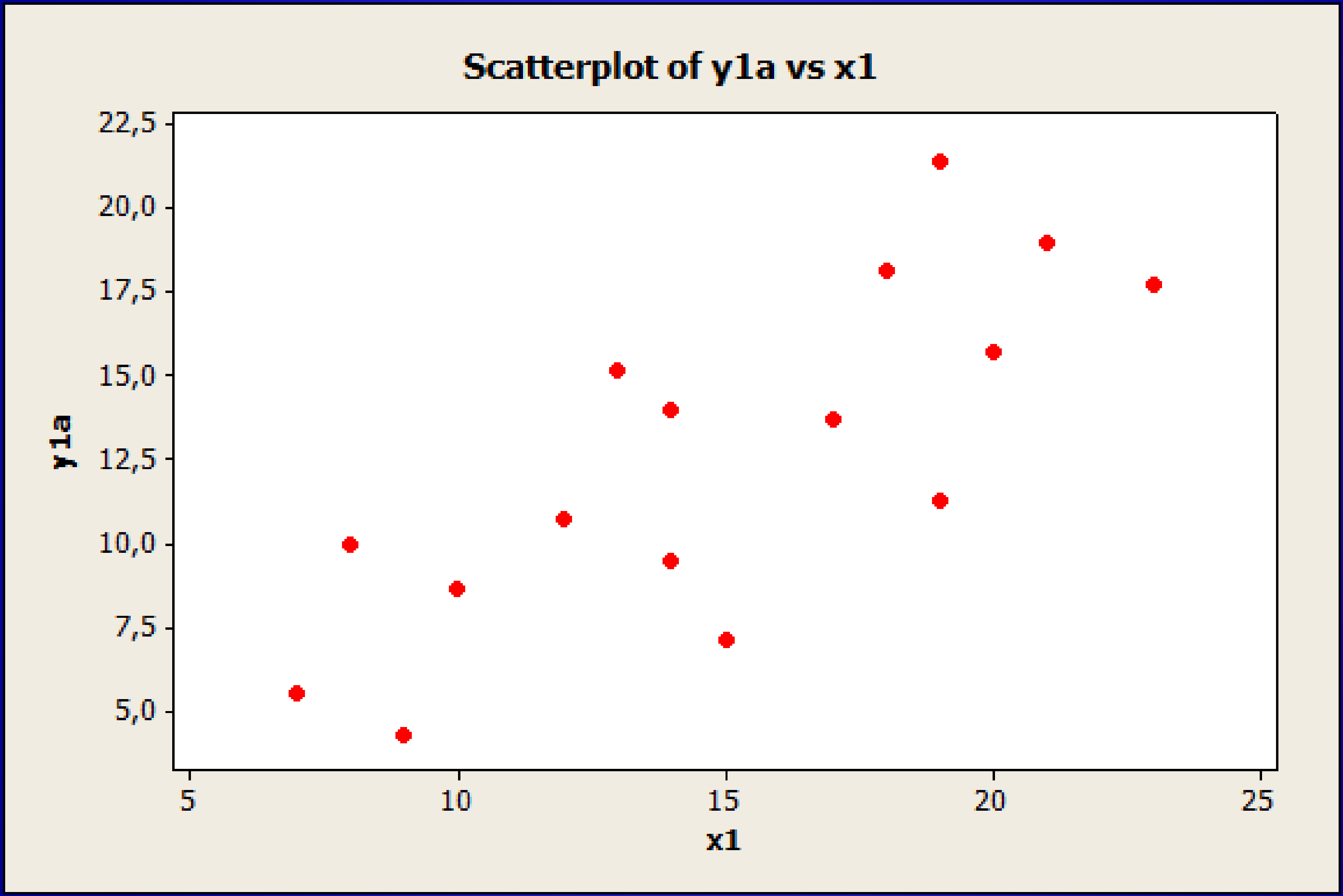


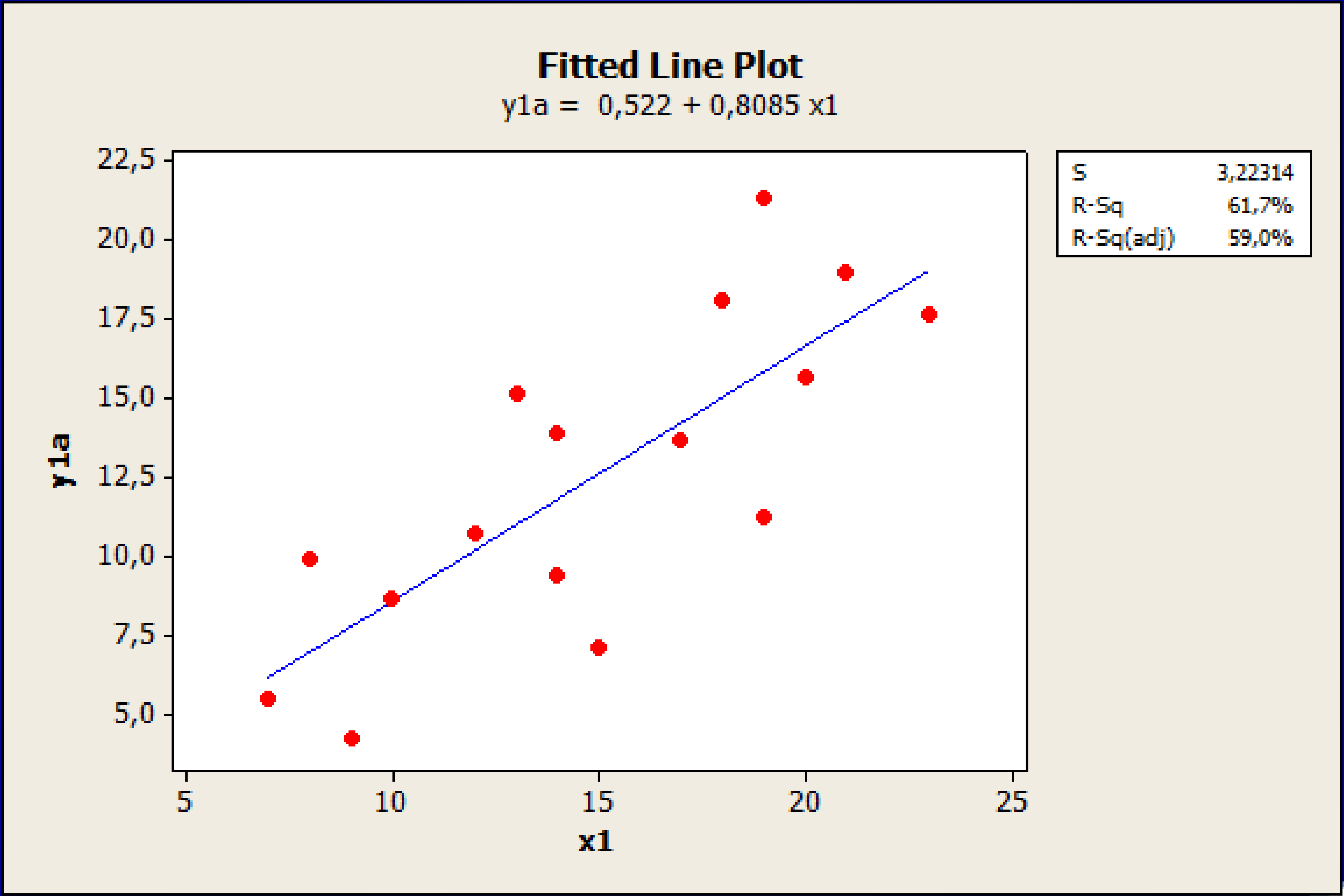
Graphical residuals analysis



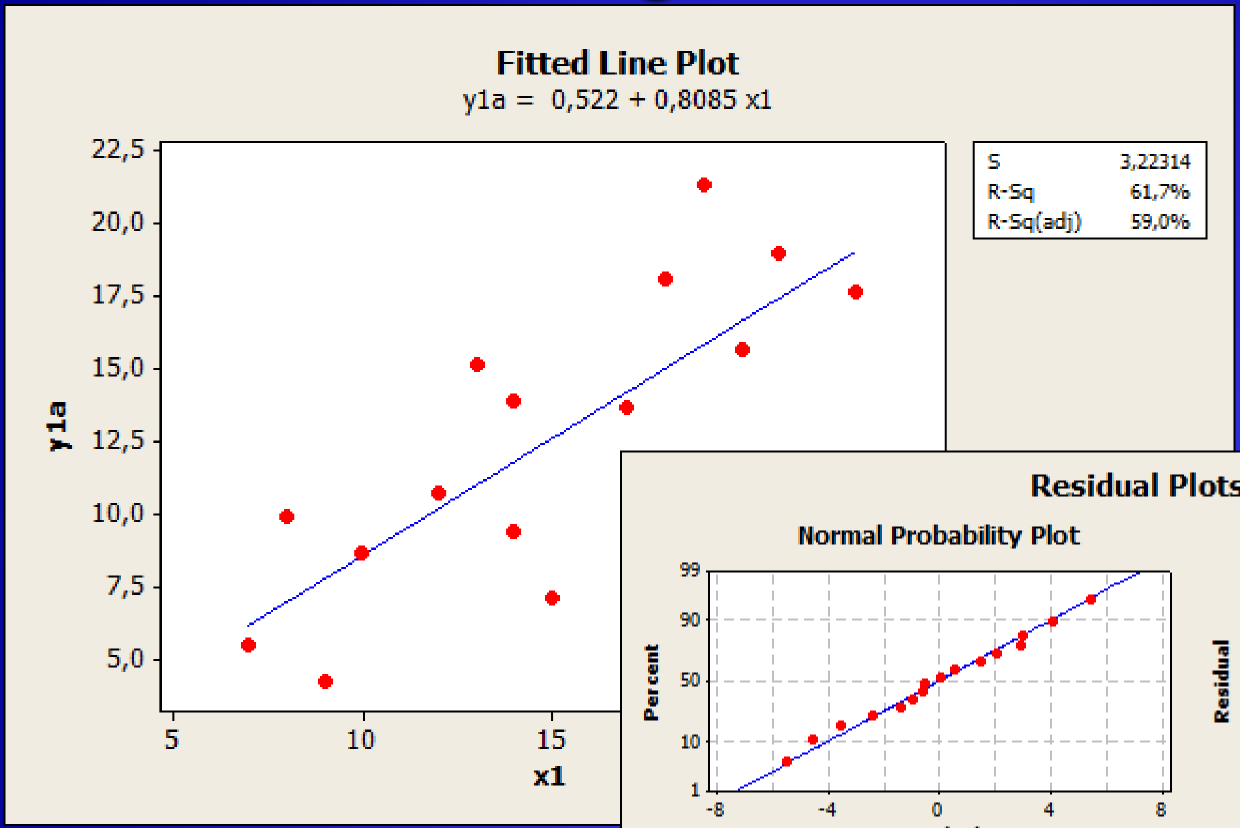
Regression

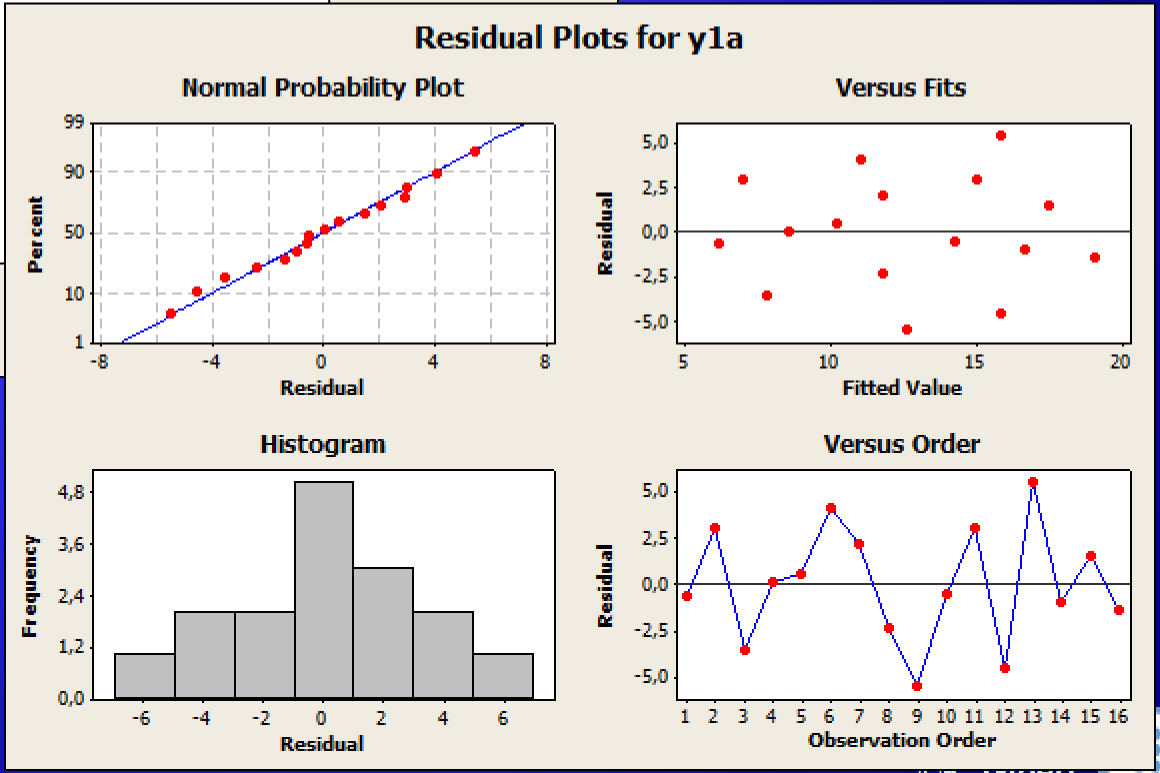
[Tomassone 56]



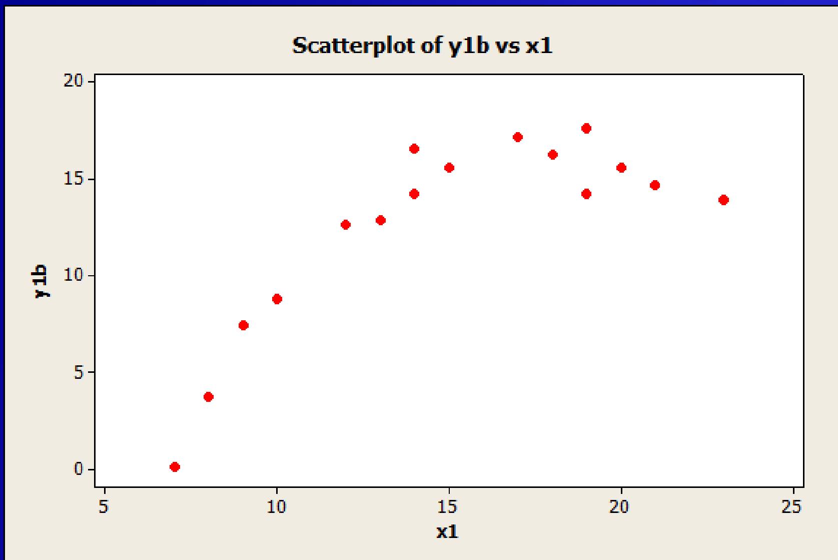

Fitted Line Plot
y1a = 0,522 + 0,8085 x1
S 3,22314
R-Sq 61,7%
R-Sq(adj) 59,0%
22,5
20,0
17,5
15,0
12,5
10,0
7,5
5,0
y1a
5
10
15
20
25
x1

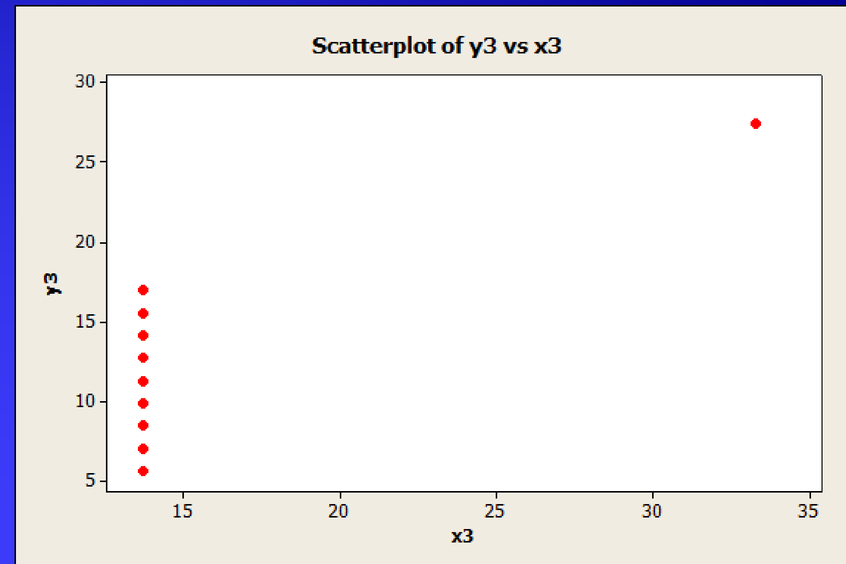
Regression

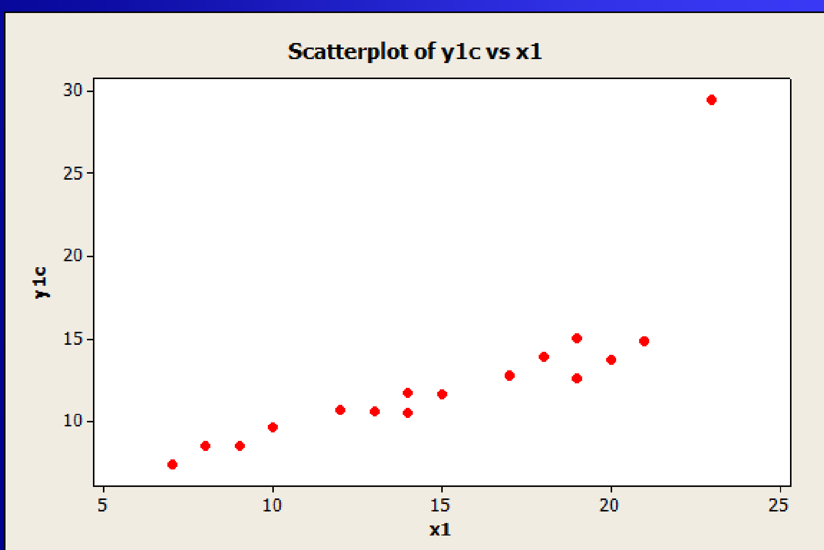

Fitted Line Plot
y1a = 0,522 + 0,8085 x1
S 3,22314
R-Sq 61,7%
R-Sq(adj) 59,0%
y1a
x1

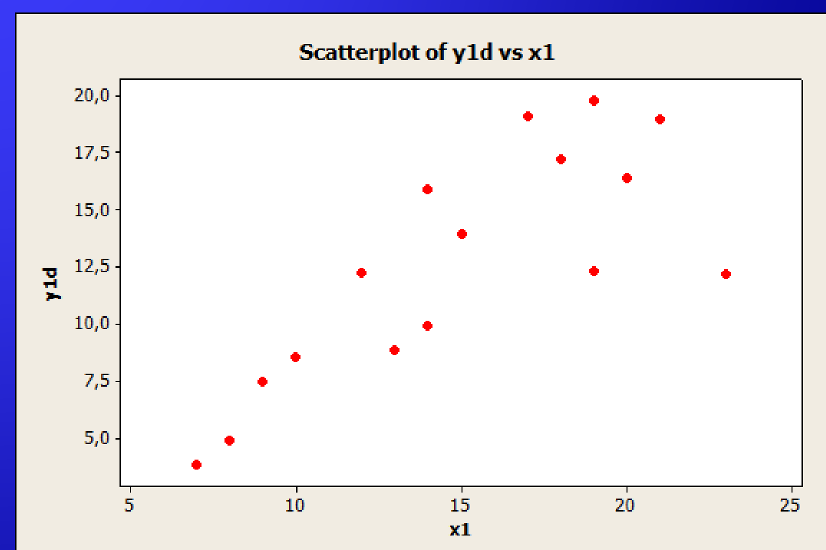

Residual Plots for y1a
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order

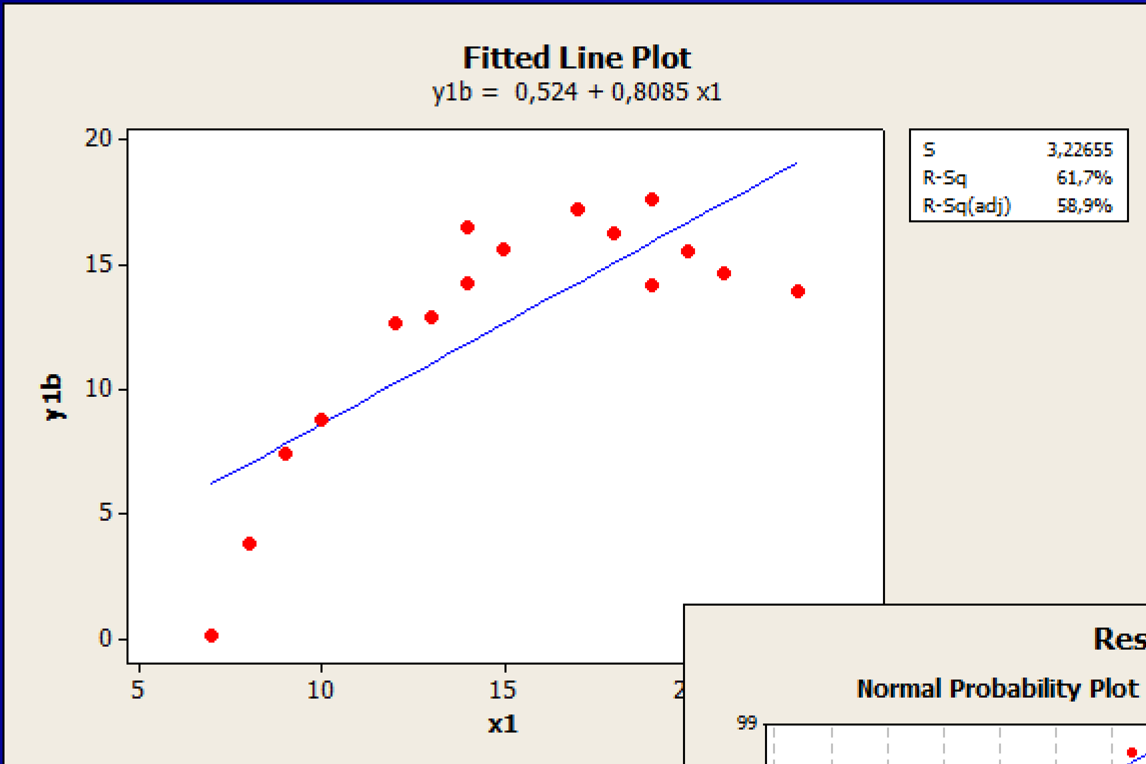
Regression

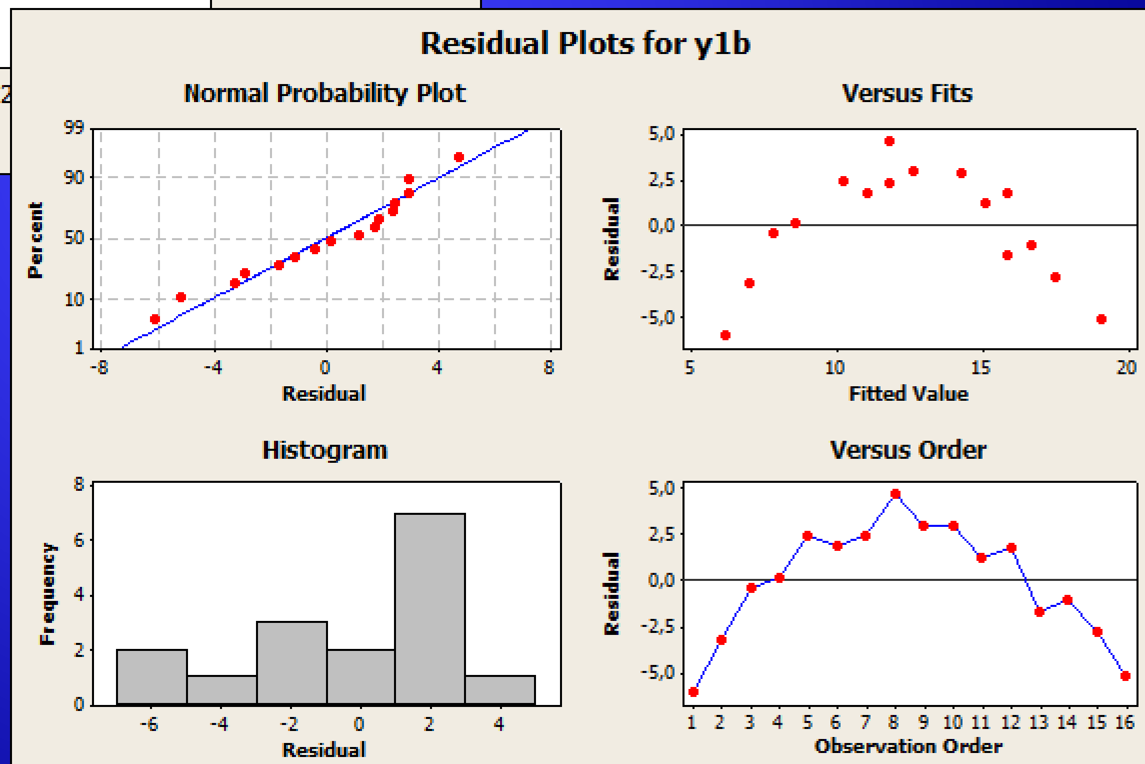

Scatterplot of y1b vs x1
y1b
20
15
10
5
0
5
10
15
20
25
x1

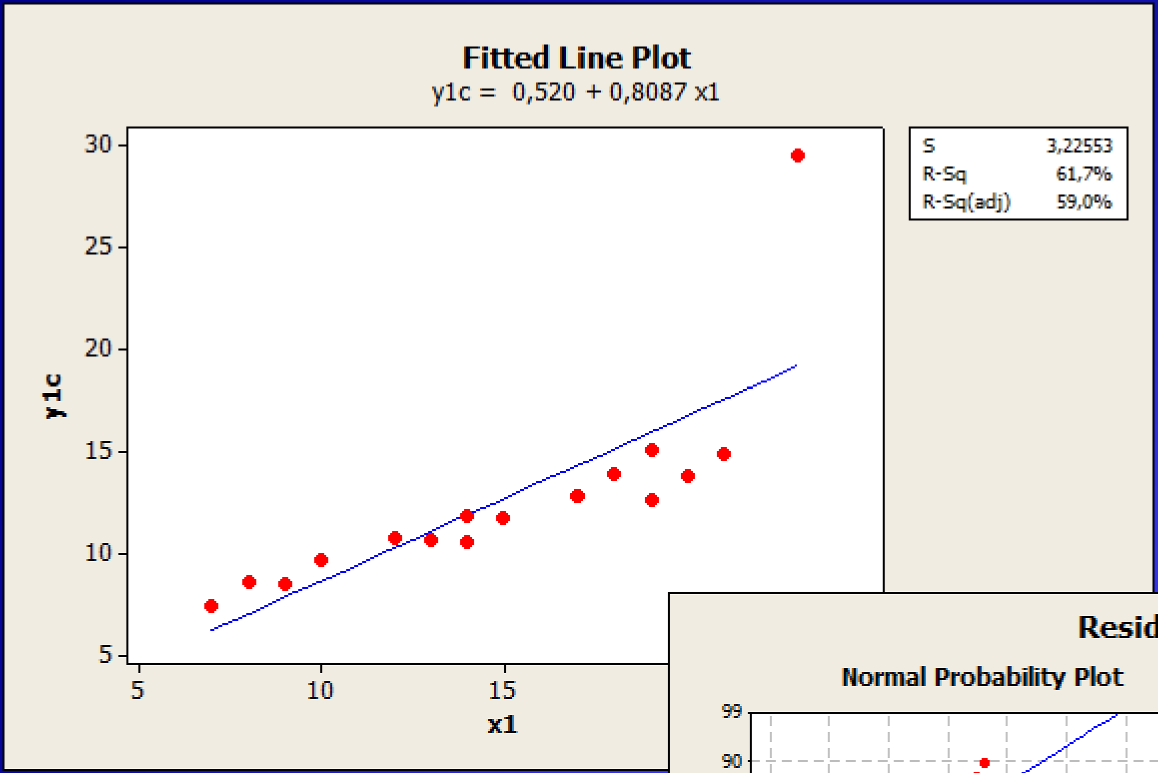

Scatterplot of y3 vs x3
y3
30
25
20
15
10
5
15
20
25
30
35
x3

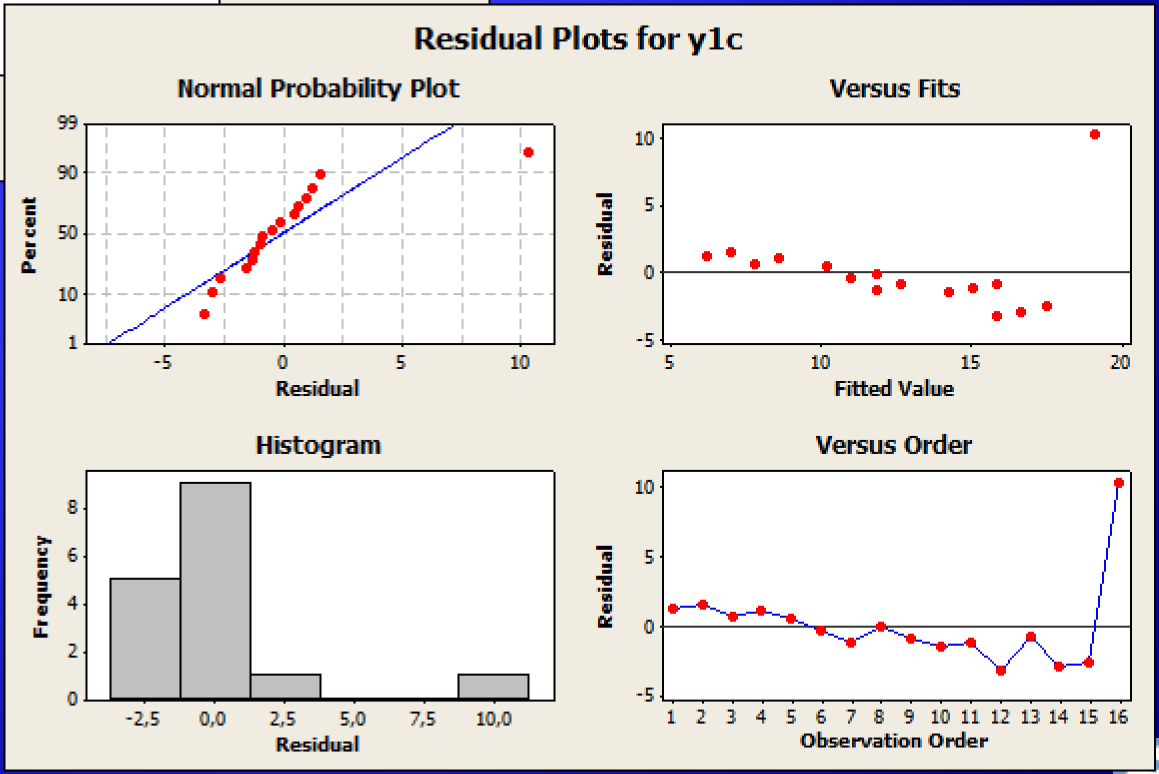

Scatterplot of y1c vs x1
y1c
30
25
20
15
10
5
10
15
20
25
x1


Scatterplot of y1d vs x1
y1d
20,0
17,5
15,0
12,5
10,0
7,5
5,0
5
10
15
20
25
x1


Fitted Line Plot
y1b = 0,524 + 0,8085 x1
S 3,22655
R-Sq 61,7%
R-Sq(adj) 58,9%
y1b
x1
0
5
10
15
20
5
10
15

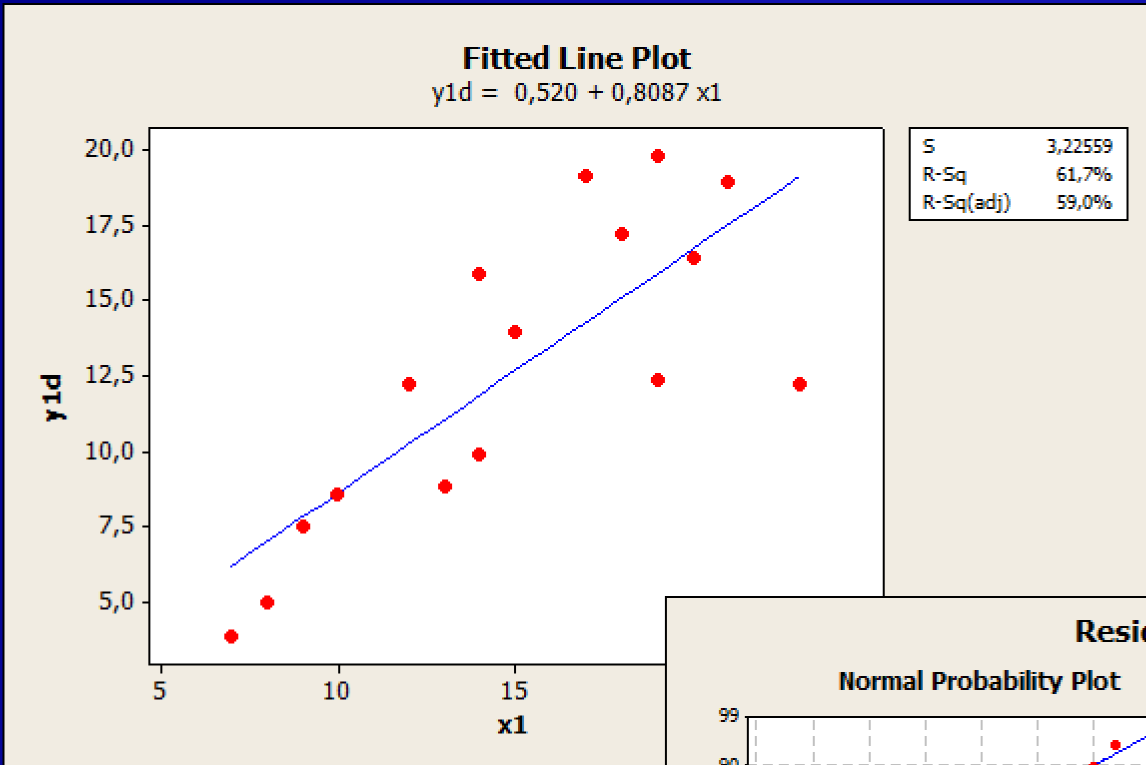

Residual Plots for y1b
Normal Probability Plot
Percent
Residual
99
90
50
10
1
-8
-4
0
4
8
Versus Fits
Residual
Fitted Value
5,0
2,5
0,0
-2,5
-5,0
5
10
15
20
Histogram
Frequency
Residual
8
6
4
2
0
-6
-4
-2
0
2
4
Versus Order
Residual
Observation Order
5,0
2,5
0,0
-2,5
-5,0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

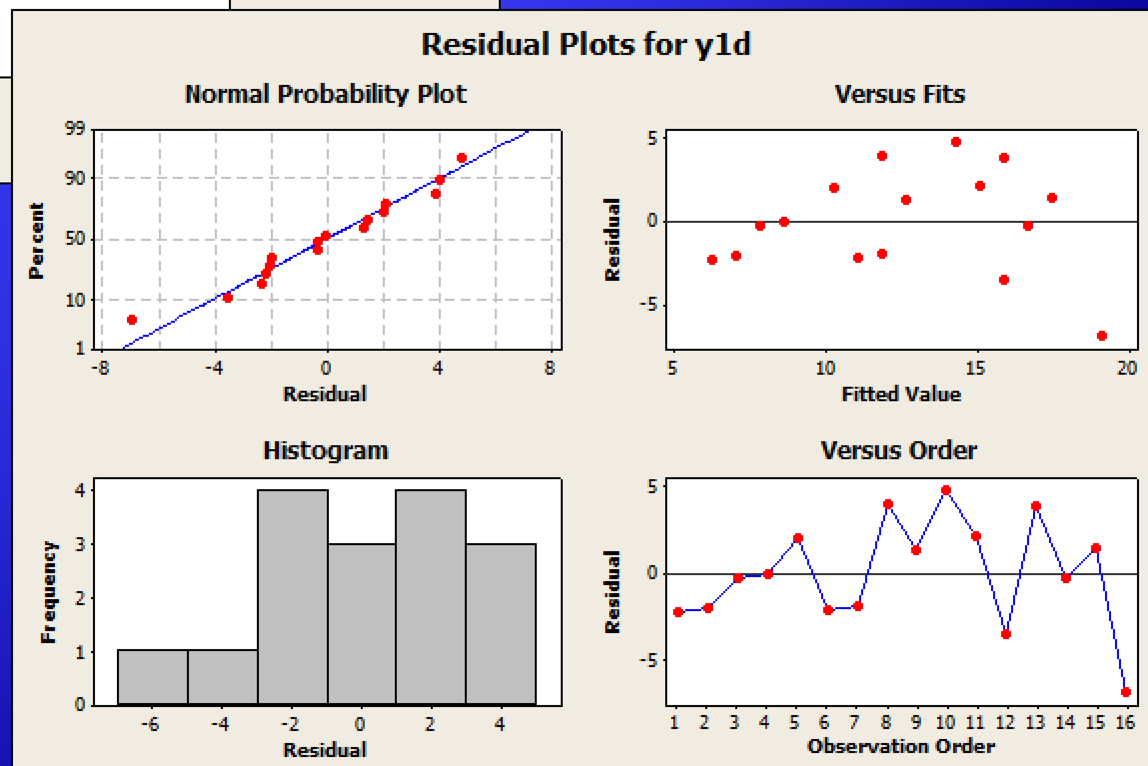


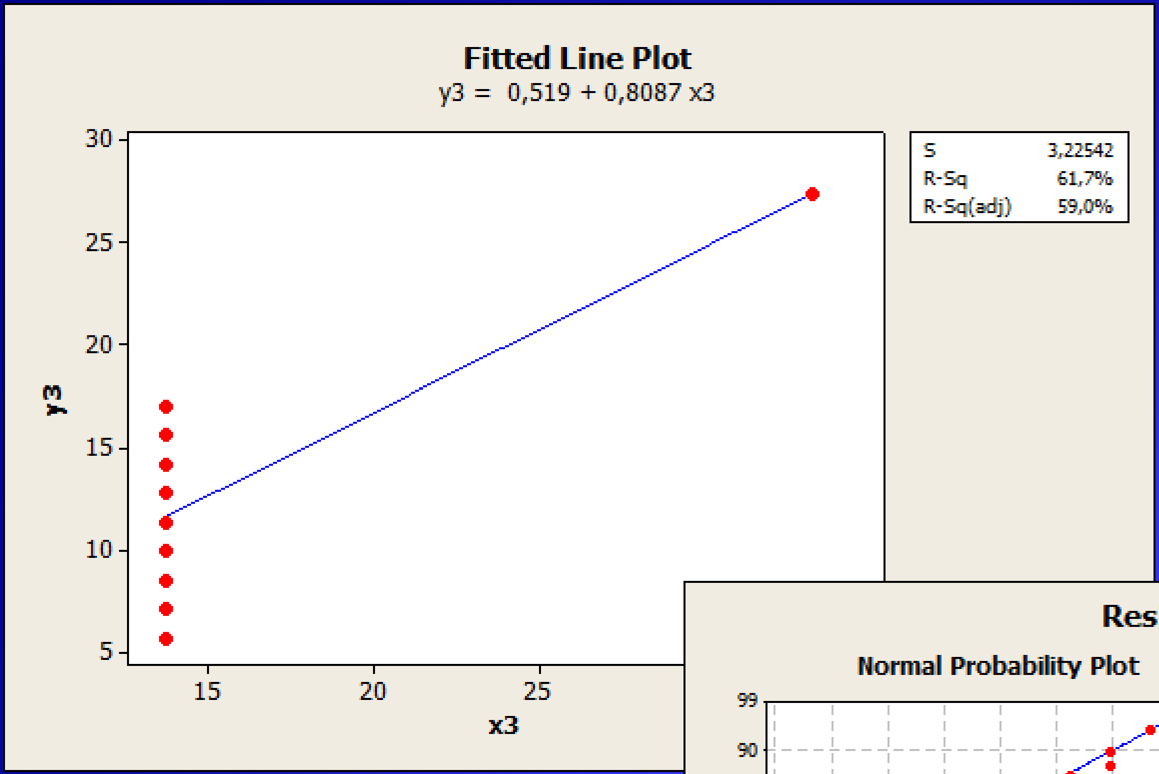


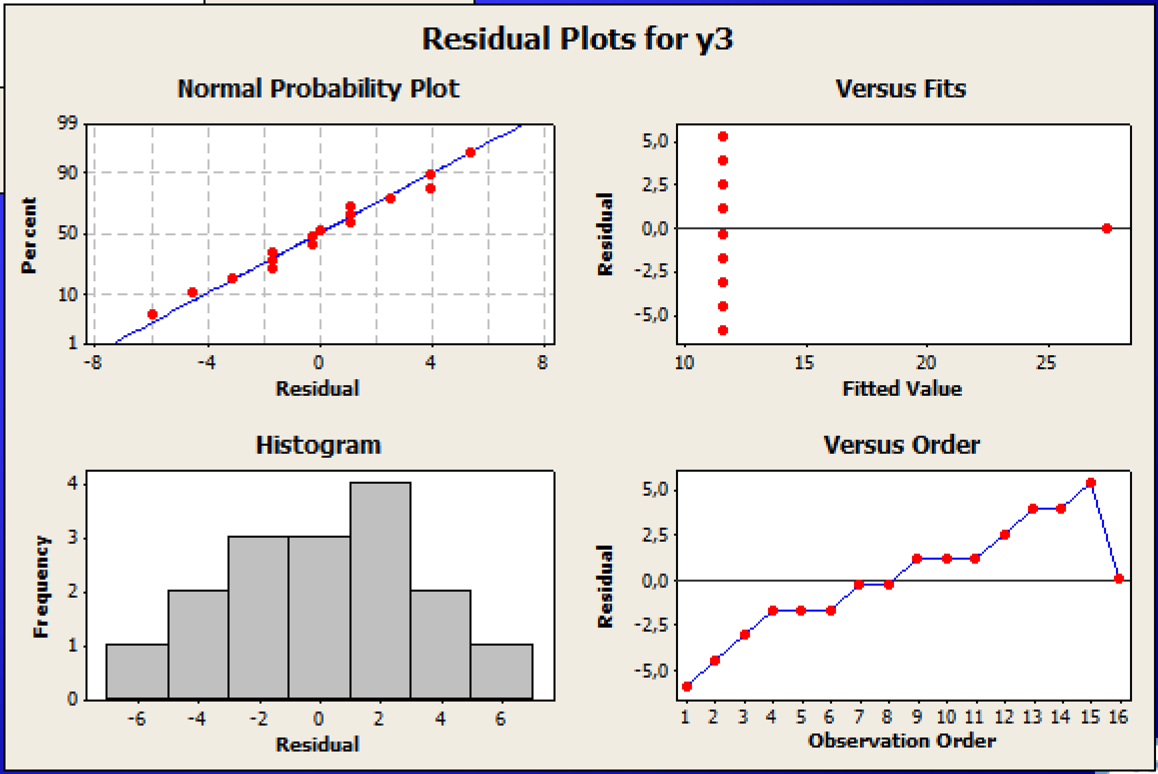
4/21/2020

©K. Gibert


Fitted Line Plot
y1d = 0,520 + 0,8087 x1
S 3,22559
R-Sq 61,7%
R-Sq(adj) 59,0%
y1d
x1
20,0
17,5
15,0
12,5
10,0
7,5
5,0
5
10
15


Residual Plots for y1d
Normal Probability Plot
Percent
Residual
99
90
50
10
1
-8
-4
0
4
8
Versus Fits
Residual
Fitted Value
5
0
-5
5
10
15
20
Histogram
Frequency
Residual
4
3
2
1
0
-6
-4
-2
0
2
4
Versus Order
Residual
Observation Order
5
0
-5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16


Fitted Line Plot
y3 = 0,519 + 0,8087 x3
S 3,22542
R-Sq 61,7%
R-Sq(adj) 59,0%
y3
x3


Residual Plots for y3
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order

Going further

Multiple linear regression

- Uses the same principles as simple linear regression
- Contains several regressions in the right hand side of regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon$$

- All properties of simple linear regression hold, except one
  - Determination coefficient (R^2) still measures goodness of fit
  - But R^2 is not equal to the squared correlation coefficient in multiple regression
- Analysis of the residuals is done by every regressor X_k

Modelling strategy

- 1) plots between Y and all X_k to chech the existence of linear association or the need of entering some X_k^2 or some 1/ X_k (provided that they make sense)
- 2) Check pvalues of all coefficients in the final regression equation. Repeat the model with all those with non signifficant p-value. Repeat till all coefficients look signifficant
- 3) With the model with all signifficant coefficients, write the equation of the final regression model
- 4) Make the graphical analysis of the residuals and verify all checks are passed
- 5) If some check fails,..... Proceed accordingly and repeat the model
- 5) Report final equation and R^2 . Also include R^2 value of the very first model
- 6) report all intermediate work done to move from the very first model to the very last and all decisions made in the meanwhile

Going much further

- ANCOVA: to introduce qualitative variables
- Interaction terms to introduce multiplicative models
- Polynomic regression to estimate higher order polynomial functions
- General Linear Model (common formulation for simple/multiple linear regression, ANOVA and ANCOVA)
- Generalized linear models: common formulation for an extension of families of models:
  - Linear,
  - Poisson
  - Logit.....
- Non linear relationships: LOESS (Locally Weighted Least Squares Regression), uses more local data to estimate the model. It uses a 'nearest neighbors' method to smooth data.
- Complex functions: Artificial Neural Networks

Supporting materials

Basic linear regression:
https://www.scribd.com/doc/304772056/BI-147-Web Chapter 2
https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&ved=2ahUKEwiQ7u_nx_noAhWE3oUKHYlQCo8QFjAIegQIChAB&url=https%3A%2F%2Fwww.parisnanterre.fr%2Fmedias%2Ffichier%2Fcours_regression(2)_1273084206969.pdf%3FID_FICHE%3D204222%26INLINE%3DFALSE&usg=AOvVaw3bXGbamKKlTIn5ewLD33DS

Linear Regression with R:
http://r-statistics.co/Linear-Regression.html

Multiple linear regression (matricial notation)
http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/401-multreg.pdf

https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwji5-qkwvnoAhV0kFwKHRbiAscQFjABegQIAxAB&url=https%3A%2F%2Fwww.uv.es%2Furiel%2FChapter%25203%2520Slides.pdf&usg=AOvVaw0qzppaA8Aj6vRa79hrIBxr

Moore, McCabe. "Craig (2012) Introduction to the Practice of Statistics." Chap 10

For the graphical analysis of the residuals:
Tomassone, Richard, Elisabeth Lesquoy, and Claude Millier. "La régression. Nouveaux regards sur une ancienne méthode statistique." (1983).