Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Bivariate data, correlation and simple linear regression
- Multiple linear regression
- Confidence intervals for a coefficient
- Prediction interval for a new observation
- Model selection
- Influential points
- Diagnostics for model assessment

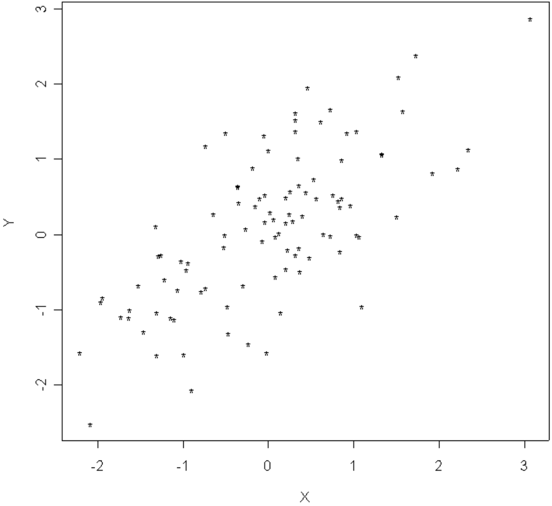
Bivariate data

- Measures on *two* variables ; *e.g.* X et Y
- We will consider the case of *continuous* variables
- We want to explore/discover the *relation*between the two variables
- We will consider sets of variables that are (at least approximately) *bivariate normal*

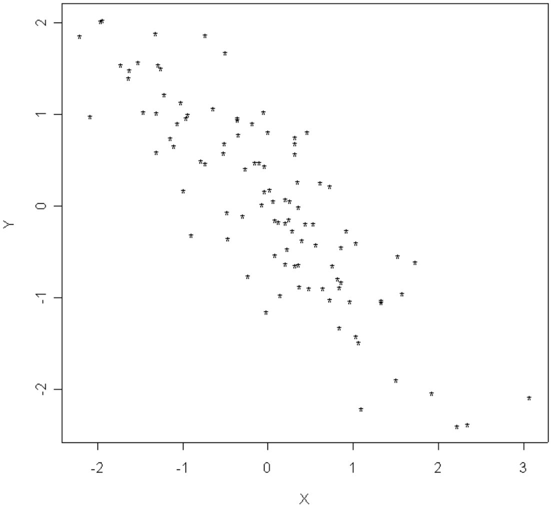
Scatterplot

- Graphical summary of bivariate data
- Values of one variable are plotted on the horizontal axis, the other on the vertical axis
- Used to visualize how the values of 2 variables are *associated*)

Scatterplot : positive association



Scatterplot : negative association



Numerical summaries

- Typically, bivariate data are summarized (numerically) with 5 statistics
- These give a good summary for oval-shaped scatterplots
- We summarize each variable *separately* : $\overline{X}, s_X; \overline{Y}, x_Y$
- But these values tell us nothing about how X and Y *vary together*

Correlation

- For random variables X and Y,with $Var(X) > 0, Var(Y) > 0$, the **correlation** $\rho(X, Y)$ is defined as :

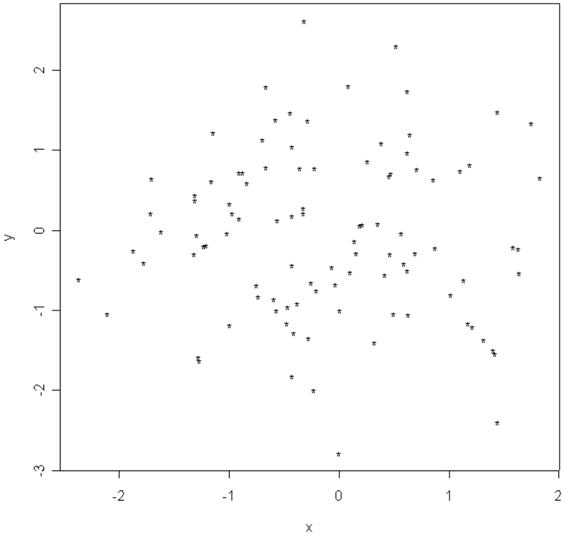
$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- ρ is a *unitless quantity*, $-1 \leq \rho \leq 1$
- ρ is a measure of **LINEAR ASSOCIATION**
- Values of ρ close to 1 or -1 indicate a strong linearity between X and Y, while values close to 0 indicate an absence of a **linear** relation
- The sign of ρ indicates the direction of association (positive or negative, corresponding to the slope of the line)
- When $\rho(X, Y) = 0$, X and Y are **uncorrelated**

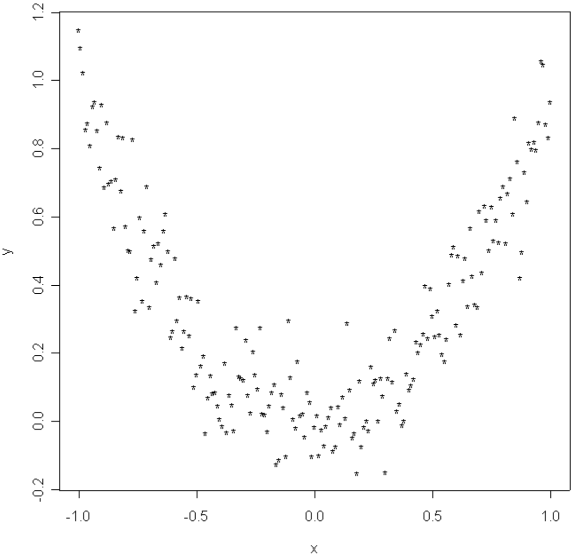
Correlation $\neq$ Causation

- We *cannot deduce* that, for X and Y strongly correlation, X causes a change in Y
- It could be that Y causes X
- X and Y could both vary as a function of a third variable, possibly unknown (whether causal or not, often time)

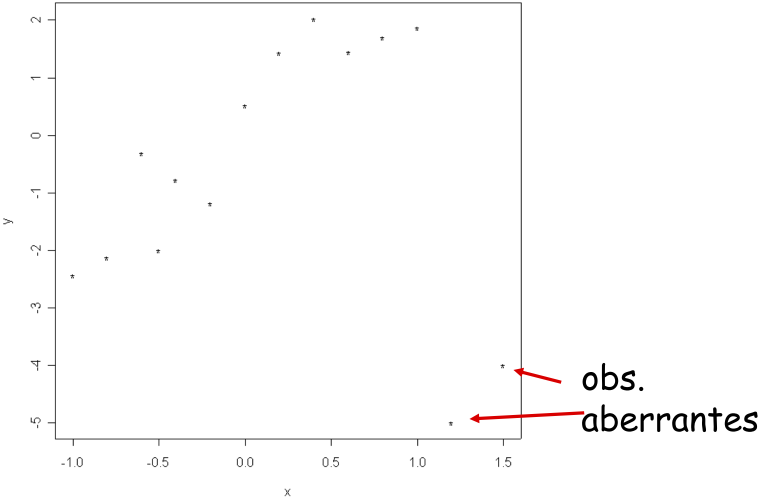
$r \approx 0$: random dispersion



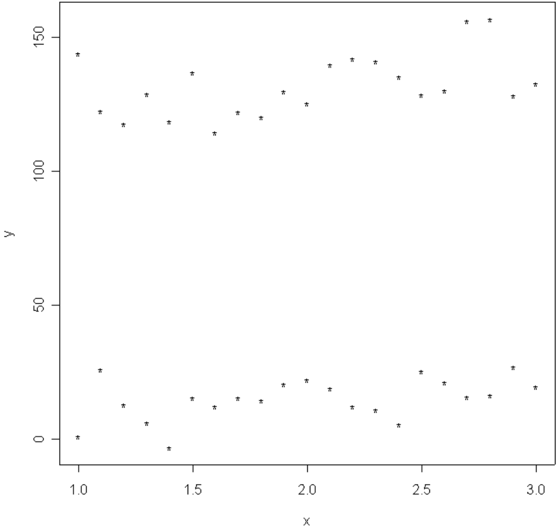
$r \approx 0$: curve



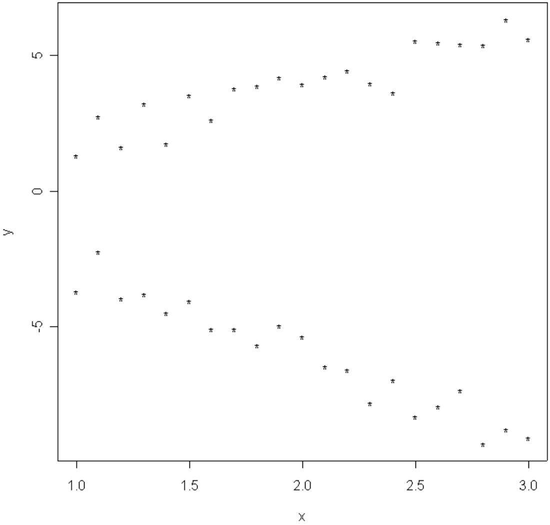
$r \approx 0$: outliers



$r \approx 0$: parallel lines



$r \approx 0$: two different lines



Simple linear regression

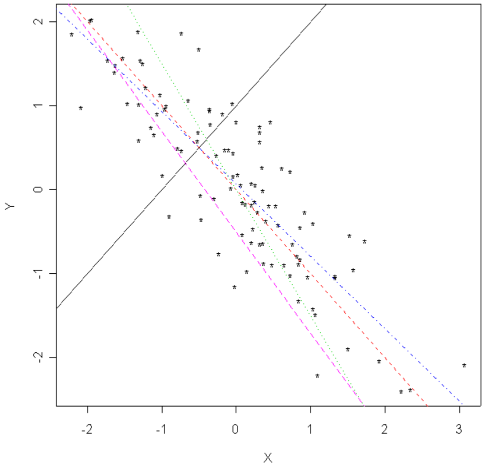
- Refers to a special line through a cloud of points in a scatterplot
- Used for 2 objectives :
  - Explanation
  - Prediction
- The equation for predicting y knowing x :

$$y = \beta_0 + \beta_1 * x$$

- $\beta_0 =$ l'*intercept* ; $\beta_1 =$ la *slope*

Which line ?

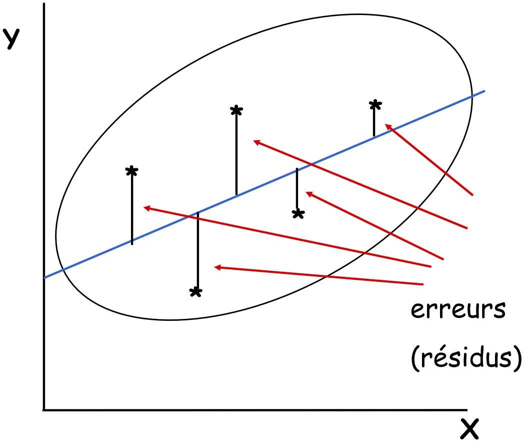
- Many possible lines can be drawn through the point cloud
- How to choose ?



Least squares

Q : How do we choose the prediction line ?

R : It is the 'best' in the sense that the sum of the *squared errors* in the vertical direction (Y) is the *minimum*



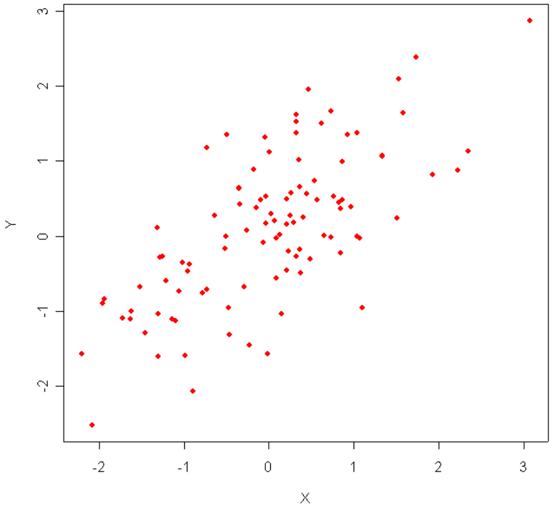
Parameter interpretation

- There are 2 parameters in the regression line :
  the *slope* and the *intercept*
- The*slope* is the average (expected) change in Y for a 1 unit change in X
- The *intercept* is the estimated value of Y when $X = 0$
- If the slope $= 0$, X does not give (linear) information for predicting Y

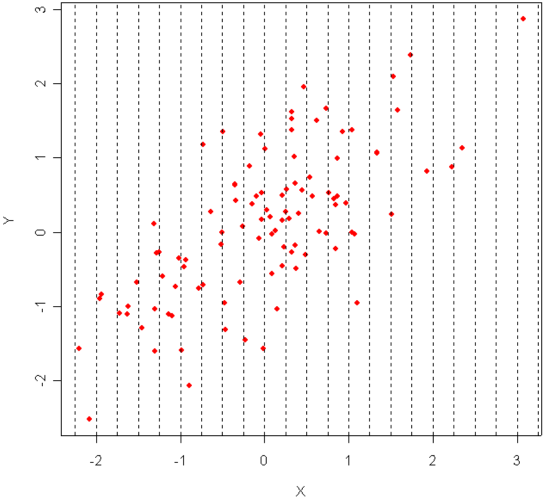
Another view of the regression line

- We can divide the scatterplot into regions (X*-bands*) based on values of X
- For each X-band, plot the average value of Y
- This is the *graph of averages*
- The regression line can be considered as a *smoothed version* of the graph of averages

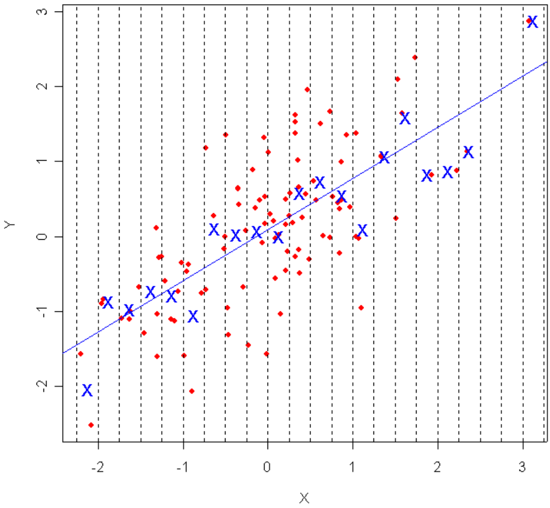
Scatterplot (again)



X-bandes



Graph of means



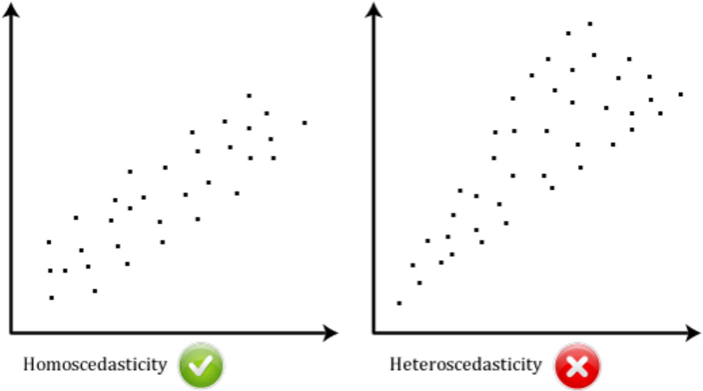
Simple linear regression – mathematics

- Here, we consider a model where the *réponse variable* y_i is linearly associated with an *explanatoryb* (or *predictor*) variable x_i :

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

- $\epsilon_1, \ldots, \epsilon_n$ are assumed to be random variables :
  - uncorrelated
  - expected value $= 0$
  - variance $= \sigma^2$ for all $i = 1, \ldots, n$ (*homoscedastic*)
- x_i are supposed constant (measured without error)
- $\rightarrow$ If the errors are also assumed to be *normally distributed*, we can carry out *hypothesis tests* and make *confidence intervals (CI)*

Homoscedastic, heteroscedastic errors



Least squares method

- The observed data are only a *sample* (not the entire population)
- Thus, we need to *estimate* the values of the population parameters β_0 (intercept) and β_1 (slope) :

$$\hat{y}_i = b_0 + b_1 x_i + \epsilon_i$$

- According to the *least squares principle*, we look for the estimators that minimize :

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

Estimation by (ordinary) least squares

- Now we have an optimization problem : find the values $\hat{\beta}_0$ et $\hat{\beta}_1$ minimizing

$$SC(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- To solve, differentiate wrt β_0, β_1 and find the zeros :

$$\begin{aligned}
\frac{d}{d\beta_0} &= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0 \\
=> &\quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
=> &\quad \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\
=> &\quad \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad (*)
\end{aligned}$$

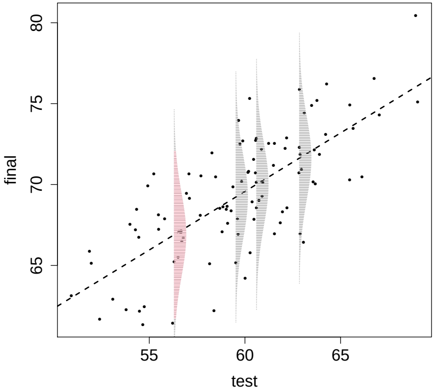
OLS, cont

$$
\begin{aligned}
\frac{d}{d\beta_1} &= \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\
=> & \sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \\
=> & \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
=> & \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad (**)
\end{aligned}
$$

Simultaneously solving (*) and (**) yields the **OLS estimates**

Conditional normal distribution

- Given x, the expected value is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Assuming homoscedasticity, the variance of y given x is the same for all x



Multivariate data

Individus	X_1	X_2	$\dots$	X_j	$\dots$	X_p
i_1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
i_2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\dots$						
i_i	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\dots$						
i_n	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

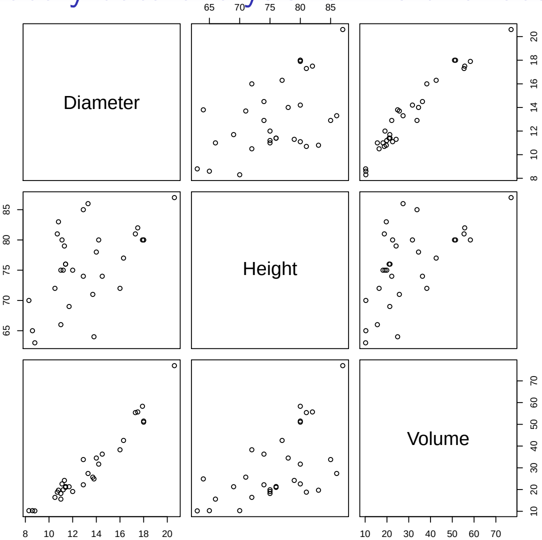
vector of means : $(\overline{x}_1, \dots, \overline{x}_p)$
matrix of variances-covariances (or *dispersion matrix*) :

$$\begin{pmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_2^2 & \cdots & s_{2,p} \\ \cdots & s_i^2 & s_{i,j} & \cdots \\ s_{p,1} & s_{p,2} & \cdots & s_p^2 \end{pmatrix}$$

Example

- A sample of cherry trees has been cut, and measures have been taken for :
  - `Diameter` (inches)
  - `Height` (feet)
  - `Volume` (cubic feet)
- The goal of of this study is to provide a prediction of volume, given measures of Height and Diameter
- Here we will use a multiple regression model

Exploratory data analysis for multivariate data



Matrix algebra for simple regression

- The model :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Multiple regression

- We could add additional predictors into the regression equation, for example :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i, i = 1, \ldots, n$$

- We use the same technique to find estimates $\hat{\beta}_j$, $j = 1, \ldots, k$, that solve the LS optimization problem. Usually this is written in matrix form :

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

  where X is the *design matrix*

(Ordinary) least squares regression

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Find a solution $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals (*OLS solution*) :

$$\min \sum_{i=1}^{n} e_i^2 \rightarrow \frac{\partial \left(\sum_{i=1}^{n} e_i^2 \right)}{\partial \hat{\beta}_j} = 0, \quad j = 0, \ldots, p$$

$$\rightarrow \sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \quad j = 0, \ldots, p$$

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \rightarrow \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

$$\rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

for $\mathbf{X}'\mathbf{X}$ nonsingular, where $\mathbf{X}$ is the *design matrix* and $\mathbf{X}'$ is the transpose of the design matrix $\mathbf{X}$

Regression estimation output

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442 
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Regression estimation output

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)
```

équation — y = Volume, x_1 = Diameter, x_2 = Height

```
Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$ → (Intercept), $\hat{\beta}_1$ → Diameter, $\hat{\beta}_2$ → Height

→ **Volume = -57.99 + 4.71 x Diameter + 0.34 x Height**

Interpretation of regression coefficients

- The regression coefficients correspond to the expected (average) change in the response variable for a unit increase in an explanatory variable :
- For simple linear regression :
  - the slope is the expected change in y when the explanatory variable x increases by 1 unit
  - the intercept is the predicted value of y when $x = 0$
- An important distinction in the case of *multiple* predictor variables :
  - each coefficient $\beta_1, \ldots, \beta_p$ corresponds to the contribution of one variable when **all other variables in the equation are held constant**
  - the coefficient β_0 is the predicted value of y when **all variables** $x_1, \ldots, x_p = 0$

OLS properties : expected value

Dans le cas

1. $E(\epsilon_i) = 0,\ i = 1, \ldots, n$;
2. $Var(\epsilon_i) = \sigma^2$ (constante) ;
3. $Cov(\epsilon_i, \epsilon_j) = Cor(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

on a :

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}(\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

OLS properties : expected value

$$
\begin{aligned}
Var(\hat{\beta}) &= Var\left((\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\,\mathbf{Var}(\mathbf{y})\,\left((\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\,\sigma^{\mathbf{2}}\mathbf{I}\,\left((\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\right)' \\
&= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(((\mathbf{X}'\mathbf{X})^{-\mathbf{1}})' \\
&= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}
\end{aligned}
$$

($(\mathbf{X}'\mathbf{X})$ symmetric)

Regression estimation output

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 
```

erreur standard ($\hat{\beta}$)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *  
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442 
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

$\hat{\sigma}$ *(s)*

n-p-1

Tests/confidence intervals for the coefficients

- In addition, assuming $\epsilon_1, \ldots, \epsilon_n \sim$ iid $N(0, \sigma^2)$, we have

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\right)$$

- Thus, $Var(\hat{\beta}_i) = \sigma^2 \left[\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\right]_{i+1,\, i+1}$
- A CI with confidence level $100(1-\alpha)\%$ for β_i takes the form :

$$\hat{\beta}_i \pm \hat{\sigma}\sqrt{\left[\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\right]_{i+1,\, i+1}}\; t_{n-p-1,1-\alpha/2}$$

- To test $H : \beta_i = 0$ vs. $A : \beta_i \neq 0$

$$t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{\left[\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\right]_{i+1,\, i+1}}}$$

- We REJECT H if : $|t_{obs}| > t_{n-p-1,1-\alpha/2}$
(equivalently, if the CI does not contain the value 0)

Prediction interval for a new observation

- In simple linear regression, a $100(1-\alpha)\%$ **prediction interval** for a new (single) observation with $x = x_0$ is given by :

$$\hat{\beta}_0 + \hat{\beta}_1 \pm \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\; t_{n-2,1-\alpha/2}$$

- A PI is *wider* than a CI for a given level
- A CI can be made as narrow as desired by increasing the sample size n
- The same is **NOT** true for a PI, since the new observation will be subject to an observation error that is not reduced by increasing n

Regression estimation output

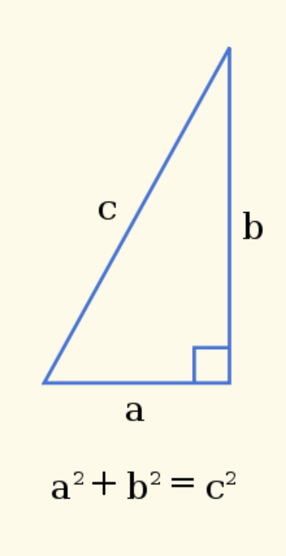
```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

t

***p*-valeur**

niveau de signification α

Pythagoren theorem



Least squares geometry

- Consider $\mathbf{y}$ as a vector in n-dimensional space
- The column vectors of $\mathbf{X}$ form a p-dim subspace
- The predicted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ represents the point in the subspace that is closest to the observations : OLS is the *orthogonal projection* of $\mathbf{y}$ on the subspace of $\mathbf{X}$
- The residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is *orthogonal* to vectors in the subspace
- $SCE = \sum e_i^2 = \mathbf{e}'\mathbf{e}$ is the square of the distance from the vector of obs. to the closest point in the subspace
- Partition $\mathbf{y}$ in *two orthogonal components* :
    - $\hat{\mathbf{y}}$ (model subspace, p dims)
    - $\hat{\mathbf{y}} - \mathbf{y}$ (error subspace, $n - p$ dims)
- (degrees of freedom correspond to the subspace dims)

Geometry of LS

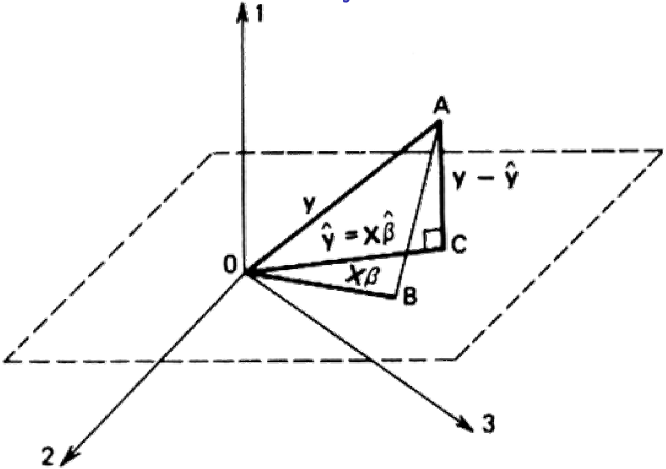


Figure 4.2 A geometrical interpretation of least squares.

PAUSE

Analysis of variance table (ANOVA)

- Uses the Pythagorean theorem to *partition the total sum of squares (SST)*
- Pythagorean theorem :

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

- equally :

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

- We can present this equality in the form of a table :

Tableau d'ANOVA

source	df	SS	MS (=SS/df)	F	p-value
regression	p	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	SSM/p	MSR/MSE	$P(F_{obs} > F_{p,n-p-1})$
error	$n-p-1$	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$SSE/(n-p-1)(= \hat{\sigma}^2)$		
total (corr.)	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$			

F-test

- The statistic $F_{obs} = MS(\text{source}/MSE)$ tests the hypothesis $H_0 : \beta_1 = \ldots = \beta_p = 0$ vs. A : at least 1 $\beta_i \neq 0$
- The distribution of F_{obs} when H is true is *the Fisher distribution $F_{p,n-p-1}$*
- The numerator of F_{obs} is *the variability explained by the regression model*
- The denominator contains *the residual variance*
- Under the null, the expected value of F is 1 and under the alternative the expected value is bigger than 1
- $\rightarrow$ REJECT the null hypothesis H for *large values of F*
- When testing a single coefficient $(H : \beta_i = 0)$, $F_{1,n-1} = t^2_{n-1}$

Regression estimation output

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

$F_{p,n-p-1}$

p-valeur

Coefficient of determination R^2

- The value y_i can be decomposed in two parts : one part *explained by the model* and one part *residual*
- The dispersion for the data can therefore be decomposed as :
  1. variance explained by the regression, and
  2. residual (unexplained) variance
- The *coefficient of determination* (or *multiple correlation*) R^2 is defined as the ratio between the explained and total variance : SSR/SST
- Equally, $R^2 = 1 - SCE/SCT$
- In *simple linear regression*, this is just the square of the correlation coefficient

Adjusted R^2

- The *adjusted R^2 (R^2_{aj})* takes into account the *number of variables* in the model
- A principal fault of R^2 is that it is *non-decreasing in the number of explanatory variables*
- Too many variables produces models that are *not robust*
- So we are more interested in the value of R^2_{adj} than R^2
- $R^2_a dj$ is not a true 'square' – it can even take on negative values

$$R^2_{aj} = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Regression estimation output

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

R^2 R^2-ajusté

R^2 or R^2-ajusté ?

Model selection

- Could fit all possible effects into a model
  - BUT : a model that is too big will be difficult to understand
- Instead, remove effects that are not important
- **HOW ? ? ?**
- A good model should
  - fit the data reasonably well
  - be as simple as possible for its intended purpose (*e.g.* descriptive, explanatory, prediction)
  - be interpretable
- Tradeoff : between *fit* and *complexity* of the model

Criteria for model comparison

- F-tests for individual effects
  - **Beware :** the *order* of the terms in the model can make a difference (nonorthogonal designs)
- Information Criteria (AIC, BIC)
  - xIC = Deviance + Complexity
  - *Deviance* = -2 × log Likelihood = measure of goodness of fit
  - *Complexity* : gives a *penalty* for including more parameters

Choosing a model

- Compare models using F-tests, AIC, BIC
- If the number of variables is small enough, could *compare all possible models*
- Usually this is not practical, use *automatic procedures*
    - forward selection
    - backward elimination
    - stepwise selection

Marginality restriction

- Lower order terms are *marginal* to higher order terms
- Need to keep terms in the model that are marginal to other terms
    - if include *polynomial* term *e.g.* x^2, need to also keep x in the model
    - if include *interaction* term, need to keep all primary variables and lower order interactions in the model

Model (variable) selection procedures I

- *Forward Selection*
  - start with *no* variables in the model
  - in successive steps, add in the 'best' unselected variable/term
  - stop when have the best model according to the chosen criterion, *e.g.* F, AIC, BIC
- *Backward Elimination*
  - start with *all* variables/terms in the model
  - in successive steps, take out the 'worst' included variable/term
  - stop when have the best model according to the chosen criterion, *e.g.* F, AIC, BIC

Model (variable) selection procedures II

- *Stepwise Selection*
  - start with the *full* model
  - use Backward Elimination to see if any term can be removed
  - use Forward Selection to see if a term can be added
  - iterate (Backward - Forward - Backward - *etc.*)
  - stop when model doesn't change

Selection procedures : problems

- The methods are *automatic*
  - do not take into account scientific knowledge
  - do not take effect size into account – can include a significant variable with an effect size that is not interesting or important
  - can lead to model that are not meaningful or unrealistic
- Not guaranteed to find the optimum
  - Stepwise : try multiple times, starting with a different model each time
- *All models are wrong, but some are useful*

HOWTO : Model Selection

- Use scientific/problem-specific knowledge to suggest important variables/terms for potential inclusion
- Then, can try automatic procedures (stepwise selection, F-tests, *etc.*)
- Observe marginality
- If you use F-tests/ANOVA tables, remember that the order of inclusion of variables matters – try different orders
- Better to use `stepAIC` function in the R package `MASS`
- (see handout, Section 6.8 in the MASS book)

Model assessment

- Important model *assumptions* :
  - Independent observations
  - Normally distributed errors
  - Constant error variance
  - Additive effects
- If the assumptions do not hold (at least approximately), then the results of the analysis will generally not be meaningful
- $\Rightarrow$ **Check assumptions ! !**

Testing submodels

- Full model : (Ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_p$
- Submodel : (ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_q,\ q < p$
- $H : \beta_{q+1} = \cdots = \beta_p = 0$ vs. A : at least 1 $\beta_i \neq 0,\ q+1 \leq i \leq p$

ANOVA table

source	df	SS	MS (=SS/df)
ω	q	$SSM(\omega)$	SSM/q
suppl. terms	$p-q$	$SSE(\omega) - SSE(\Omega)$	$(SSE(\omega) - SSE(\Omega))/(p-q)$
errorr	$n-p-1$	$SSE(\Omega)$	$SSE(\Omega)/(n-p-1)$
total (corr.)	$n-1$	SST	

- The F-statistic for testing the significance of the extra terms in Ω is :

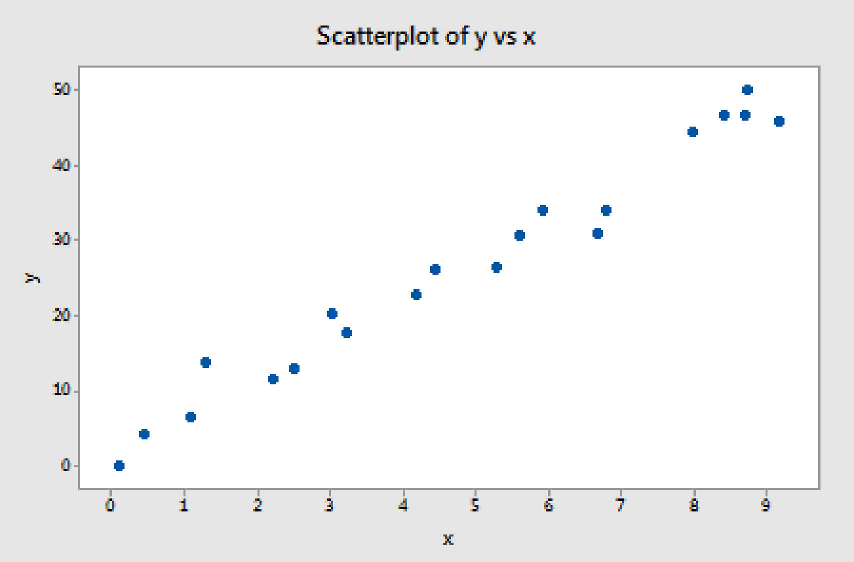
$$F_{obs} = \frac{(SSE(\omega) - SSE(\Omega))/(p-q)}{SSE(\Omega)/(n-p-1)} \sim F_{p-q,n-p-1} \text{ under } H$$

- We REJECT H when $F_{obs} > F_{p-q,n-p-1}(1-\alpha)$

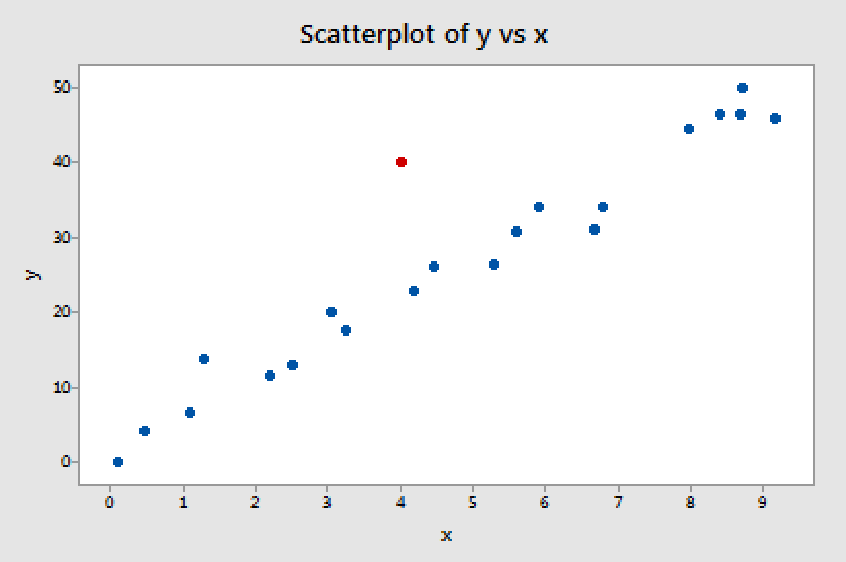
Another regression estimation output

```
> trees.fit1 <- lm(Volume ~ Diameter, trees.dat)
> summary(trees.fit1)

Call:
lm(formula = Volume ~ Diameter, data = trees.dat)

Residuals:
   Min     1Q Median     3Q    Max 
-8.065 -3.107  0.152  3.495  9.587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Diameter      5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared: 0.9353,     Adjusted R-squared: 0.9331 
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

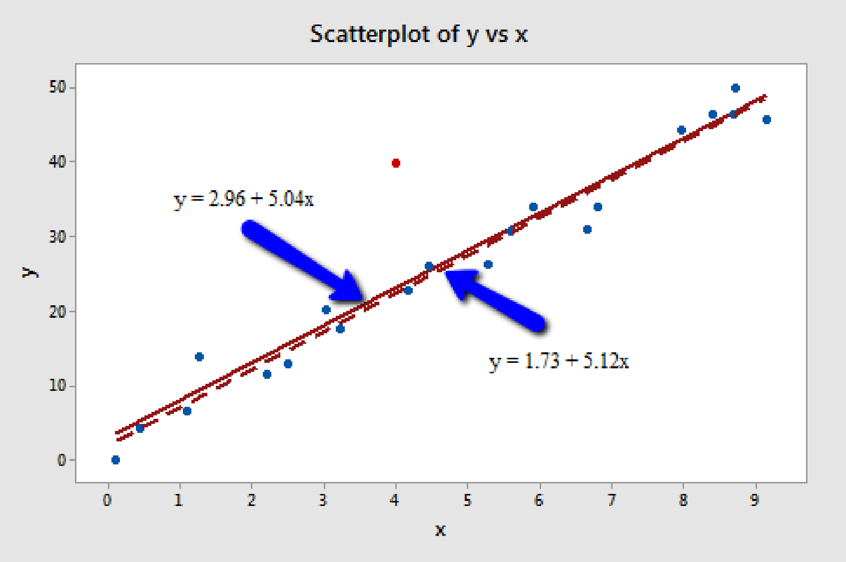
Influential points : example 1



Influential points : example 2



Influential points : example 2 without red point



Influential points : results comparison 2

With red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.71075	91.01%	90.53%	89.61%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.96	2.01	1.47	0.157	
x	5.037	0.363	13.86	0.000	1.00

Regression Equation

y = 2.96 + 5.037 x

Without red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

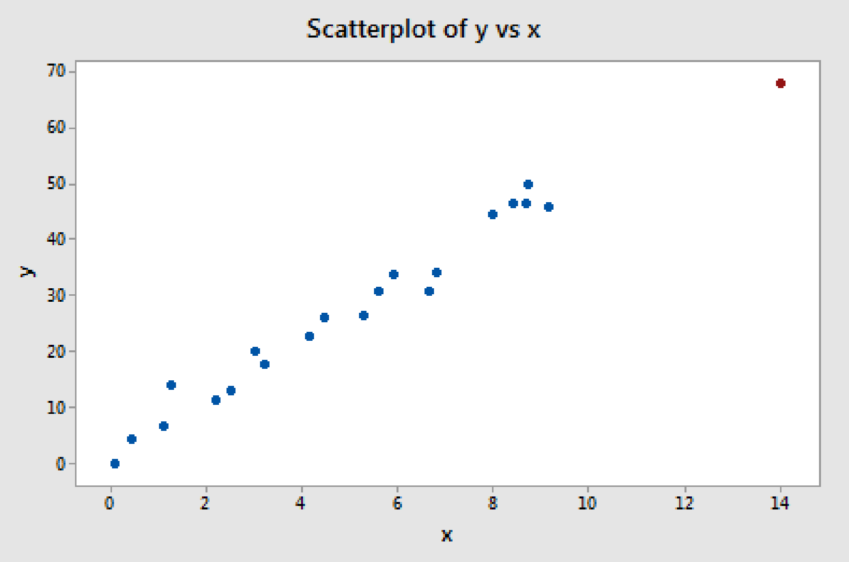
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

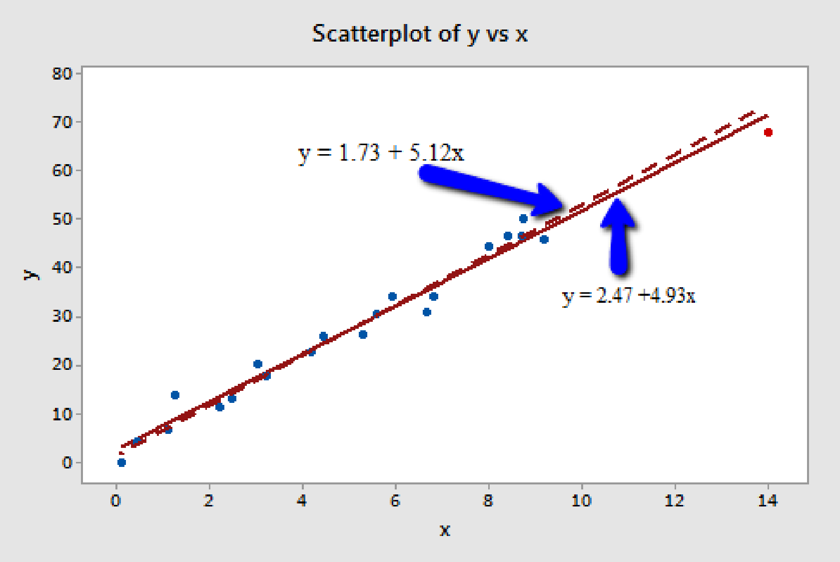
Regression Equation

y = 1.73 + 5.117 x

Influential points : example 3



Influential points : example 2 without red point



Influential points : results comparison 3

With red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.71075	91.01%	90.53%	89.61%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.96	2.01	1.47	0.157	
x	5.037	0.363	13.86	0.000	1.00

Regression Equation

y = 2.96 + 5.037 x

Without red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

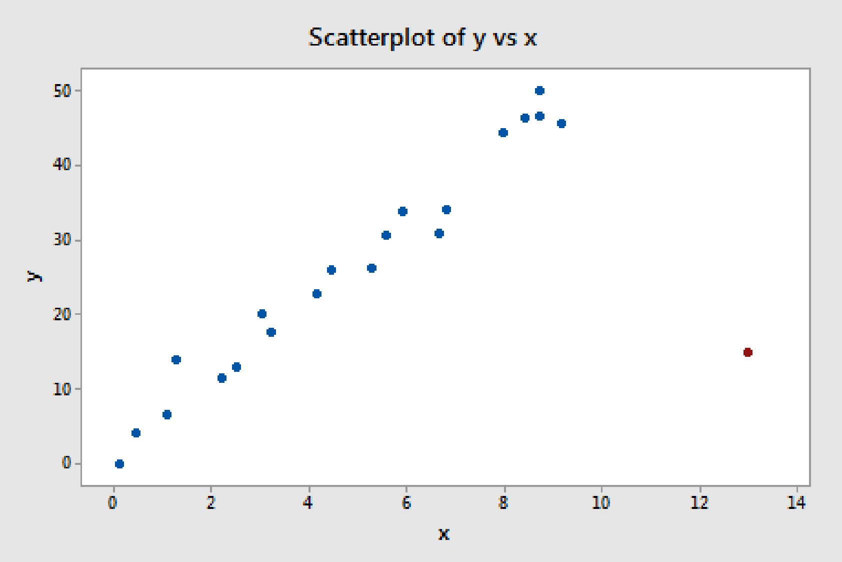
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

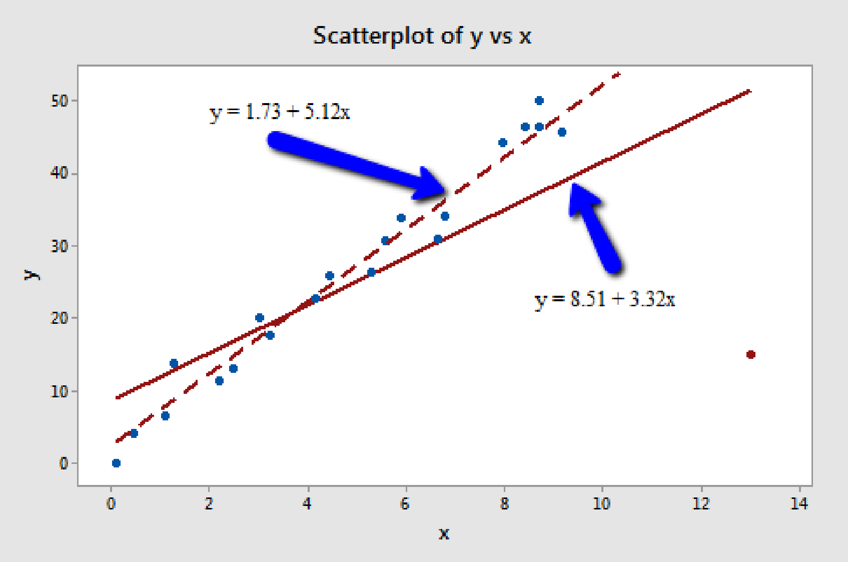
Regression Equation

y = 1.73 + 5.117 x

Influential points : example 4



Influential points : example 2 without red point



Influential points : results comparison 4

With red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.71075	91.01%	90.53%	89.61%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.96	2.01	1.47	0.157	
x	5.037	0.363	13.86	0.000	1.00

Regression Equation

y = 2.96 + 5.037 x

Without red point:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

Regression Equation

y = 1.73 + 5.117 x

Leverage

- We can write the OLS prediction for y as $\hat{y} = Hy$, where H is the 'hat matrix' $(X'X)^(-1)X'$
- Each predicted response can be written as $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n, \ i = 1, \ldots, n$
- Therefore, the leverage h_{ii} quantifies the *influence* that the *observed response* y_i has on its predicted value y_i
- The leverage depends only on the predictor values x_{ij}
- Whether the data point is influential or not *also depends* on the observed value of the reponse y_i

Outliers

- One way to identify ($y-$) outliers by considering *standardized residuals* :
$$r_i = \frac{e_i}{SE(e_i)} = \frac{y - \hat{y}}{\sqrt{MSE(1 - h_{ii})}}$$
- Thus, the standardized residuals are represented in the number of standard deviations away from the mean
- Some might consider points whose standardized residual r_i larger than 2 or 3 to be *outliers*

Studentized residuals for identifying outliers

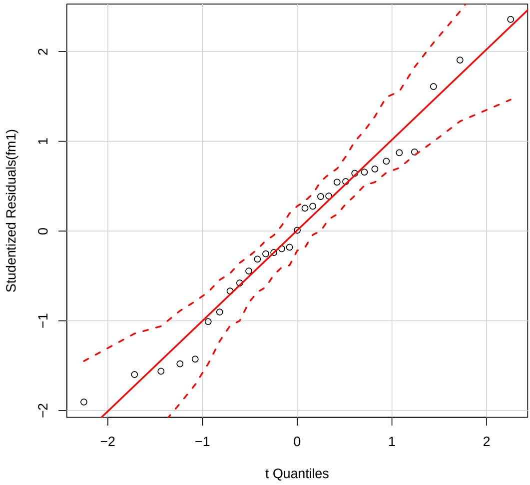
- A better way to identify $(y-)$ outliers by considering *studentized residuals* :

$$t_i = \frac{e_(i)}{SE(e_(i))} = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}},$$

  where $e_(i)$ is the residual obtained when observation i is left out : $y - \hat{y}_{(i)}$

- In general, *studentized* residuals are going to be more effective for detecting outlying Y observations than standardized residuals
- Observation with studentized residual larger than 3 (in absolute value) can be considered as *outliers*

Trees example : studentized residuals



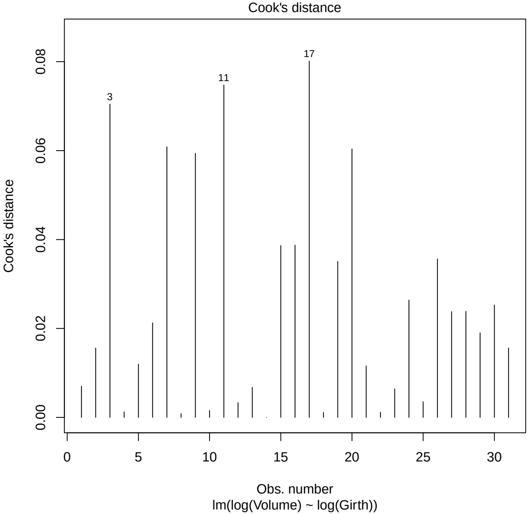
Cook's distance

- Another useful diagnostic is Cook's distance :

$$D_k = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{i=1}^{n} (\hat{y}_{i(k)} - y_i)^2$$

- These values assess the impact of the kth observation on the *estimated regression coefficients* $\hat{\beta}_i$
- Values of D_k larger than 1 are suggestive that the corresponding observation has undue influence on the estimated coefficients

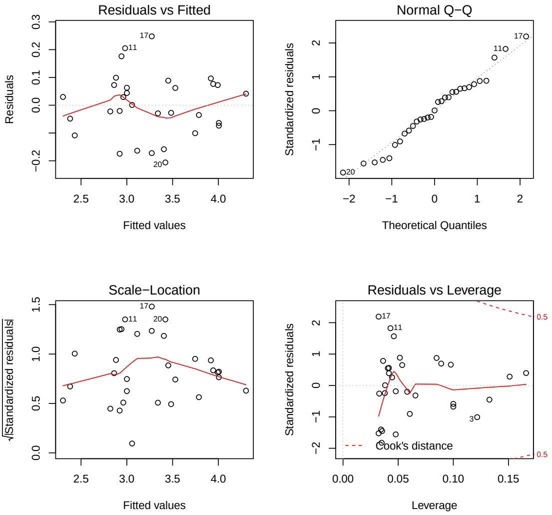
Trees example : Cook's distance



Other diagnostic plots

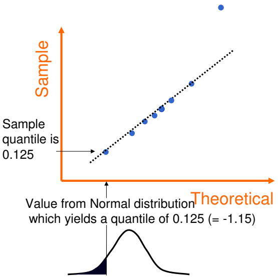
- In addition to the *exploratory plots* you make at the beginning of the analysis, you will also need *additional diagnostic plots* in the model assessment phase
- There should not be any *structure* in the residuals
- Plot residuals against predicted values, variables in the model, variables *not* in the model (*e.g.* to see if some important variable is left out, assess dependence), normal QQ-plot
- Look for outliers, constant variance, patterns, normality

Some diagnostic plots



QQ-plot

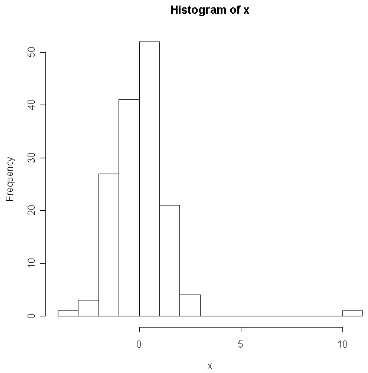
- Quantile-quantile plot
- Used to determine whether a sample follows a particular distribution (*e.g.* normal) or to compare 2 samples
- A graphical method for the identification of outliers when the data are approximately normal

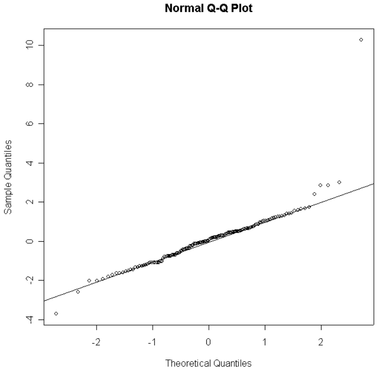


Typical deviations from a straight line

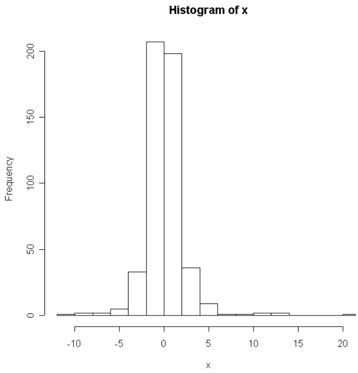
- Outliers
- Curvature at both extremes (long or short tails)
- Convexe/concave curvature (asymmetry)
- Horizontal segments (discretization)

Outliers

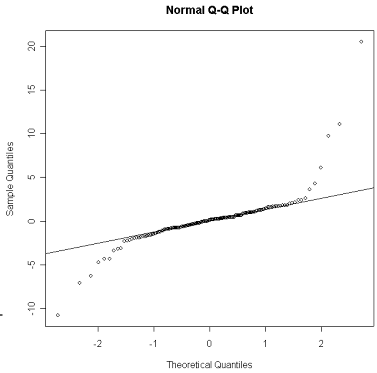

Histogram of x
Frequency
0
10
20
30
40
50
0
5
10
x


Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

Long tails

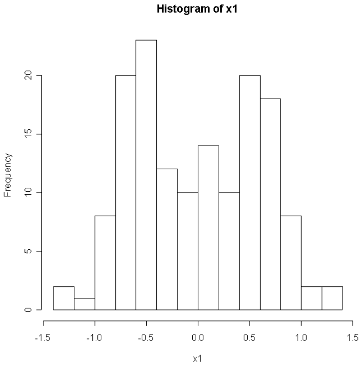


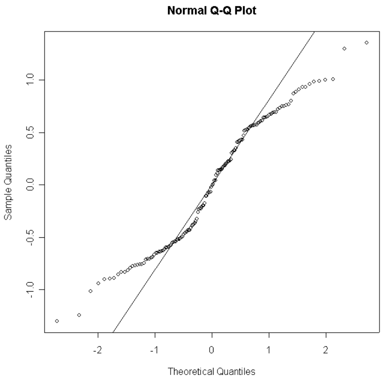
Histogram of x
Frequency
0
50
100
150
200
-10
-5
0
5
10
15
20
x



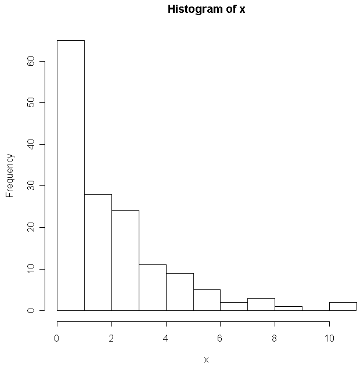
Normal Q-Q Plot
Sample Quantiles
-10
-5
0
5
10
15
20
-2
-1
0
1
2
Theoretical Quantiles

Short tails

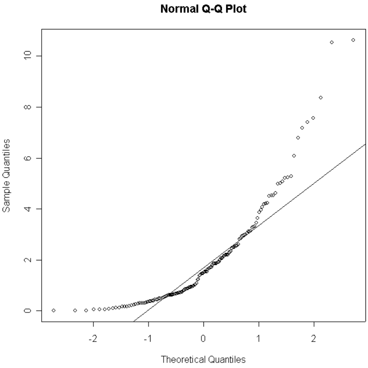

Histogram of x1
Frequency
x1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2

Asymmetry

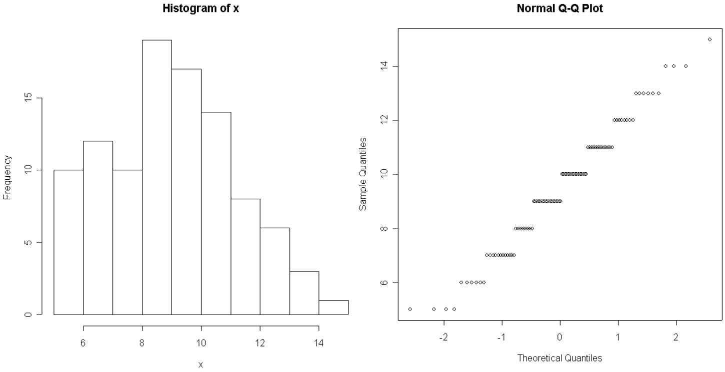


Histogram of x
Frequency
0
10
20
30
40
50
60
0
2
4
6
8
10
x



Normal Q-Q Plot
Sample Quantiles
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

Discretization



Dealing with problematic data points

- Check for obvious errors and correct them
- Consider the possibility that you might have misformulated your regression model : do you need additional predictors or interaction terms ?
- Decide whether or not deleting data points is warranted – BUT : must have *objective* reason
- If you do delete any data after you've collected it, *justify and describe it* in your reports
- If you are not sure what to do about a data point, analyze the data twice – once with and once without the data point – and report the results of both analyses
- Use *common sense* and *knowledge* about the specific context

Pitfalls in regression

- Regression effect/regression fallacy
    - It is unlikely to have a very high/low value in X
    - The associated Y value is more likely to be closer to the mean ('regression toward the mean')
    - The *regression fallacy* consists in thinking that this *regression effect* needs a special theory to explain it
- Correlation is not *causation*
- Extrapolation – relation may not continue to hold outside the range where it is estimated
- Nonlinearity
- Missing variables, confounding