# STATISTICAL MODELS FOR CLINICAL AND HEALTH DATA

V. Monbet

[1] Université de Rennes/UFR Mathématiques

# Multiple Linear Regression

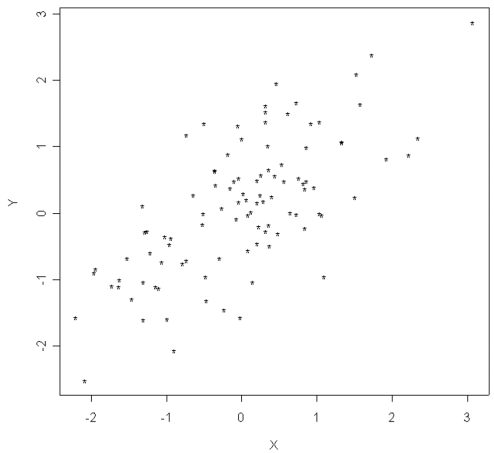
- Multiple linear regression
- Confidence intervals for a coefficient
- Prediction interval for a new observation
- Model selection
- Influential points
- Diagnostics for model assessment

Source : Goldstein, EPFL

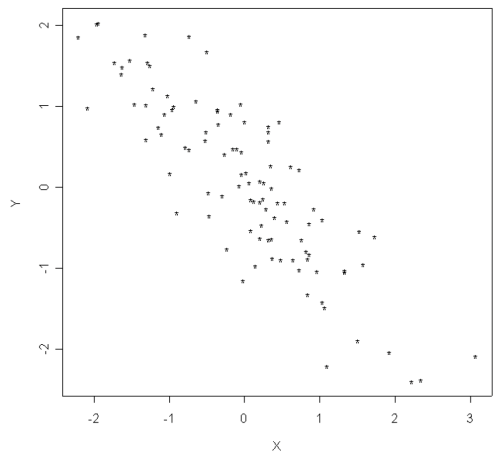
# Scatterplot

- 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)

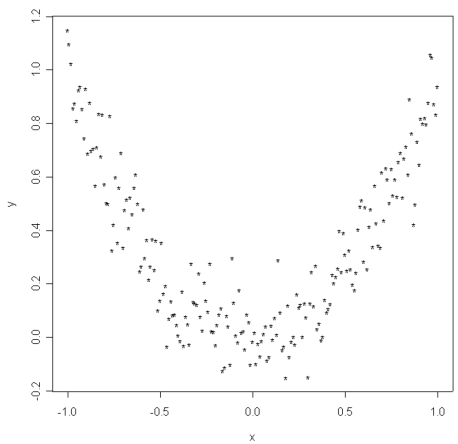
Positive ass.




Negative Ass.




Non Lin. Ass.



# Multivariate data

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

vector of means : $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$

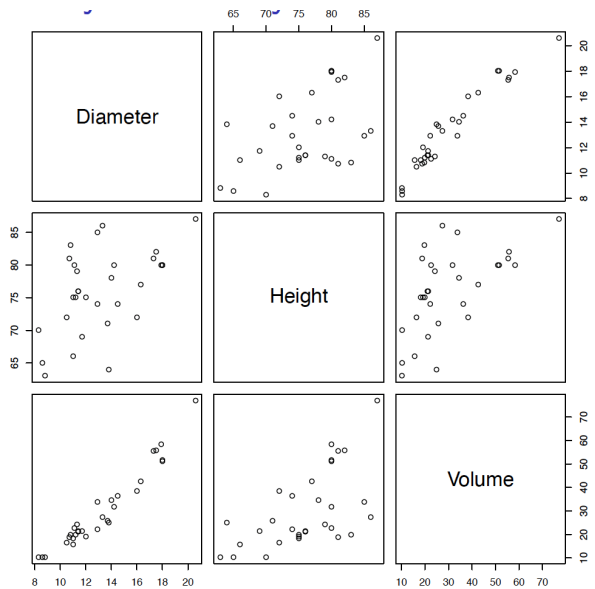
matrix of variance-covariances

# 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



# Multiple regression

- The model

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

- Technique to find estimates $\hat{\beta}_j,\ j = 1, \cdots, p$, solve the Least Square optimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}\mathbf{X}_i\right)^2$$

- Result $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$
where $\mathbf{X}$ is the design matrix.

# 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)

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
```

équation → y ($x_1$, $x_2$)

$\hat{\beta}_0$ → (Intercept)

$\hat{\beta}_1$ → Diameter

$\hat{\beta}_2$ → Height

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

# Interpretation of regression coefficients

```
> 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
```

équation → $y$ $x_1$ $x_2$

$\hat{\beta}_0$ → (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 $\beta_1, \cdots, \beta_p$ corresponds to the contribution of one variable when all other variables in the equation are held constant
  - the coefficient $\beta_0$ is the predicted value of $y$ when all variables $x_1, \cdots, x_p = 0$

# OLS properties : expected value

When

- $E(\epsilon_i) = 0,\ i = 1, \cdots, n$;
- $Var(\epsilon_i) = \sigma^2$ (constante) ;
- $Cov(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

one has :

$$\begin{aligned} E(\widehat{\beta}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(y) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= \beta \end{aligned}$$

The OLS estimator is unbiaised.

# OLS properties : expected value

$$
\begin{aligned}
Var(\widehat{\beta}) &= Var(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(y)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbb{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
$$

Note that $(\mathbf{X}^T\mathbf{X})$ is 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

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
```

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

$\hat{\sigma}$ (s)

n-p-1

# Tests/confidence intervals for the coefficients

In addition, assuming $\epsilon_1, \cdots, \epsilon_n$ iid $\simeq \mathcal{N}(0, \sigma^2)$ we have

$$\widehat{\boldsymbol{\beta}} \simeq \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

Thus, $Var(\hat{\beta}_j) = \sigma^2[\mathbf{X}^T\mathbf{X})^{-1}]_{j+1,j+1}$

A confidence interval (CI), with confidence level $100(1-\alpha)\%$ for $\beta - j$ takes the form

$$\hat{\beta}_j \pm \hat{\sigma}\sqrt{[\mathbf{X}^T\mathbf{X})^{-1}]_{j+1,j+1}}t_{n-p-1,1-\alpha/2}$$

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

$$S_{obs} = \frac{\beta_j}{\sigma\sqrt{[\mathbf{X}^T\mathbf{X})^{-1}]_{j+1,j+1}}}$$

H is rejeted 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 (PI) for a new (single) observation with $x = x_0$ is given by :

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

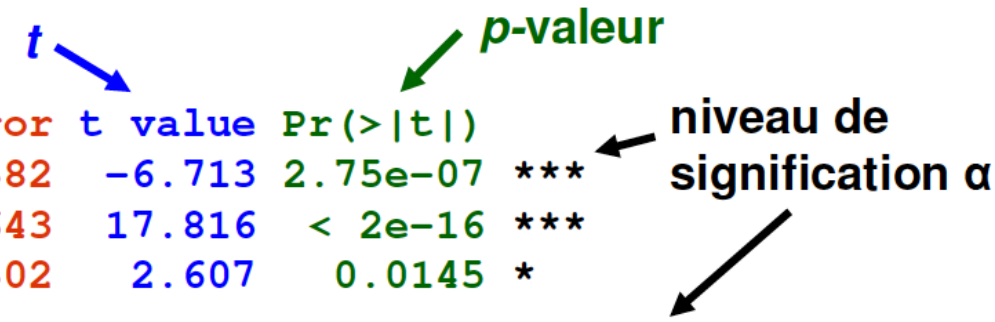
A PI is wider than a CI for a given level.

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
```



# Analysis of variance table (ANOVA)

Uses the Pythagorean theorem

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

to partition the total sum of squares (SST).

It can also be written

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

We can present this equality in the form of a table (ANOVA table)

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

# F test

The statistic $F_{\text{obs}} = MS/(source - MSE)$ tests the hypothesis
H :$\beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs. A at least 1 $\beta_i \neq 0$

The distribution of $F_{\text{obs}}$ when H is true is the Fisher distribution $F_{p;n-p-1}$

The numerator of $F_{\text{obs}}$ is the variability explained by the regression model

The denominator contains the residual variance

Under the null, the expected value of $F_{\text{obs}}$ is 1 and under the alternative the expected value is bigger than 1
then REJECT the null hypothesis H for large values of $F_{\text{obs}}$.

When testing a single coefficient $F_{1;n-1} = t^2_{n-1}$

# 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 :

- variance explained by the regression, and
- 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 $R^2_{aj}$ than $R^2$

$R^2_{aj}$ is not a true "square" - it can even take on negative values

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

# 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-test 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 $= -2 \times$ log Likeliohood = 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

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

# Selection procedures : problems

The methods are automatic

- do not take into account scientific knowledge
- do not take effect size into account - can include a signiffcant variable with aneffect 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 assesment

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 **Check assumptions ! !**

# Testing submodels

Full model : $(\Omega) : y = \beta_0 + \beta_1 + \cdots + \beta_p$

Submodel model : $(\Omega) : y = \beta_0 + \beta_1 + \cdots + \beta_q$, $q < p$

H : $\beta_{q+1} = \cdots = \beta_p = 0$ vs. A : atleat one $\beta_j \neq 0$, $q + 1 \leq j \leq p$

**ANOVA table**

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

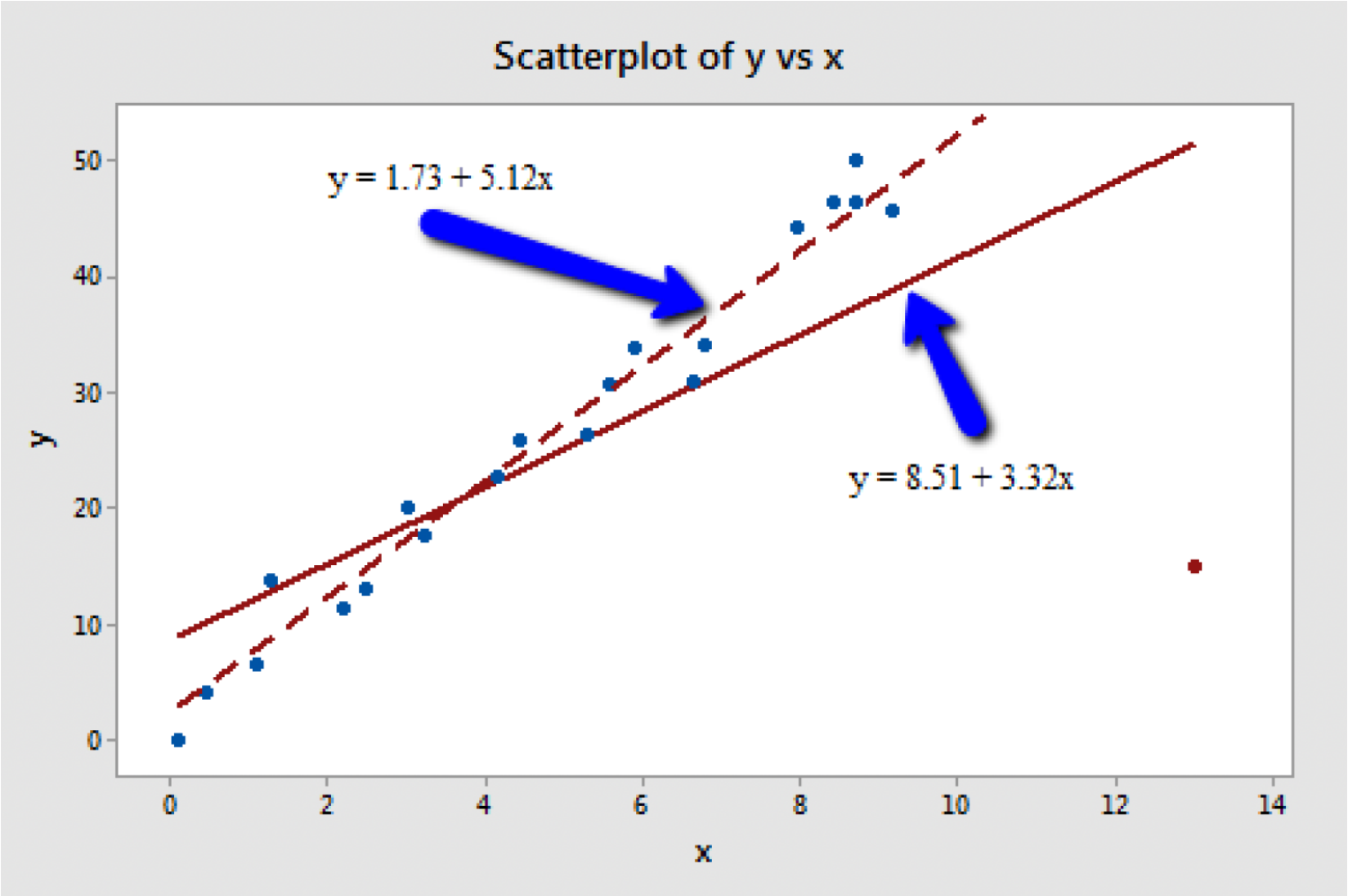
The F-statistic for testin the significance of the estra terms in $\Omega$ is

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

We REFECT H when $F_{\text{obs}} > F_{p-q;n-p-1}(1 - \alpha)$

```
> 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
```

# Influencial points



# Studentized residuals for identifying outliers

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

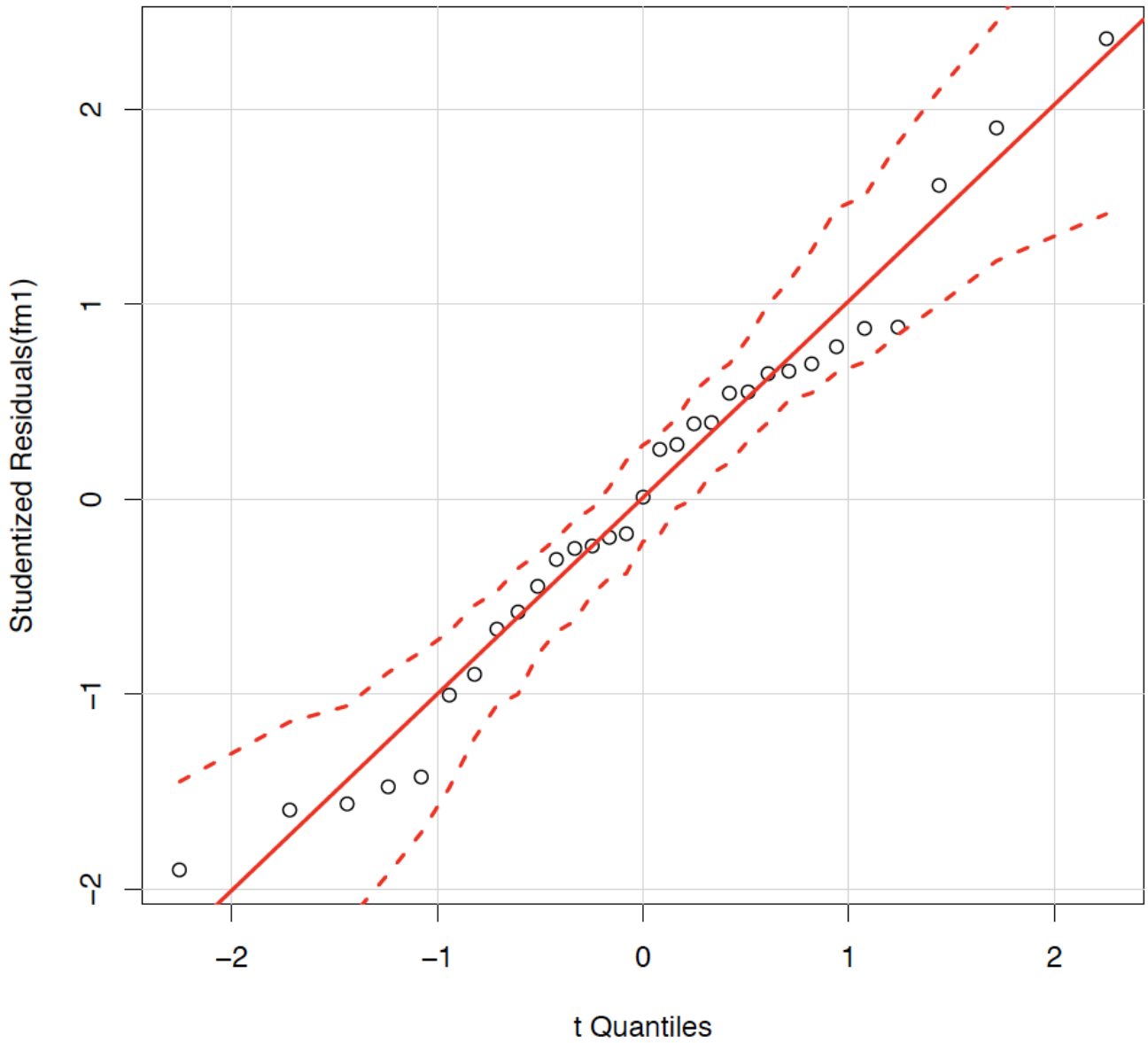
$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

where $H$ is the hat matirx $(X^T X)^{-1} X^T$ and $MSE_{(i)}$ is the mean square error computed when point $i$ is not used for fitting.

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

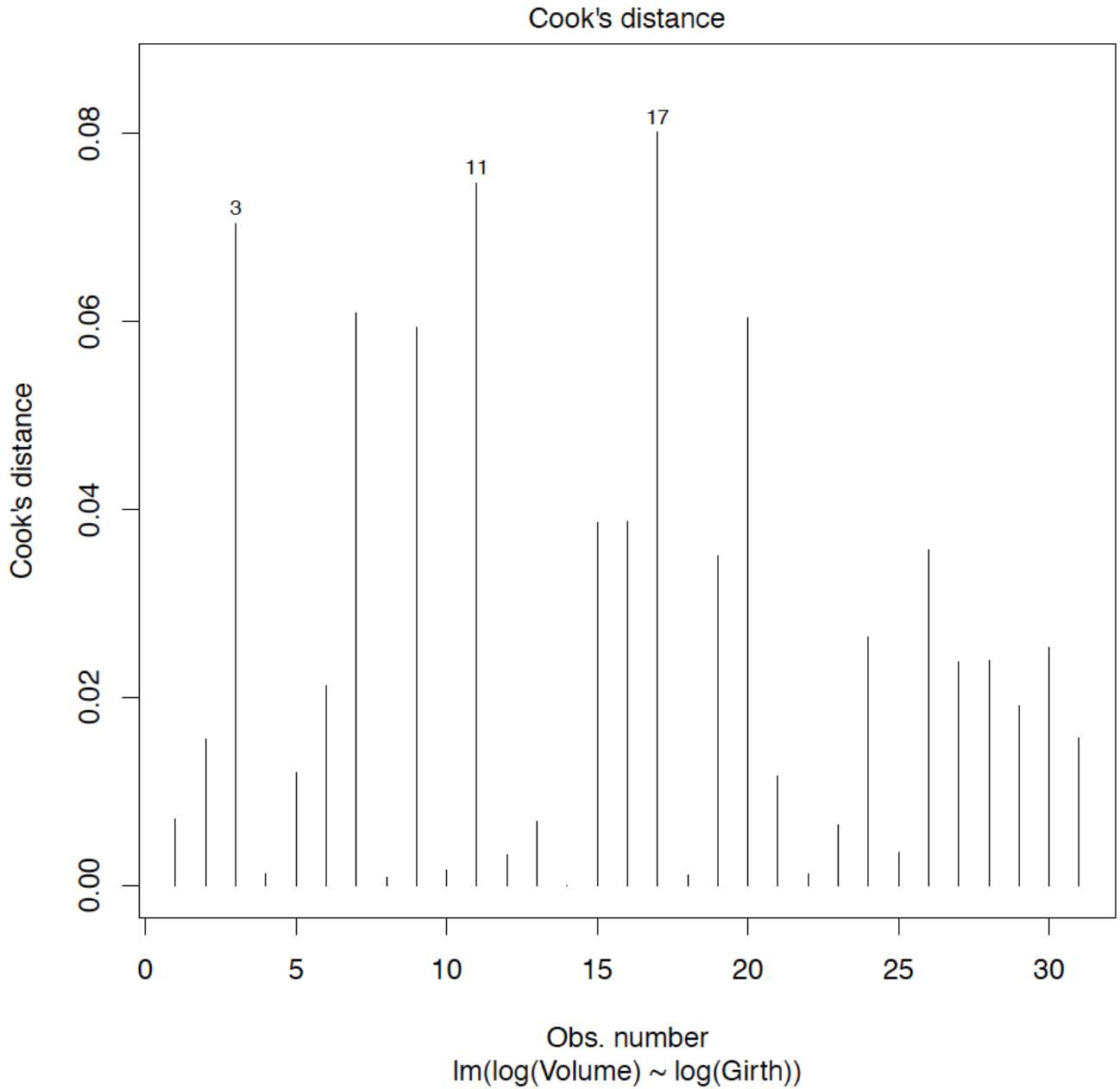
Another useful diagnostic is Cook's distance

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

These values assess the impact of the kth observation on the estimated regression coefficients $\hat{\beta}_j$ je ne comprends pas : a voir

Values of $D_k$ larger than 1 are suggestive that the corresponding observation has undue influence on the estimated coeffiients

# Trees example : Cook's distance



# Other diagnostic plots

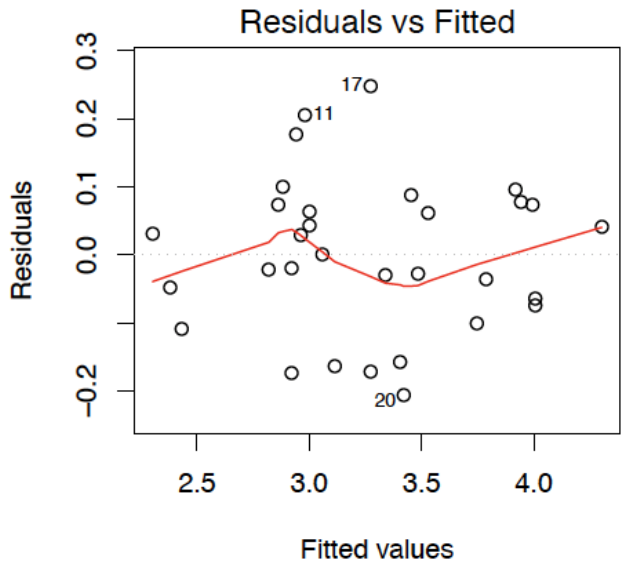
In addition to the exploratory plots you make at the beginning of the analysis, you will also need additional diagnostic plots in the model assessment phase

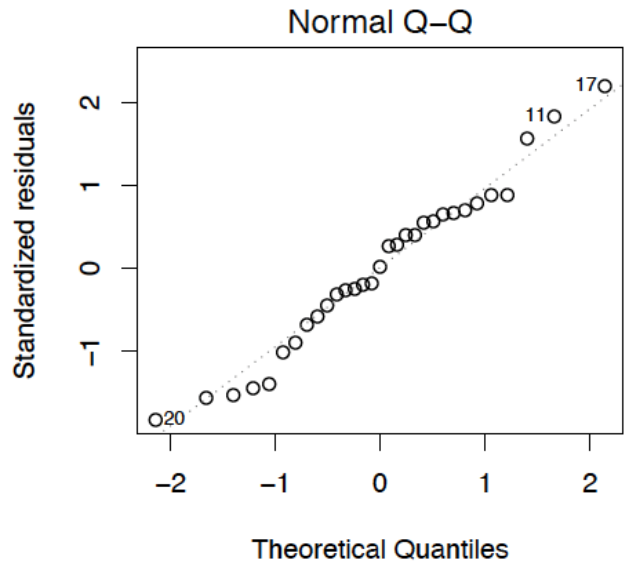
There should not be any structure in the residuals

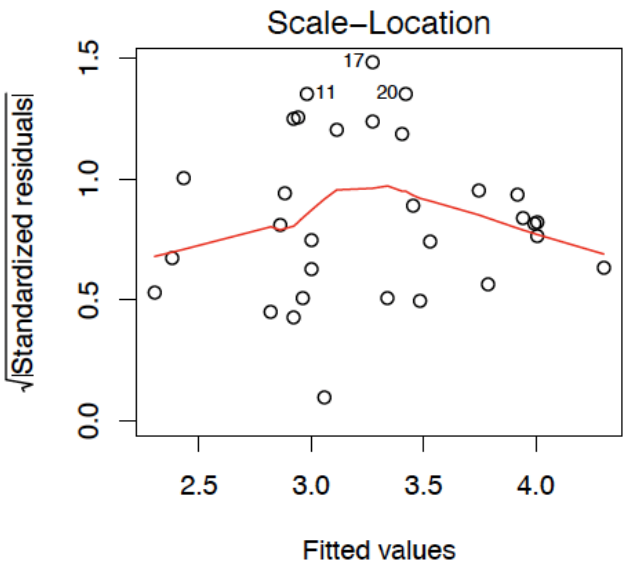
Plot residuals against predicted values, variables in the model, variables not in the model (e.g. to see if some important variable is left out, assess dependence), normal QQ-plot

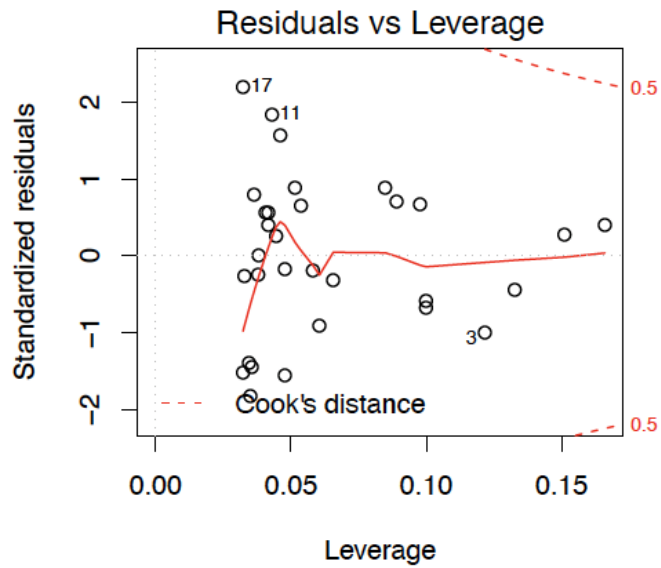
Look for outliers, constant variance, patterns, normality

# Diagnostic plots



# 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