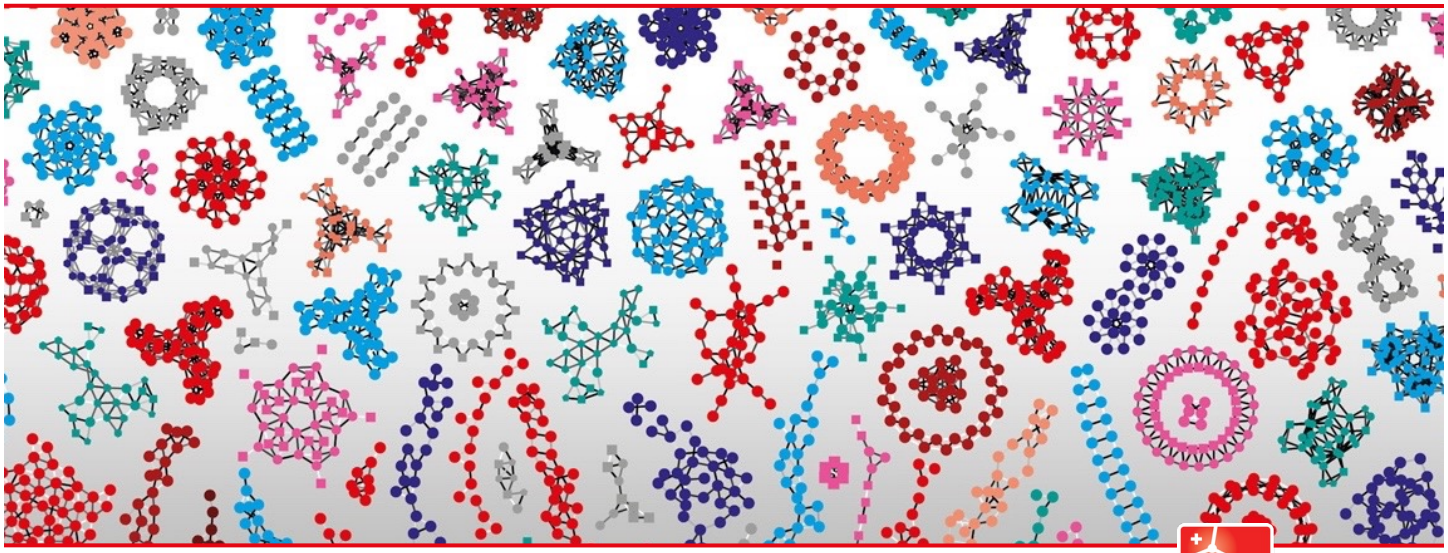


SIB Swiss Institute of Bioinformatics

Advanced statistics: Statistical modeling 2023

Rachel Marcone (Jeitziner) and Mauro Delorenzi

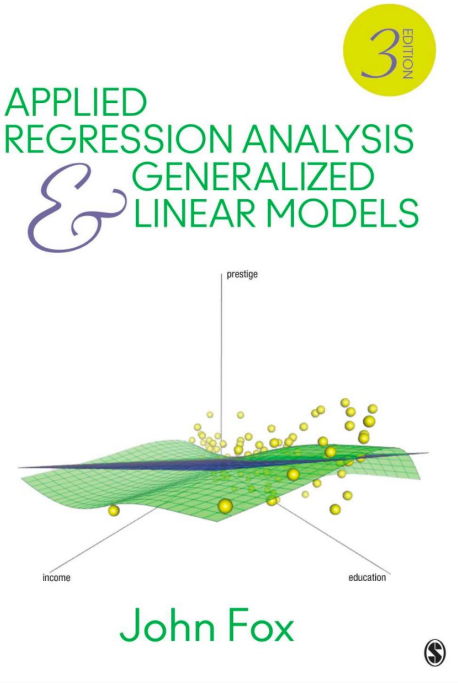
Slides credit also to Linda Dib, Frédéric Schütz, Isabelle Dupanloup a.o.

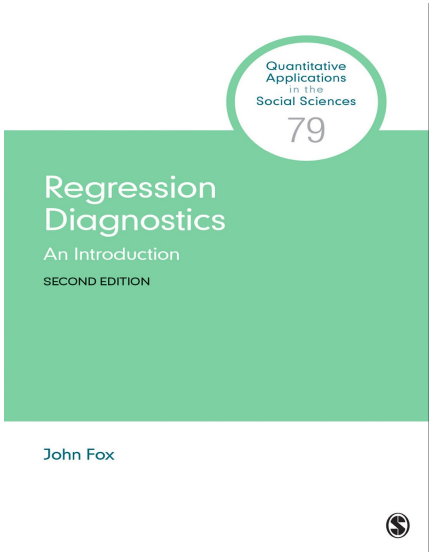
Advanced statistics: Statistical modeling

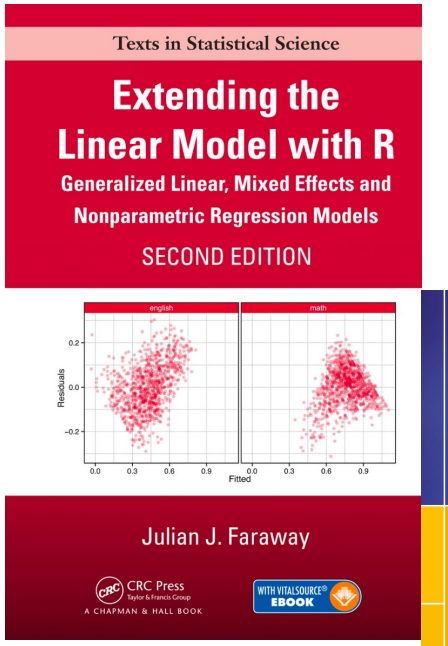
- Introductory statistics course: models and tools (such as linear regression) to analyze “simple” datasets (not appropriate for all types of data)
- Goal of the course: learn beyond classical linear modelling
- Program of the course:
  - Review of the basics of linear regression (**LM**), uni- and multi-variable
  - Extensions of **LM**: complex functional relations (non-linear), polynomial / spline regression
  - Generalized linear models (**GLMs**) : logistic / Poisson regression

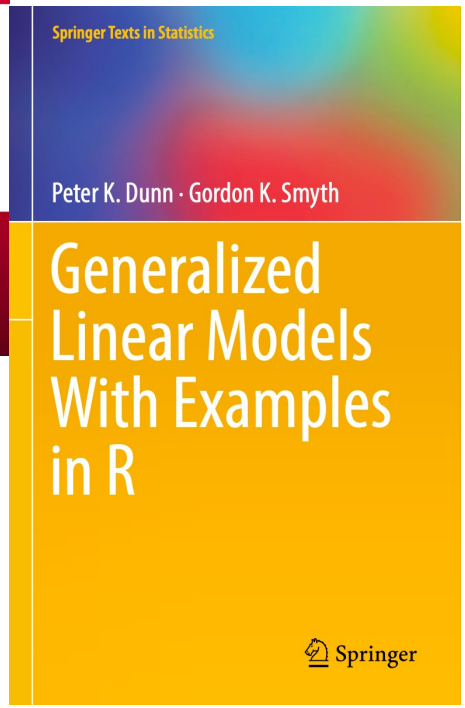
  - Mixed-effects models (fixed and random effects components)
  - Analysis of longitudinal data (application of mixed-effects models)
  - Generalized Additive Models (**GAMs**)

Books


3 EDITION
APPLIED
REGRESSION ANALYSIS
& GENERALIZED
LINEAR MODELS
prestige
income
education
John Fox

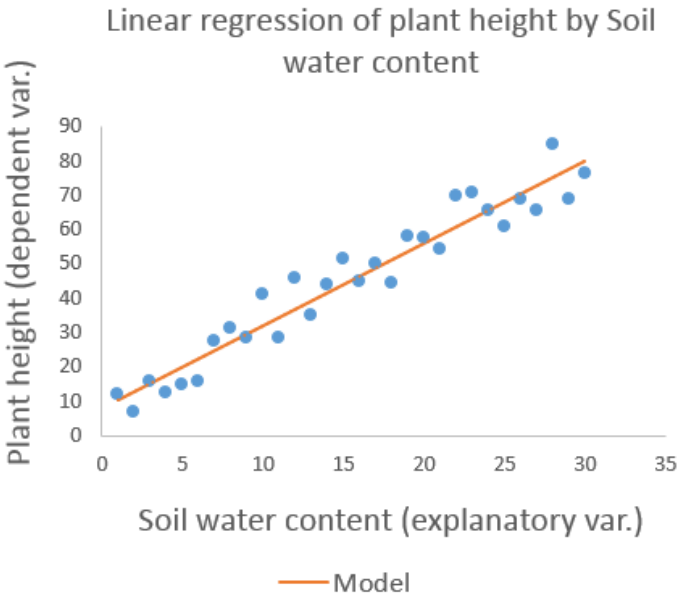

Quantitative Applications in the Social Sciences
79
Regression Diagnostics
An Introduction
SECOND EDITION
John Fox


Texts in Statistical Science
Extending the Linear Model with R
Generalized Linear, Mixed Effects and Nonparametric Regression Models
SECOND EDITION
english
math
Residuals
Fitted
Julian J. Faraway
CRC Press
A CHAPMAN & HALL BOOK
WITH VITALSOURCE EBOOK


Springer Texts in Statistics
Peter K. Dunn · Gordon K. Smyth
Generalized Linear Models With Examples in R
Springer

Statistical models

What is a statistical model ?



Modeling:

- process of developing / applying mathematically-formalized way to represent certain aspects of "reality" (the machinery that generates the data),
- in a simplifying approximate fashion,
- in order to describe and "understand" certain relations and (potentially) to make predictions from the model about future events

Statistical:

Based on principles and methods developed in statistical / data analysis sciences

What is a statistical model ?

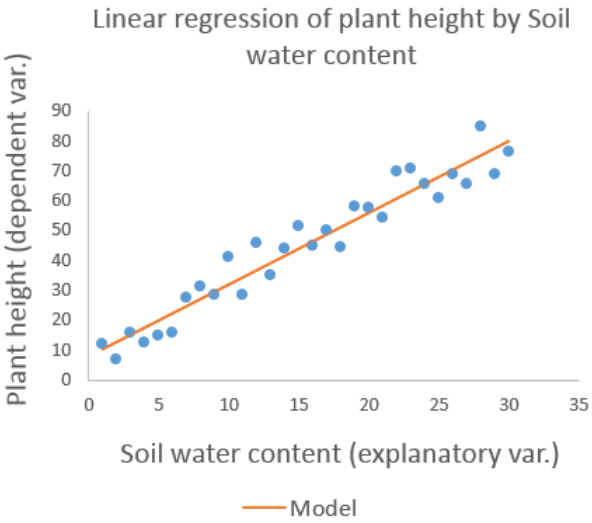
A **statistical model** is a set of equations involving random variables, with associated distributional assumptions,

devised in the context of a **question** and a body of **data concerning some phenomenon**,

with which **tentative answers** can be derived, along with **measures of uncertainty** concerning these answers.

questions + ***data*** $\xrightarrow[\textbf{\textit{model}}]{}$ ***answers*** + ***measures of uncertainty***

Height =
intercept + slope * **soil water content**



Type / Role of variables:

Dependent variables (or responses): variables we want to describe, understand, explain, model, predict

Explanatory variables
(or independent variables or predictors or covariates):
variables we use to explain, to describe or to predict the dependent variable(s)

Both variables may be quantitative or qualitative

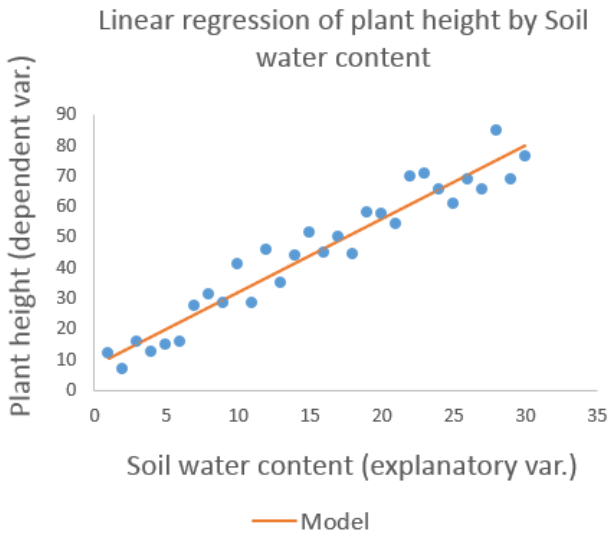
What is a model parameter ?

Statistical model: refers to the equations used with quantities called model parameters

"Model": includes or not a specific set of values estimated for the parameters

Statistical modeling

1. Estimation of model parameter
2. Prediction of the dependent variable(s)

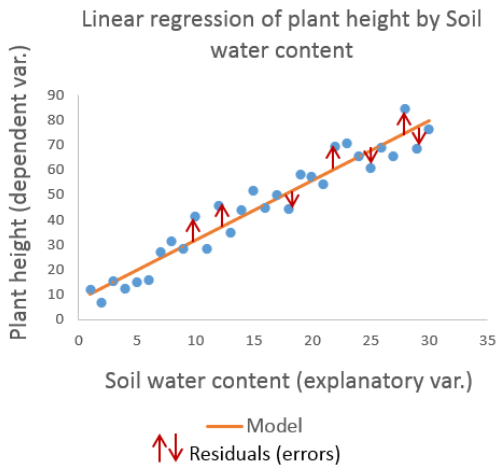


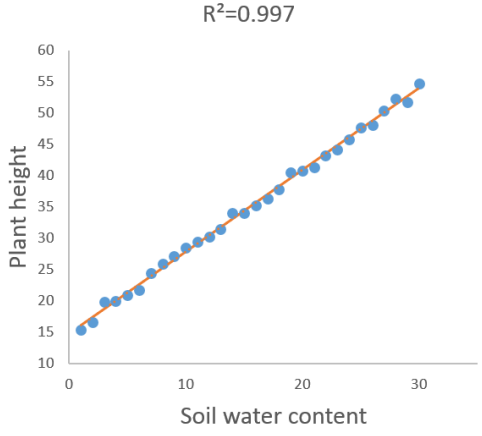
Height =
intercept +
slope * soil water content

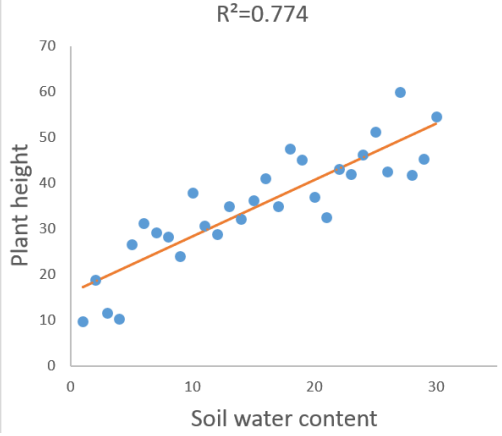
What is a model residual ?

Model residuals (or “errors”): distances between data points and the expected values based on the model (equation with fitted parameters)

Model residuals represents the part of variability in the data the model was unable to capture







Modeling overview

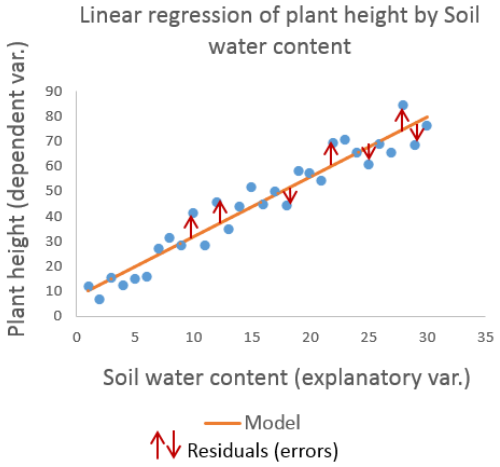
Want to capture important features of the relationship between a (set of) variable(s) and one or more response(s)

Many simple models are of the form

$$Y = f(x) + \text{error} \quad , \quad \text{or } g(Y) = f(x) + \text{error}$$

with differences in the form of g and f
and distributional assumptions about the error term

Modeling overview revised



Given a response g(Y) or **Y**
that (might) depend on a variable **X**

X → ? → individual values Y_i

X → E(Y | X) the expected value for Y
given a value for X ("conditional on the X") → individual values Y_i

(for the same X we can have several points with different Y values)

Model formulas in R

A simple *model formula* in R looks something like:

```
yvar ~ xvar1 + xvar2 + xvar3
```

Can read **~** as "*described (or modeled) by*".

We could write a model (algebraically) as

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Model formulas in R

By default, an intercept is included in the model – you don't have to include a term in the model formula

If you want to leave the intercept out:

```
yvar ~ -1 + xvar1 + xvar2 + xvar3
```

Amodel with only the intercept (the overall mean)

```
yvar ~ 1
```

Model formulas in R

The generic form is response ~ predictors

The predictors can be numeric or factor

Other symbols to create formulas with combinations of variables (e.g. interactions)

+ to add more variables (a + b)

: to introduce interactions between two terms (a:b)

- to include both interactions and the terms (a*b is the same as a + b + a:b)

- to leave out variables (a*b - a:b is the same as a + b)

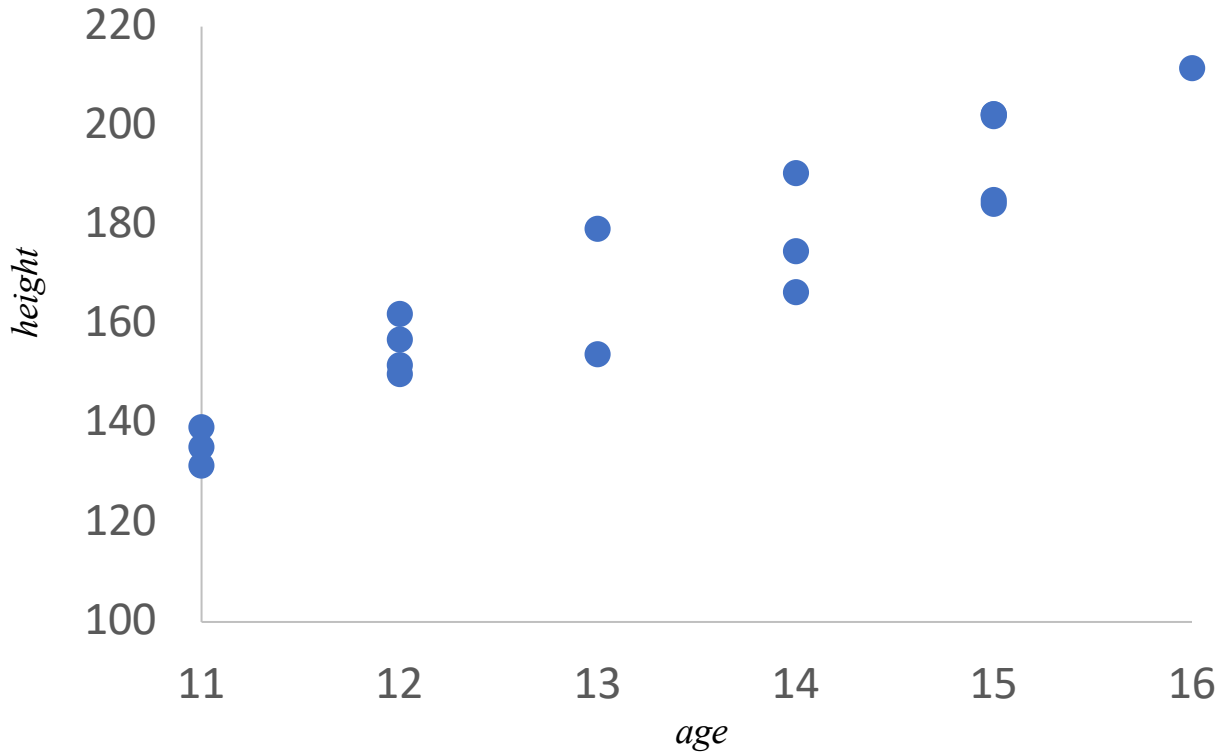
^n to add variables to the power of n

I() treats what's in () as a mathematical expression (a + b versus I(a + b))

Linear models

Can we predict the height of a teenager using his age ?

Example: scatterplot of age vs height in teenagers



(Simple) Linear Regression

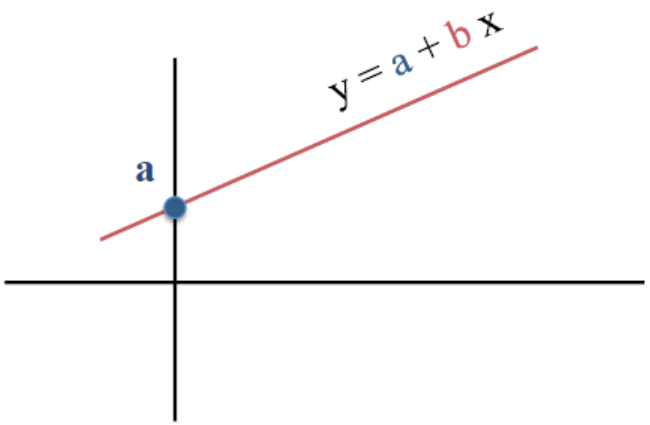
Simple linear regression refers to drawing a (particular, special) line through a scatterplot

It is used for 2 broad purposes: **explanation** and **prediction**.

The equation for a line to predict y knowing x (in slope- intercept form) looks like

$$y = a + b\,x$$

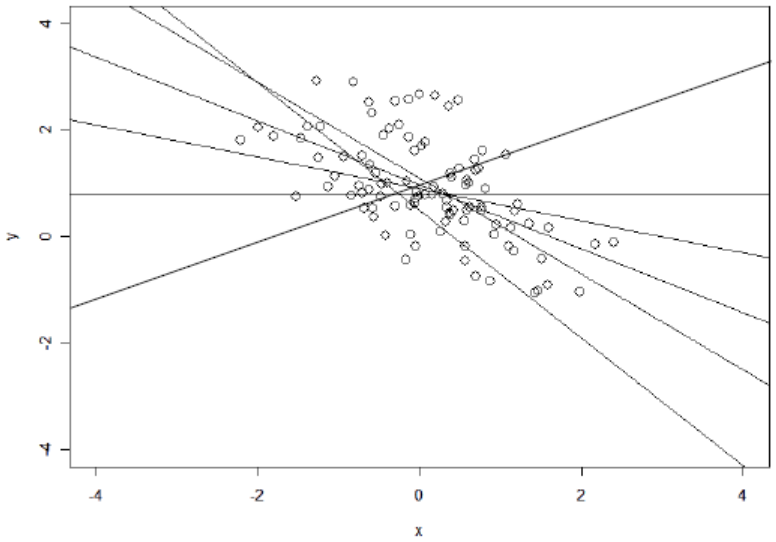
where a is called the intercept and b is the slope.



(Simple) Linear Regression

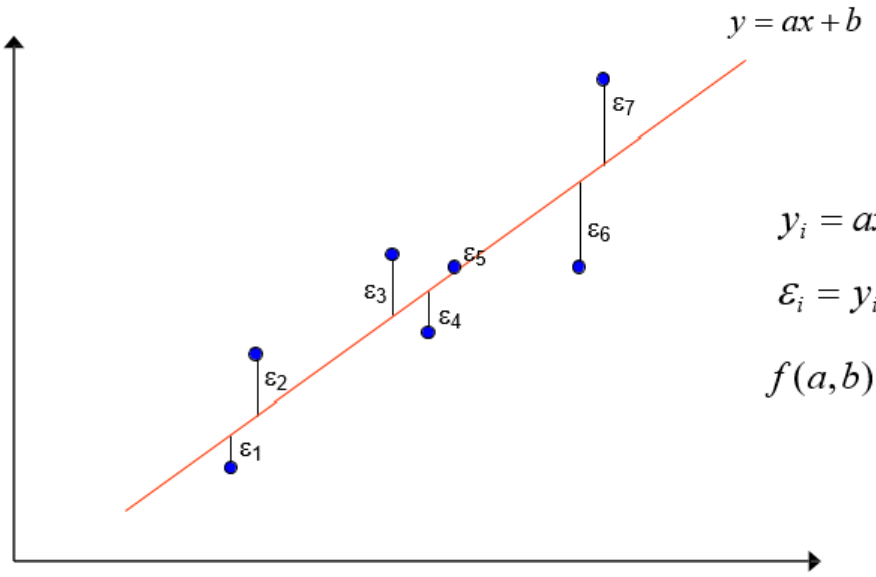
What is the “best” line which fits this data ?

Can we use it to summarize the relation between x and y ?



Linear regression: least-squares fitting

Least-square fitting



Regression line such that:

$$\sum_i \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \ldots$$

minimum

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i = y_i - (ax_i + b)$$

$$f(a,b) = \sum_i \varepsilon_i^2 = \sum_i \left[y_i - (ax_i + b)\right]^2$$

$$\partial f(a,b)/\partial a = 0$$
$$\partial f(a,b)/\partial b = 0$$

The least-squares procedure finds the straight line with the **smallest sum of squares of vertical errors**.

Linear regression: least-squares fitting (LS)

Formalization and extension of linear regression

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

$$i = 1, \cdots, n$$

Y represents **one** data point

Y_i : response (known)

β_0, β_1 : model parameters (estimated)

X_i : predictor (known)

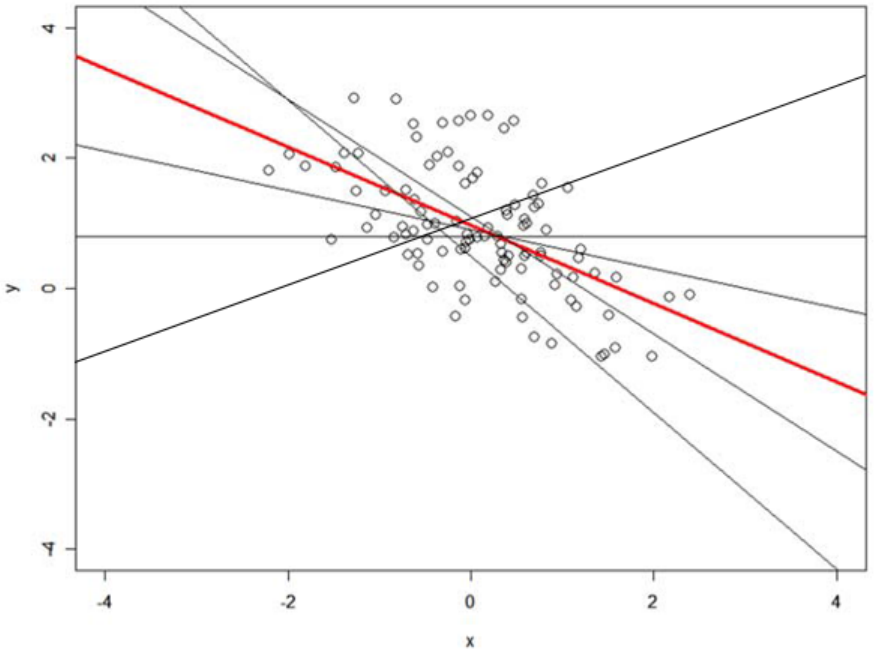
ε_i : error term $\sim N(0, \sigma^2)$ (estimated)

Minimizing $\sum_i \varepsilon_i^2$ yields b_0 and b_1 estimators of β_0 and β_1

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

Linear regression: least-squares fitting (LS)

Over all possible straight lines, y= 1 - 0.6x is the “best” possible line according to this criterion.



y = 0.9 + 0.6x

y = 0.8 + 0x

y = 0.9 – 0.3x

y = 1 – 0.6x

y = 1.1 – 0.9x

y = 0.5 – 1.2x

(Simple) Linear Regression: interpretation of parameters

The regression line has two parameters: the slope and the intercept

The regression slope is the average change in Y when X increases by 1 unit

The intercept is the predicted value for Y when X = 0

If the slope = 0, then X does not help in predicting Y (linearly, in the linear model)

(Simple) Linear Regression: residuals

There is an error in making a regression prediction:

$$\text{error} = \text{observed } Y - \text{predicted } Y = y - (a + bX)$$

These errors are called residuals

The regression equation by LS has this property:
the sum of the residuals is 0 ⇔ the mean of the residuals is 0

Ideally, we want the regression to include all the predictable variance,
so than the distribution of the residuals is “pure random”
and does not depend on X nor on the predicted Y.

Linear models (general case)

p parameter linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i \qquad i = 1, \cdots, n$$

or $$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \qquad \text{with} \quad X_{i0} \equiv 1$$

Y_i response (e.g. expression of a gene)

X_{ik} predictor variables (e.g. dose of drug [continuous], or KO vs wt)

β_k model parameter (measurement of magnitude of effect associated to predictor variable)

ε_i error term (measurement of departure from ideal case)

Linear models: matrix form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$

is equivalent to

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p-1} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Linear models: parameter estimation

Least-square estimation of regression coefficients

$\{\beta_k\}$ such that

$$Q = \sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{ip-1})^2 \quad \text{minimum}$$

$\mathbf{b} = (b_0 \cdots b_{p-1})'$ estimator of $\boldsymbol{\beta}$ is computed as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y} \qquad E\{\boldsymbol{\varepsilon}\} = \mathbf{0}$$

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

Linear models: parameter estimation

Vector of fitted values = $\hat{Y}$ = $E[Y \mid X]$

Matrix notation = $\hat{Y} = X\beta = X\,(X'X)^{-1}X'\,Y = H\,Y$

$H = X\,(X'X)^{-1}X'$ ***is called the hat matrix***

The diagonal values h_{ij} ***of the matrix are good indicators of the influence (impact) of the i-observation on the results of the regression fit.***

Linear models: linearity

Linearity is about the model parameters

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$
$$Y_i = \beta \sin X_i + \varepsilon_i$$

} Linear in βs

$$Y_i = \beta_0 + \log(\beta_1 X_{i1} + \beta_2 X_{i2}) + \beta_3 X_{i3} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \exp(\beta_2 X_i + \beta_3) + \varepsilon_i$$

} Not linear in βs

A concrete example in R

Using the CLASS dataset, from the program SAS
(units have been modified from imperial to metric)

Use statistical models to answer the question:

"Can we predict the height of a teenager, using his age, sex and weight ?"

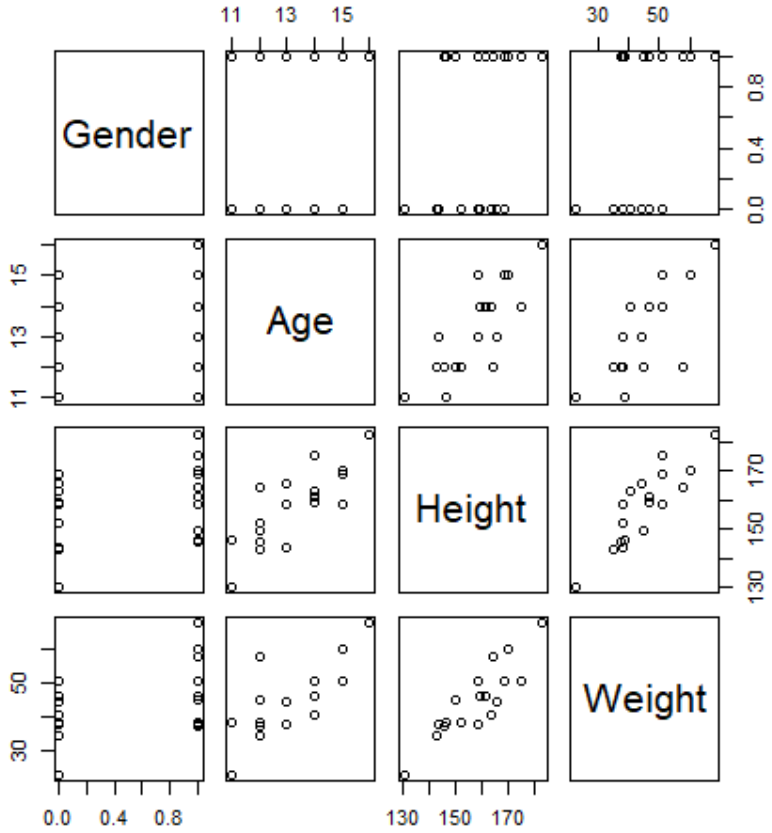
The CLASS dataset from SAS

```
> class
      Name Gender Age  Height  Weight
1    JOYCE      F  11 130.302 22.8765
2   THOMAS      M  11 146.050 38.5050
3    JAMES      M  12 145.542 37.5990
4     JANE      F  12 151.892 38.2785
5     JOHN      M  12 149.860 45.0735
6   LOUISE      F  12 143.002 34.8810
7   ROBERT      M  12 164.592 57.9840
8    ALICE      F  13 143.510 38.0520
9  BARBARA      F  13 165.862 44.3940
10 JEFFREY      M  13 158.750 38.0520
11   CAROL      F  14 159.512 46.4325
12   HENRY      M  14 161.290 46.4325
13  ALFRED      M  14 175.260 50.9625
14    JUDY      F  14 163.322 40.7700
15   JANET      F  15 158.750 50.9625
16    MARY      F  15 168.910 50.7360
17  RONALD      M  15 170.180 60.2490
18 WILLIAM      M  15 168.910 50.7360
19  PHILIP      M  16 182.880 67.9500
```

The CLASS dataset

```
> summary(class[,-1])

Gender      Age             Height          Weight
F: 9   Min.   :11.00   Min.   :130.3   Min.   :22.88
M:10   1st Qu.:12.00   1st Qu.:148.0   1st Qu.:38.17
       Median :13.00   Median :159.5   Median :45.07
       Mean   :13.32   Mean   :158.3   Mean   :45.31
       3rd Qu.:14.50   3rd Qu.:167.4   3rd Qu.:50.85
       Max.   :16.00   Max.   :182.9   Max.   :67.95


> pairs(class[,-1])
```

The CLASS dataset from SAS



The CLASS dataset from SAS

```
> summary( lm( Height ~ Age, data = class) )

Call:
lm(formula = Height ~ Age)

Residuals:
      Min        1Q    Median        3Q       Max
-12.59000  -3.57300  -0.07867   3.49000  15.57133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.069     16.565   3.868  0.00124 **
Age            7.079      1.237   5.724 2.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

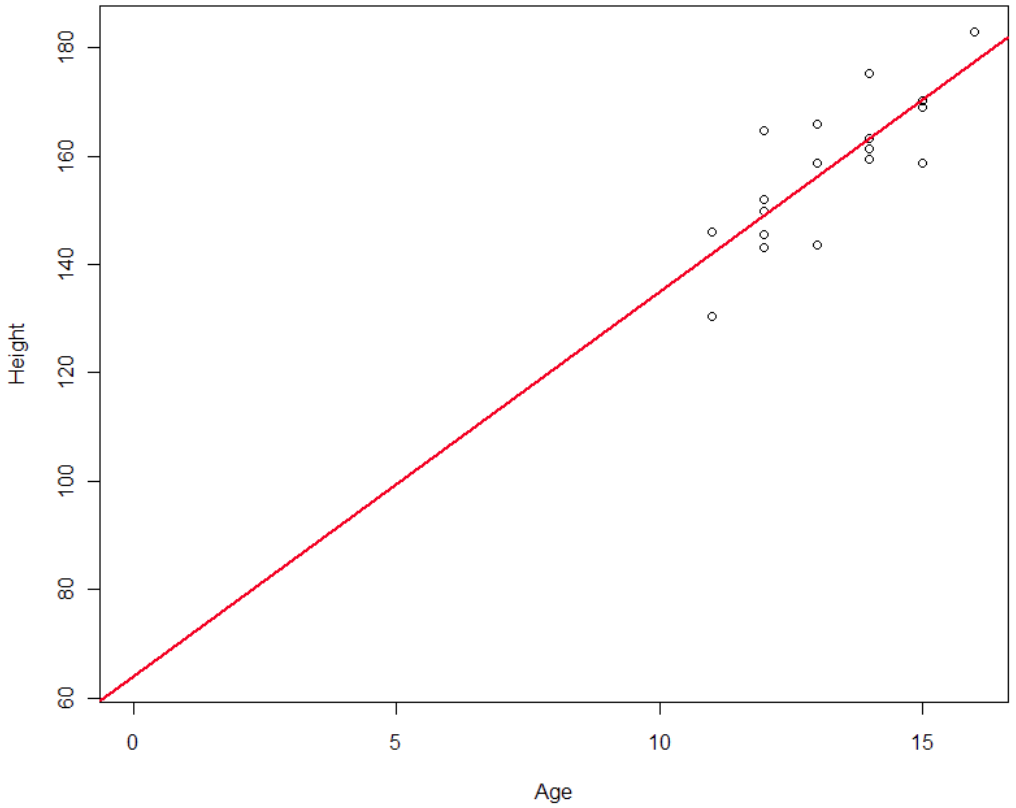
The CLASS dataset from SAS

```
> model <- lm( Height ~ Age, data=class)
> model

Call:
lm(formula = Height ~ Age, data = class)

Coefficients:
(Intercept)          Age
      64.07         7.08
```

Model: Height = 64.07 + 7.08 x Age

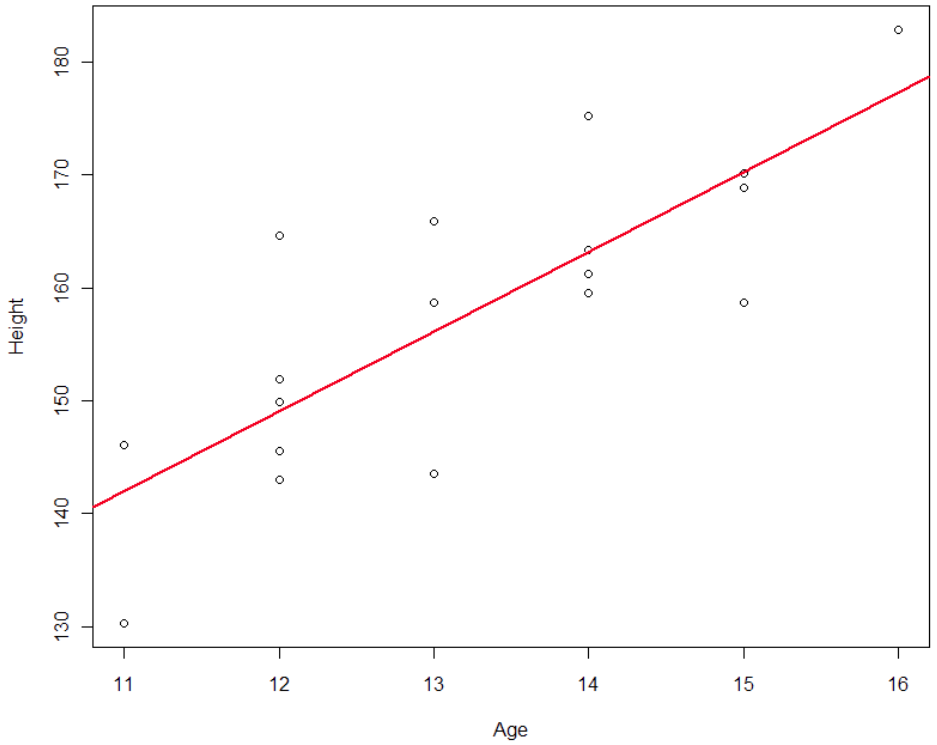
The CLASS dataset from SAS

```
plot( Age, Height, xlim=range(0,Age),
ylim=range(coef(model)[1], Height) )

abline(model, col="red", lwd=2)
```



The CLASS dataset from SAS

```
plot( Age, Height )

abline(model, col="red", lwd=2)
```



The CLASS dataset from SAS

$$se(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.069     16.565   3.868  0.00124 **
Age            7.079      1.237   5.724 2.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These statistical tests tell us if the parameters are significantly different from 0.

Estimate and Std. Error are used for hypothesis testing
t-value = Estimate / Std. Error (of this estimate) follows a t-distribution
(under assumptions: the residuals should follow a normal distribution)

The CLASS dataset from SAS

```
Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

The residual standard error is the standard deviation of the residuals

It is not exactly equal to what the `sd` command would return:

```
> sd(residuals(model)) [1] 7.611075
> sqrt(sum(residuals(model)^2)/18)
[1] 7.611075
```

Here, we must divide by the number of degrees of freedom to get the same number:

```
> sqrt(sum(residuals(model)^2)/17) [1]
7.831732
```

The CLASS dataset from SAS

```
Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

The *number of degrees* of freedom indicates the number of independent pieces of data that are available to estimate the error

While we have 19 residuals here, they are not all independent: for example, the last one is constrained because the sum of all residuals must be 0.

The number of DF is

total observations – number of parameters estimated

Two parameters are estimated (intercept + coefficient), so 19-2 = 17

The CLASS dataset from SAS

```
Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

R^2 is the proportion of the total variance in the response data that is explained by the model (if R^2=1, the data fits perfectly on a straight line).

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} \qquad R^2 = SSR / SST$$

In the case of simple regression, it is equal to the square of the correlation coefficient between the two variables.

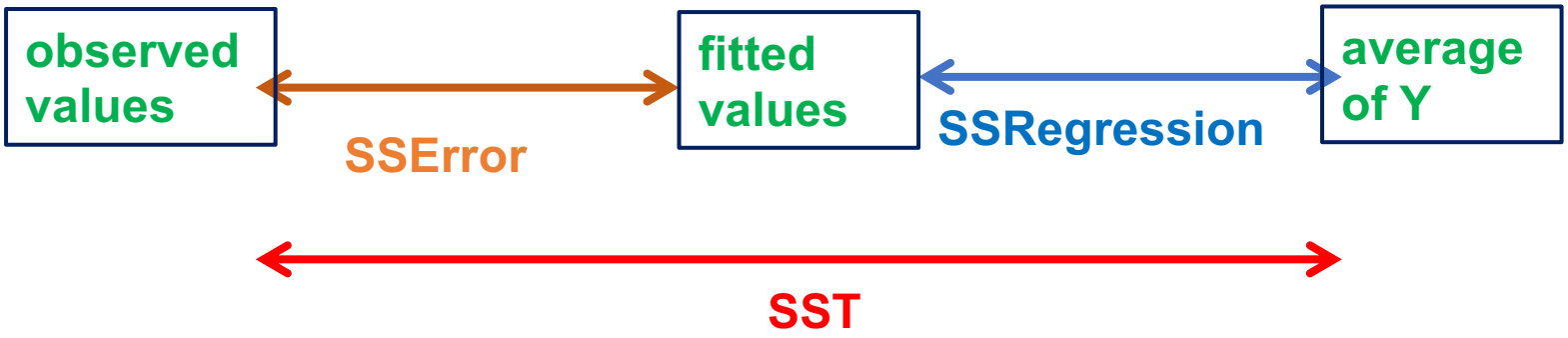
SST = SSR + SSE

Total sum of squares = regression SS + residual SS

$$\underset{\text{SST}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \underset{\text{SSR}}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} + \underset{\text{SSE}}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

R^2 = SSR / SST

HEURISTIC REPRESENTATION



The CLASS dataset from SAS

```
Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

Analysis of variance:

$$\underset{\text{SST}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \underset{\text{SSR}}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} + \underset{\text{SSE}}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean squares (or variance)*	*F*
Regression Model	p=1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	n-2	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	n-1	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$MST = \frac{SST}{n-1}$	

MSR : mean sum of squares of the regression
MSE : mean sum of squares of the errors

RATIO: MSR / MSE is high if the regression reduces the errors considerably compared to what could be expected (by random fitting) given the degrees of freedom
It should follow (given assumptions) a F-distribution

The CLASS dataset from SAS

```
> summary( lm( Height ~ Age, data = class) )

Call:
lm(formula = Height ~ Age)

Residuals:
      Min        1Q    Median        3Q       Max
-12.59000  -3.57300  -0.07867   3.49000  15.57133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.069     16.565   3.868  0.00124 **
Age            7.079      1.237   5.724 2.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

CONFIDENCE INTERVALS FOR THE FITTED PARAMETER VALUES

Usual method

Estimate ± coefficient * Std- Error

coefficient from t-distribution and degrees of freedom
and for desired coverage
for 95% coverage coefficient ~ 1.96

Example

Beta (Age) : Point Estimate = 7.079 ; ± which interval width ?

DF = 17, for 95% coverage :
qt(0.975,17) = 2.109816
Width = 2.110 * 1.237 = 2.610

Beta (Age) 7.079 ± 2.610 ; [4.469, 9.689]

> confint(mm)

	2.5 %	**97.5 %**
(Intercept)	**29.119381**	**99.017952**
Age	**4.470168**	**9.688499**

The CLASS dataset from SAS

Multiple regression: assessing the effect of several variables *together*

Two separate simple regressions

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.069     16.565   3.868  0.00124 **
Age             7.079      1.237   5.724 2.48e-05 ***
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.12816    6.80692  15.885 1.24e-11 ***
Weight        0.50194    0.06644   7.555 7.89e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What happens if both,
age and weight variables
were included in the same model ?

One multiple regression with two variables

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.069     16.565   3.868  0.00124 **
Age             7.079      1.237   5.724 2.48e-05 ***
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.12816    6.80692  15.885 1.24e-11 ***
Weight        0.50194    0.06644   7.555 7.89e-07 ***
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.77355   12.90896   6.335 9.92e-06 ***
Age           3.11575    1.34668   2.314  0.03431 *
Weight        0.35064    0.08827   3.973  0.00109 **
```

While both age and weight seem significant by themselves, age is much less significant when weight is already included.

It is not surprising that a lot of the information provided by the age is also provided by the weight, so that there may be little need to have both terms in the model.

One multiple regression with two variables

```
lm(formula = Height ~ Age)
Multiple R-squared: 0.658,Adjusted R-squared: 0.6383
lm(formula = Height ~ Age + Weight)
Multiple R-squared: 0.828,Adjusted R-squared: 0.8065
```

As before, R^2 is the proportion of the total variance in the response data that is explained by the model.

Adding a new variable in the model will always increase R^2, up to 1 when there the number of degrees of freedom is 0 (number of parameters to estimate = number of observations).

One multiple regression with two variables

```
Multiple R-squared: 0.828,Adjusted R-squared: 0.8065
```

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

The adjusted R-squared adjusts for the number of variables in the model and does not necessarily increase when the number of variables increase.

It is always equal or smaller than R^2 ; for large n (n >> p) about the same.

One multiple regression with n variables

```
Call:
lm(formula = y ~ x1 + x2 + x2 + x2 + x3 + x4 + x5 + x6 + x7 +
    x8 + x9)

Residuals:
ALL 10 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.15582         NA      NA       NA
x1           3.07968         NA      NA       NA
x2          -1.43406         NA      NA       NA
x3          -2.19318         NA      NA       NA
x4           1.48186         NA      NA       NA
x5           1.24668         NA      NA       NA
x6           0.08936         NA      NA       NA
x7           1.43718         NA      NA       NA
x8          -1.22919         NA      NA       NA
x9           1.21790         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 9 and 0 DF,  p-value: NA
```

One multiple regression with two variables

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.77355   12.90896   6.335 9.92e-06 ***
Age          3.11575    1.34668   2.314  0.03431 *
Weight       0.35064    0.08827   3.973  0.00109 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-statistic: 38.52 on 2 and 16 DF,  p-value: 7.646e-07
```

The **F-statistic** allows us to test if the whole regression (adding all variables *vs* having only the intercept in) is significant.

If any of the tests for the individual variables is significant, the F-test will generally be significant as well.

However, even if no individual variable is significant (e.g. $p < 0.05$), the F-test can still be significant.

Categorical variables, dummy variables and contrasts

Categorical variables

We'd like to use categorical variables in a linear model, as in:

Height = b0 + b1 Age + **b2 « Gender »** + error

Intuitively, we want to estimate a « Male » and a « Female » effect.

In practice, categorical variables (factors in R) are turned (by default, based on alphabetical order) into dummy variables of the form.

$$\textbf{Gender} = \begin{cases} 0 \text{ if Female} \\ 1 \text{ if Male} \end{cases}$$

and the model can be interpreted as follows:

- b_0 is the baseline for height among women (at Age = 0)
- $\mathbf{b_2}$ represent the increase/decrease of this baseline for men.

- .

Categorical variables

```
Call:
lm(formula = Height ~ Age + Gender)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8462 -4.8523 -0.8102  3.3677 13.5058

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.291     14.957   4.165  0.00073 ***
Age            6.928      1.117   6.202 1.27e-05 ***
GenderM        7.204      3.251   2.216  0.04152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.061 on 16 degrees of freedom
Multiple R-squared: 0.7387,     Adjusted R-squared: 0.706
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

baseline for height among Female

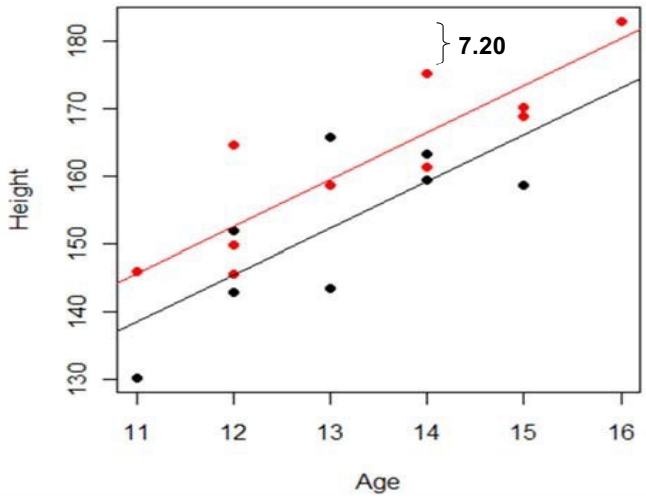
The factor GenderM corresponds to the difference in baseline for Males compared to females

Categorical variables

The model specifies 2 straight lines, with the same slope but different y-intercepts:

For women: Height = 62.29 + 6.93 Age (in black)

For men: Height = 69.49 + 6.93 Age (in red)



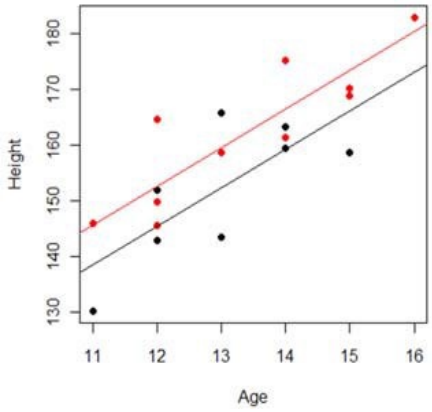
Interaction

So far, we have assumed a difference between the lines, but the same slope; that is, for both men and women, the effect of age is the same.

If this assumption is incorrect, it means that there is an ***interaction*** between the factors « age » and « gender », that is, the effect of age is different depending on the gender.

Interactions are modeled in R in the following way:

```
lm(formula = Height ~ Age + Gender +
Age:Gender)
```

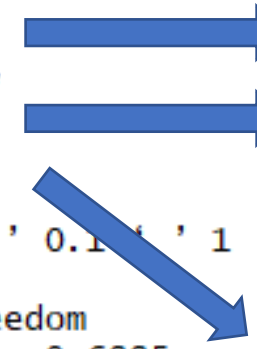


which is equivalent to

```
lm(formula = Height ~ Age * Gender)
```

Interaction

```
Call:
lm(formula = Height ~ Age * Gender, data = class)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7449 -4.5324 -0.9265  3.4873 13.6071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.2610    24.4880   2.297  0.03640 *
Age           7.3841     1.8429   4.007  0.00114 **
GenderM      17.1304    31.5238   0.543  0.59483
Age:GenderM  -0.7468     2.3583  -0.317  0.75585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.269 on 15 degrees of freedom
Multiple R-squared:  0.7404,    Adjusted R-squared:  0.6885
F-statistic: 14.26 on 3 and 15 DF,  p-value: 0.0001152
```



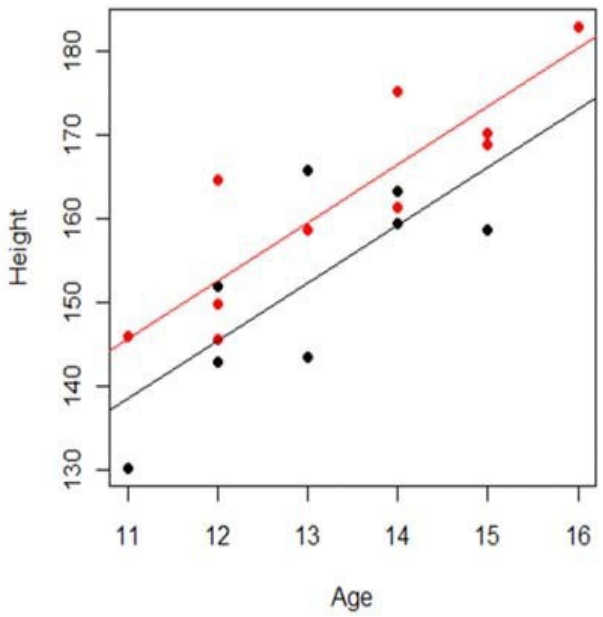
baseline for height among Female

difference in baseline for Males compared to females

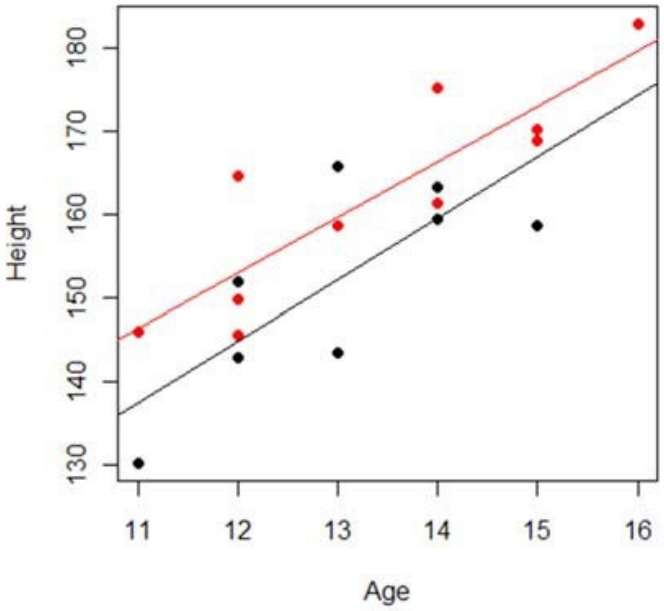
age effect only for males

Height = 56.26 + 7.38*Age + 17.13 (only for males) – 0.75*Age (only for males)

Interaction



No interaction



With interaction

What if my variable has more than 2 levels ?

The interpretation was straightforward with two levels: one was the baseline, and we estimated the difference between the second one and the baseline.

With more than two levels, there are different ways, termed contrasts, of looking at the coefficients.

The most common one is called **treatment contrasts** and corresponds to taking the first level as the baseline (as a control), and all the other coefficients correspond to differences of each level with the control (treatments).

Diagnostic tools

Basic model checking

It is always possible to fit a linear model and find a slope and intercept
… but it does not mean that the model is “meaningful” or “optimal” !

Model checking **questions**:

- How good is the fitting , could it be improved ?
- Are there **outliers**, which might “disturb” during fitting ?
- Are there points that have a **high impact** and might decrease the model quality ?

- Does the the model fir look “perfect”: the residuals are “normal” (“Gaussian”) and have “constant variance” ? And are independent from each other ?

technically $E_i \sim$ i.i.d $N(0,$ sigma constant $| i)$

Note: the statistical tests (p-values) and confidence intervals are calculated using this assumptions, they are unreliable if this is not at least approximately satisfied.

Basic model checking

Examination of ***Residuals***:

- If they show a **pattern** => maybe can improve the model, there is still a systematic trend that could be captured by a better model

- If they have **variable variance**:

Is the model missing something compared to reality (is miss-specified): another explanatory variable ?, a data transformation?,
Or maybe there are some outliers that impact the parameter estimation?

- If they are **NOT normally** distributed: same questions

- If they are **NOT independent**: were the data collected in a "good way" ?

- If there are **Outliers**: which points (is OK? eliminate ? Can be verified ?) (or consider using "robust regression" methods instead of regular LS ?)

Residuals

Types of *Residuals*:

- **Raw residuals** Ri = Observed - Fitted (Expected) = Y - Y_i

- Rescaled (specifically to each data point) to have expected sd = 1 :
  Studentized Residuals
- Should follow about a t-distribution (resp. approx. N(0,1))

Basic model checking

Examination of ***Influence***:

– Are there "overly influential points" ?, which / why ? Bad "design" ?
– Repeat / eliminate ?

– Detection of *influential observations*: *Hat matrix*
*(potential influence, **leverage**)*

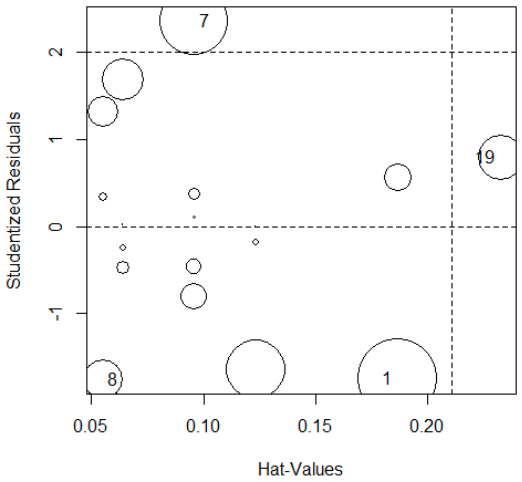
LEVERAGE**:* ***Hat

*OUTLIERS / **WEIRDNESS**:* ***(Stud.) Residuals***

***INFLUENCE**:*
Leverage x Weirdness ***Cook's d.***

influencePlot

```
library(car)

influencePlot(model,
xlab="Hat-Values",
ylab="Studentized
Residuals")
```



Description

This function creates a “bubble” plot of Studentized residuals versus hat values, areas of the circles proportional to the value “Cook's distance”.

Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the Studentized-residual scale.

Value

If points are identified, returns a data frame with the hat values, Studentized residuals and Cook's distance of the identified points. If no points are identified, nothing is returned. This function is primarily used for its side-effect of drawing a plot.

Hat values

High leverage ('influential') : example :
points far from the center, have potentially greater influence ("leverage effect")

One way to identify these points is through the *hat values*
(obtained from the *hat matrix H*):

h_{ij}: contribution of the ith observation to the jth fitted value

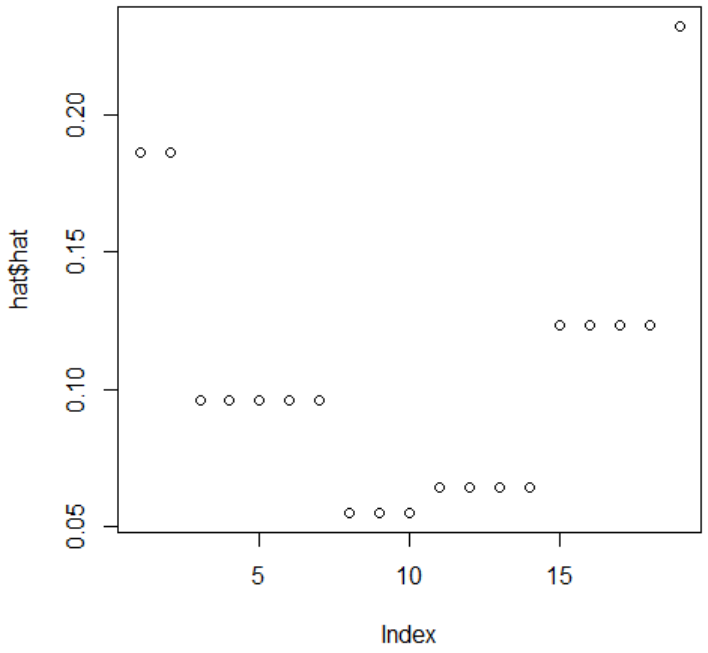
Most informative are the diagonal values of the hat matrix

$H_i = h_{ii}$: "leverage" of the ith observation to the fitted values (via the fitted model parameters)

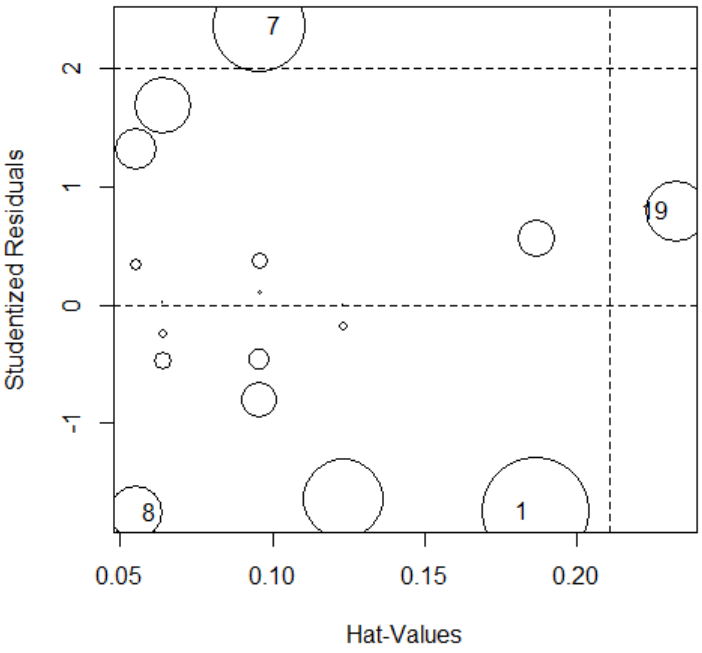
Average value of h = nb of predictors p / nb of points n (**p / n**)

Cutoff typically 2*p/n to 3*p/n:
points beyond are considered worthy of a careful examination

Hat values

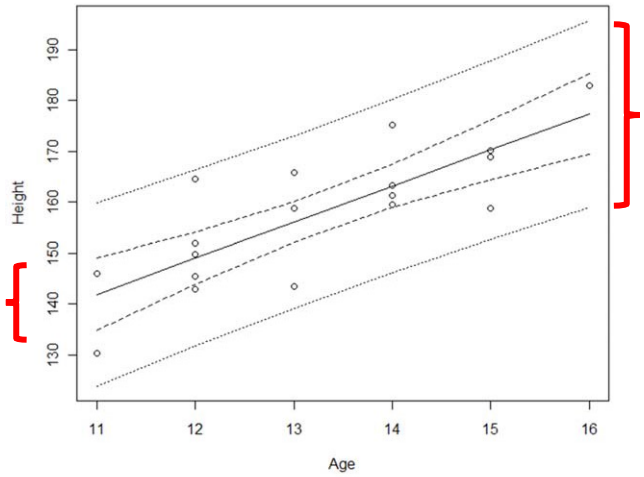


id	Name	Gender	Age	Height	Weight
1	JOYCE	F	11	130.302	22.877
2	THOMAS	M	11	146.050	38.505
3	JAMES	M	12	145.542	37.599
4	JANE	F	12	151.892	38.279
5	JOHN	M	12	149.860	45.074
6	LOUISE	F	12	143.002	34.881
7	ROBERT	M	12	164.592	57.984
8	ALICE	F	13	143.510	38.052
9	BARBARA	F	13	165.862	44.394
10	JEFFREY	M	13	158.750	38.052
11	CAROL	F	14	159.512	46.433
12	HENRY	M	14	161.290	46.433
13	ALFRED	M	14	175.260	50.963
14	JUDY	F	14	163.322	40.770
15	JANET	F	15	158.750	50.963
16	MARY	F	15	168.910	50.736
17	RONALD	M	15	170.180	60.249
18	WILLIAM	M	15	168.910	50.736
19	PHILIP	M	16	182.880	67.950




```
hat <- lm.influence(model)
plot(hat$hat)

library(car)
influencePlot(model, xlab="Hat-Values", ylab="Studentized Residuals")
```

Confidence bands



Prediction interval

Confidence interval

Narrow bands: describe the uncertainty about the regression line

Wide bands: describe where most (95% by default) predictions would fall, assuming normality and constant variance.

```
predict.lm(model, newdata=data.frame(Age=new_age), interval="confidence")
predict.lm(model, newdata=data.frame(Age=new_age), interval="prediction")
```

CONFIDENCE INTERVALS FOR SINGLE VALUES, CURVE AND PREDICTIONS

1) What is the “precision” of the fitted values E[Y | X] ?

2) What is the “precision” of the regression line E[Y | X] “as a whole” ?

3) If new values are sampled:
where are they likely to fall into ? Where should we expect them to be ?

CONFIDENCE INTERVALS FOR SINGLE VALUES, CURVE AND PREDICTIONS

"precision" of the estimated values E[Y | X] (the regression line)

1) Let's define (separately) at each value of X
a (vertical) confidence interval CI , such that
the true value E[Y | X] lies within the CI in 95% of the times
(times we do the fitting with such kind of data resampled from the same distribution)
Positions-specific "error" of the estimation.
E[Y | X] being the mean of the Y (given the X): the more data we have the more precise the estimate, the narrower the CI.
Width of the CI -> 0 as n -> ∞
Pointwise / single CI (together they form a band around the regression line)
The width comes from the imprecision in the estimate of the parameters

2) Let's define (together, simultaneously) a **band** around the regression line
Around the regression line such that the true curve of the E[Y | X] values lies completely within the defined band, **simultaneously** for all positions, at no position it is outside.
For this we need a larger band than for 1) as we are asking for more.
The width comes from the imprecision in the estimate of the parameters and here also:
Width of the band -> 0 as n -> ∞

CONFIDENCE INTERVALS FOR SINGLE VALUES, CURVE AND PREDICTIONS

"precision" of the estimated values E[Y | X] (the regression line)

1) **Pointwise**

In mathematical terms, a pointwise confidence band $\hat{f}(x) \pm w(x)$ with coverage probability 1 – *α* satisfies the following condition separately for each value of *x*:

$$\Pr\left(\hat{f}(x) - w(x) \leq f(x) \leq \hat{f}(x) + w(x)\right) = 1 - \alpha,$$

where $\hat{f}(x)$ is the point estimate of *f*(*x*).

2) **Simultaneously** for all position points

In mathematical terms, a simultaneous confidence band $\hat{f}(x) \pm w(x)$ with coverage probability 1 – *α* satisfies the following condition:

$$\Pr\left(\hat{f}(x) - w(x) \leq f(x) \leq \hat{f}(x) + w(x) \text{ for all } x\right) = 1 - \alpha.$$

CONFIDENCE INTERVALS and CONFIDENCE BANDS

Confidence bands are closely related to confidence intervals, which represent the uncertainty in an estimate of a single numerical value.

A **confidence band** is used to represent the uncertainty in an estimate of a curve .

"As confidence intervals, by construction, only refer to a single point, they are narrower (at this point) than a confidence band which is supposed to hold simultaneously at many points."

Similarly, a **prediction band** is used to represent the uncertainty about the value of a new data-point on the curve, subject to sampling variability.

[https://en.wikipedia.org/wiki/Confidence_and_prediction_bands]

CONFIDENCE FOR THE CURVE AND FOR NEW VALUES (PREDICTIONS)

A prediction band is used to represent the uncertainty about the value of a new data-point on the curve, subject to sampling variability.

[https://en.wikipedia.org/wiki/Confidence_and_prediction_bands]

Predict a new value:

Given a value of X,

a) consider the imprecision about the difference between the true value of E[Y|X] and the one estimated by the fitting, take a draw from this distribution (the one that gives us the confidence interval of E[Y|X] at the position X)

b) now with this for this E[Y|X] generate a new point using the (estimation of the) underlying distribution (that is the standard deviation of the residuals, the scatter of the single points)

In b) we have the variability due to the data-generation process we are studying, it is something that exists "outside" of the modeling and is given.

Only for a) : standard deviation -> 0 as n -> ∞

Also: most of the given data points must be inside the "prediction band", while only a few might be within the the "confidence intervals band".

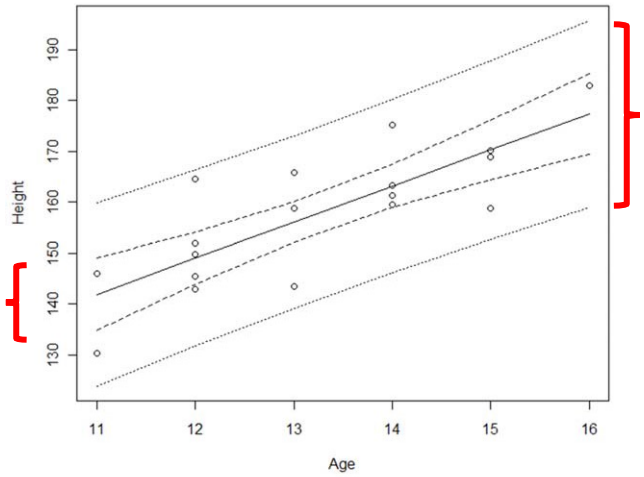
CONFIDENCE INTERVALS FOR SINGLE VALUES, CURVE AND PREDICTIONS

Analogy
Data ~ N (μ, sigma)

Estimate μ with a ^μ
confidence interval for ^μ to include real μ :
this gets smaller with increasing sample size (standard errors)
and is smaller than sigma (usually ~ sigma / sqrt(n) }

New data are ~ N (^μ, sigma)
scattering width is due to (imprecision in ^μ) + sigma

Confidence bands



Prediction interval

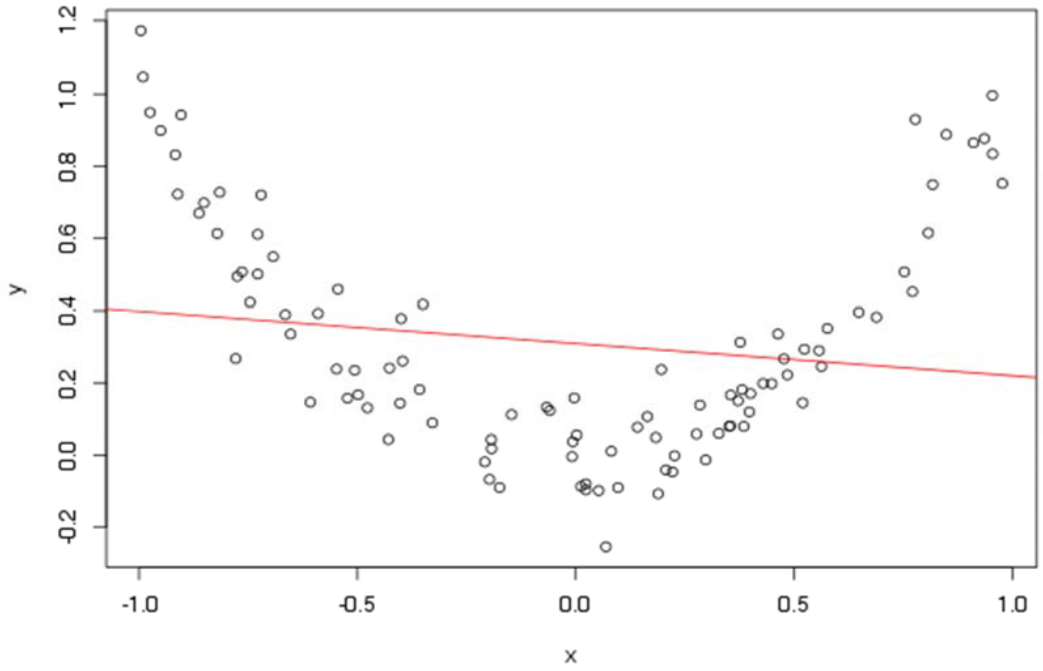
Confidence interval

Narrow bands: describe the uncertainty about the regression line

Wide bands: describe where most (95% by default) predictions would fall, assuming normality and constant variance.

```
predict.lm(model, newdata=data.frame(Age=new_age), interval="confidence")
predict.lm(model, newdata=data.frame(Age=new_age), interval="prediction")
```

What if the data is not linear ?

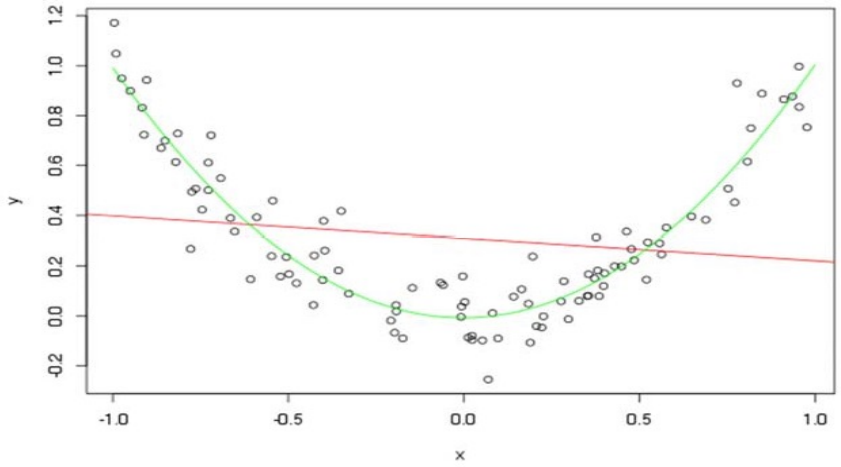


What if the data is not linear ?

Use a polynomial regression

$$y = b_0 + b_1 x + b_2 x^2$$

This is still linear for b_i; it is as if we had added a new variable.

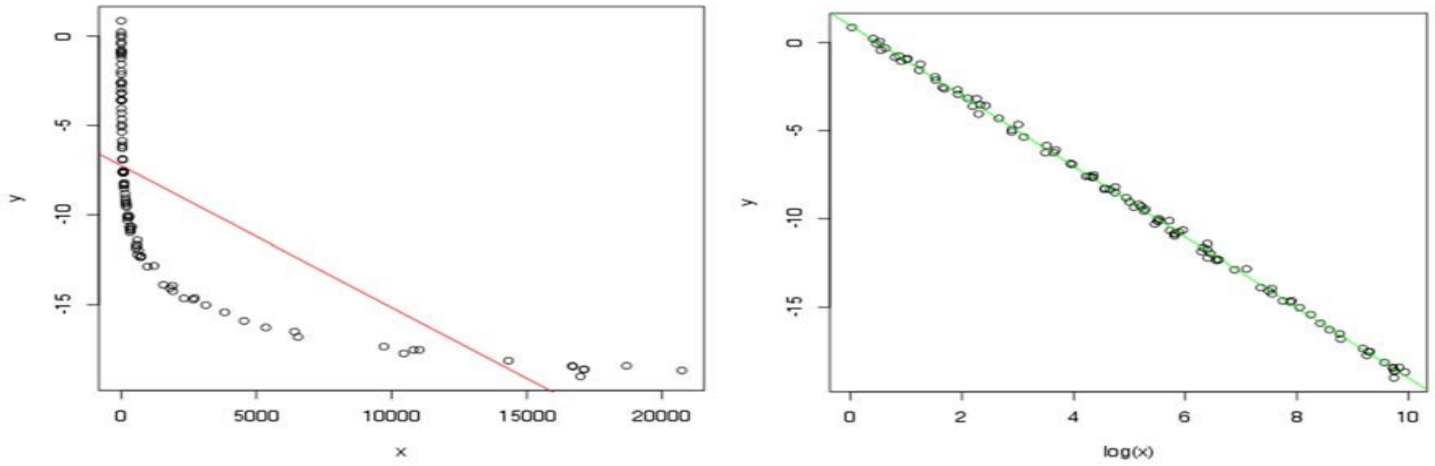


What if the data is not linear ?

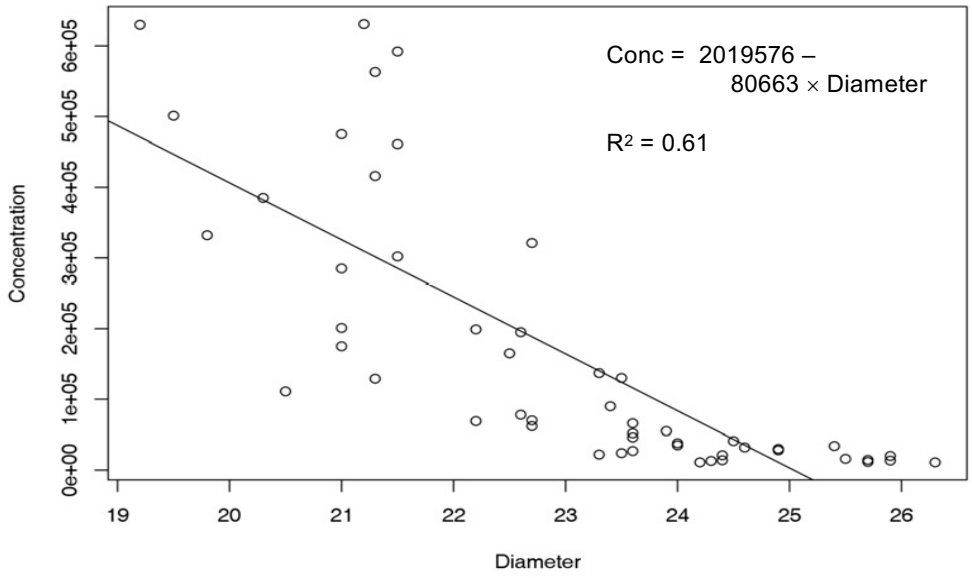
Consider transforming the data (log)

$$\log(y) = a + b\,x$$

$$y = a + b\,\log(x)$$



Linear model predicting Concentration from Diameter




```
> model <- lm( conc ~ diameter, data=hellung
)
> abline(model)
```

Example: predicting cell concentration

The hellung dataset

" Diameter and concentration of *Tetrahymena* cells with and without glucose added to growth medium."

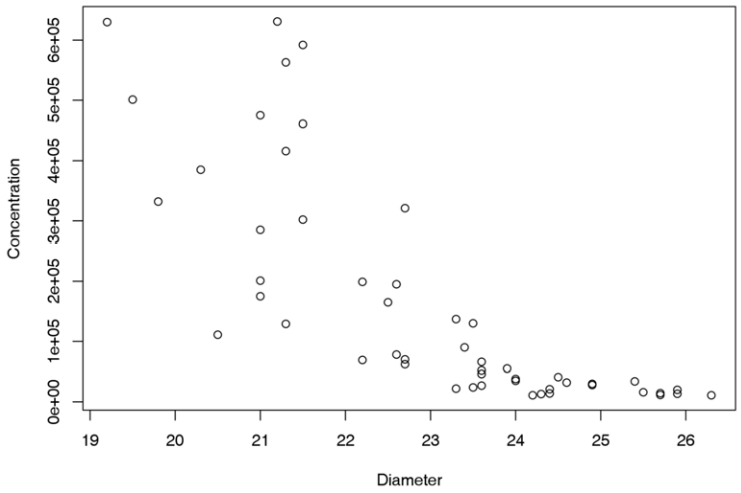
```
> library(ISwR); data(hellung)
```

Can we predict the concentration of cells using the diameter and the presence/absence of glucose ?

The Hellung data in R

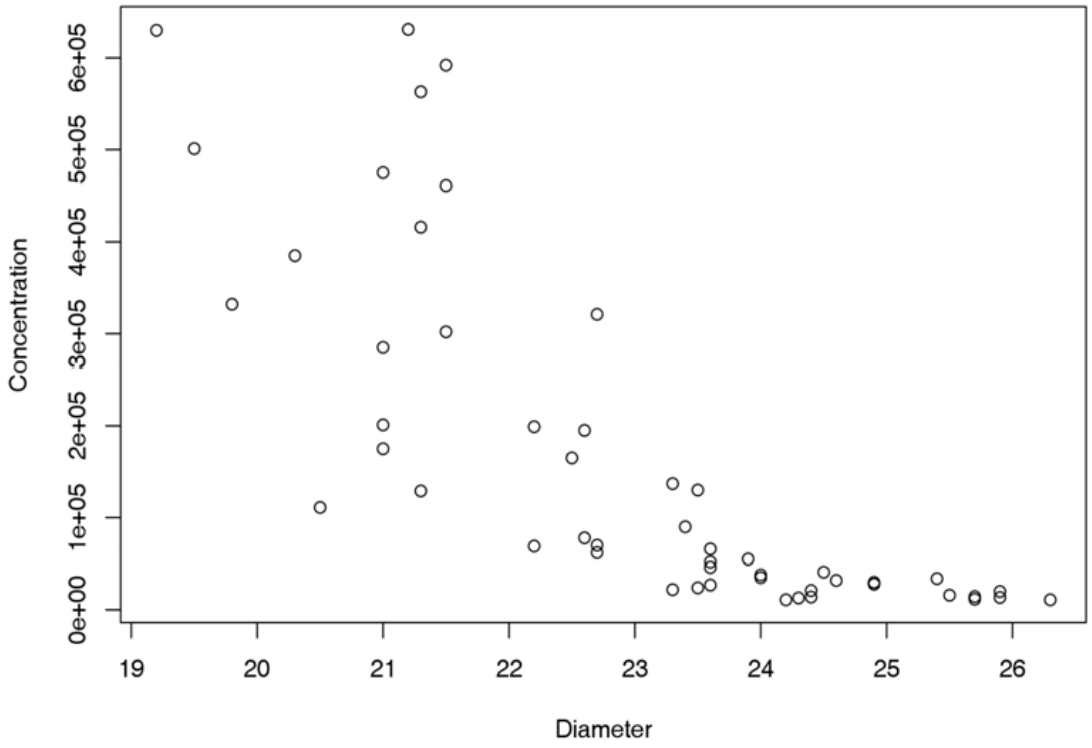
```
> hellung
    glucose           conc     diameter
1           1      631000           21.2
2           1      592000           21.5
3           1      563000           21.3
4           1      475000           21.0
5           1      461000           21.5
   [...]
33          2      630000           19.2
34          2      501000           19.5
35          2      332000           19.8
36          2      285000           21.0
37          2      201000           21.0
```

Hellung dataset: Diameter vs Concentration



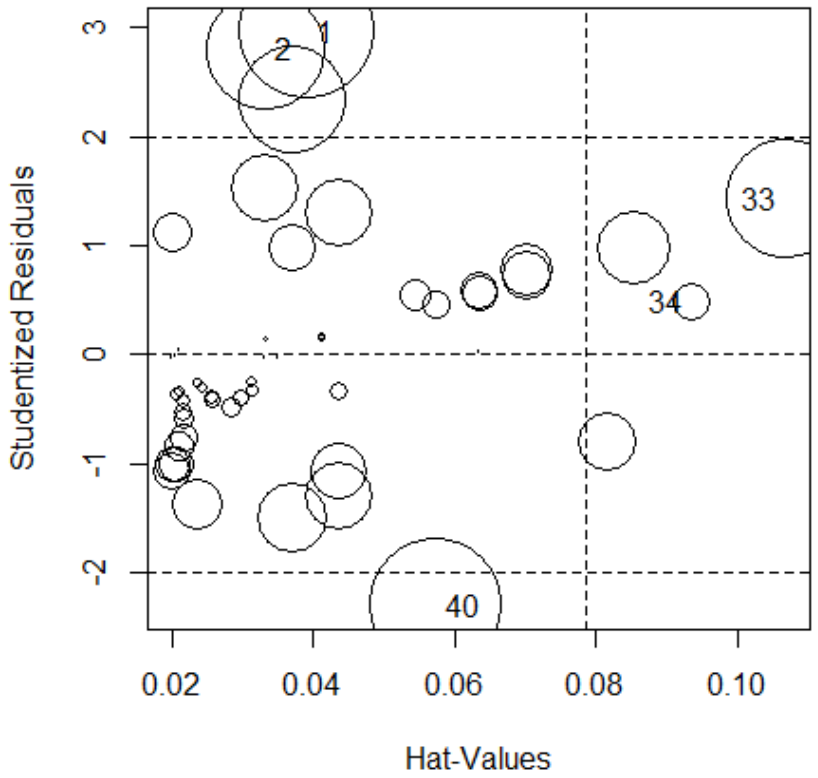

```
> plot(hellung$diameter, hellung$conc,
     xlab="Diameter",
     ylab="Concentration")
```

Can we predict the concentration given the diameter of the cells ?

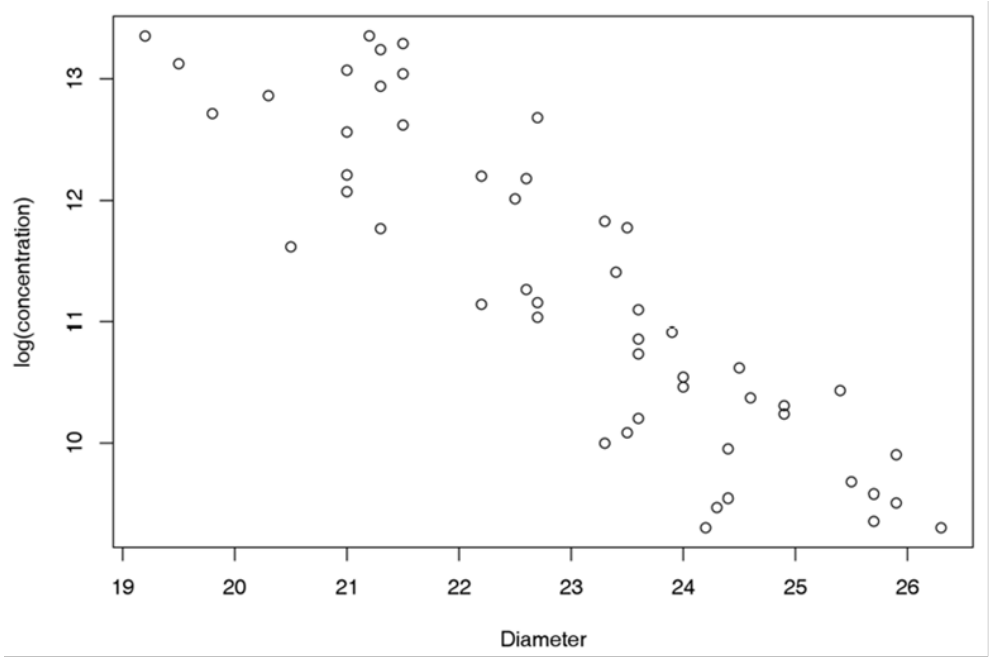


Residuals and hat values

```
> influencePlot(model, xlab="Hat-Values", ylab="Studentized Residuals")
      StudRes        Hat      CookD
1   2.9625032 0.03915889 0.15434569
2   2.7930627 0.03318496 0.11756602
33  1.4280137 0.10674277 0.11931146
34  0.4752678 0.09352771 0.01183991
40 -2.2980607 0.05732206 0.14766395
```

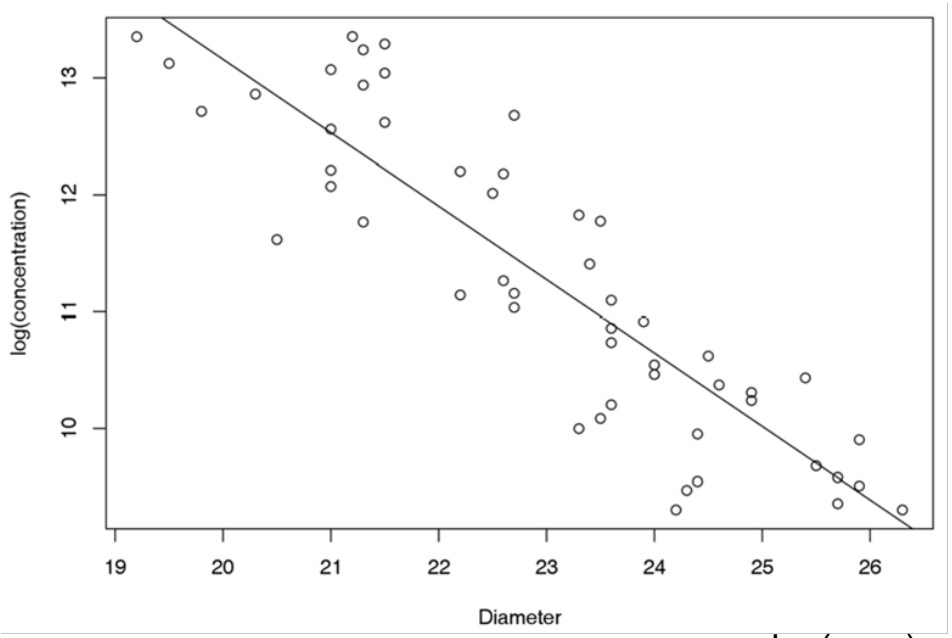


Transforming the data to improve the fit




```
logconc <- log(hellung$conc)
plot(hellung$diameter, logconc,
     xlab="Diameter", ylab="log(concentration)" )
```

Linear model predicting log(Concentration) from Diameter



log(conc) = 25.7 – 0.62 × Diameter

```
modellog <- lm(logconc ~ diameter, data=hellung)
abline(modellog)
```

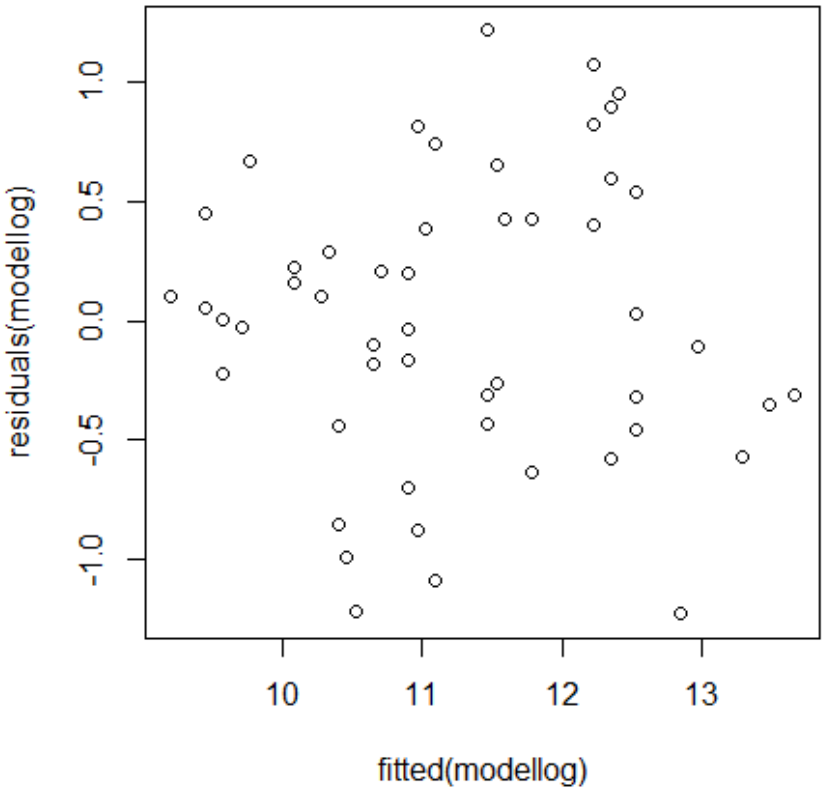
$R^2 = 0.78$

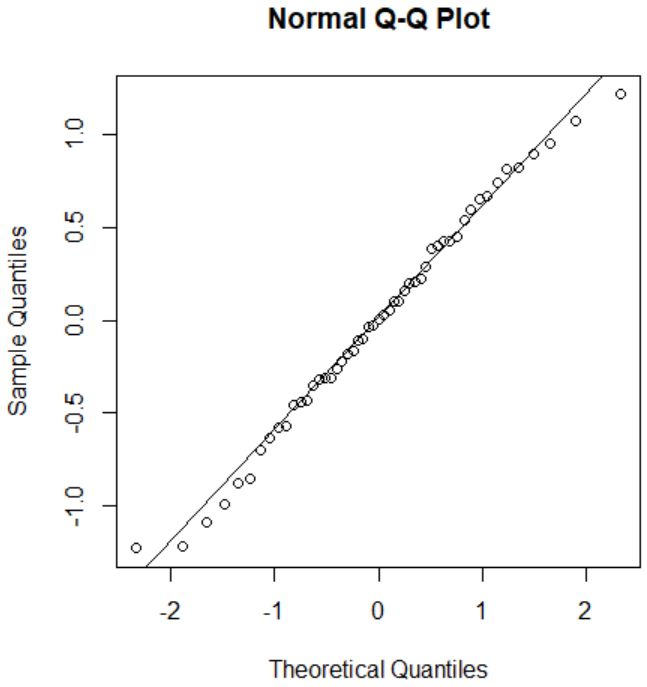
Details of the linear model

$$\log(\text{concentration}) = 25.7 - 0.63 \times \text{diameter}$$

```
summary(modellog)

    Call:
    lm(formula = logconc ~ diameter)

    Residuals:
          Min        1Q    Median        3Q       Max
    -1.227992 -0.388761  0.003015  0.424183  1.215852

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) 25.72239    1.09418   23.51   <2e-16 ***
    diameter    -0.62815    0.04743  -13.24   <2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 0.6105 on 49 degrees of freedom
    Multiple R-squared: 0.7817,     Adjusted R-squared: 0.7772
    F-statistic: 175.4 on 1 and 49 DF,  p-value: < 2.2e-16
```

Diagnostic plots





Predicting Concentration from diameter

We have a linear model for predicting the log of the concentration:

$$\log(\text{concentration}) = 25.7 - 0.63 \times \text{diameter}$$

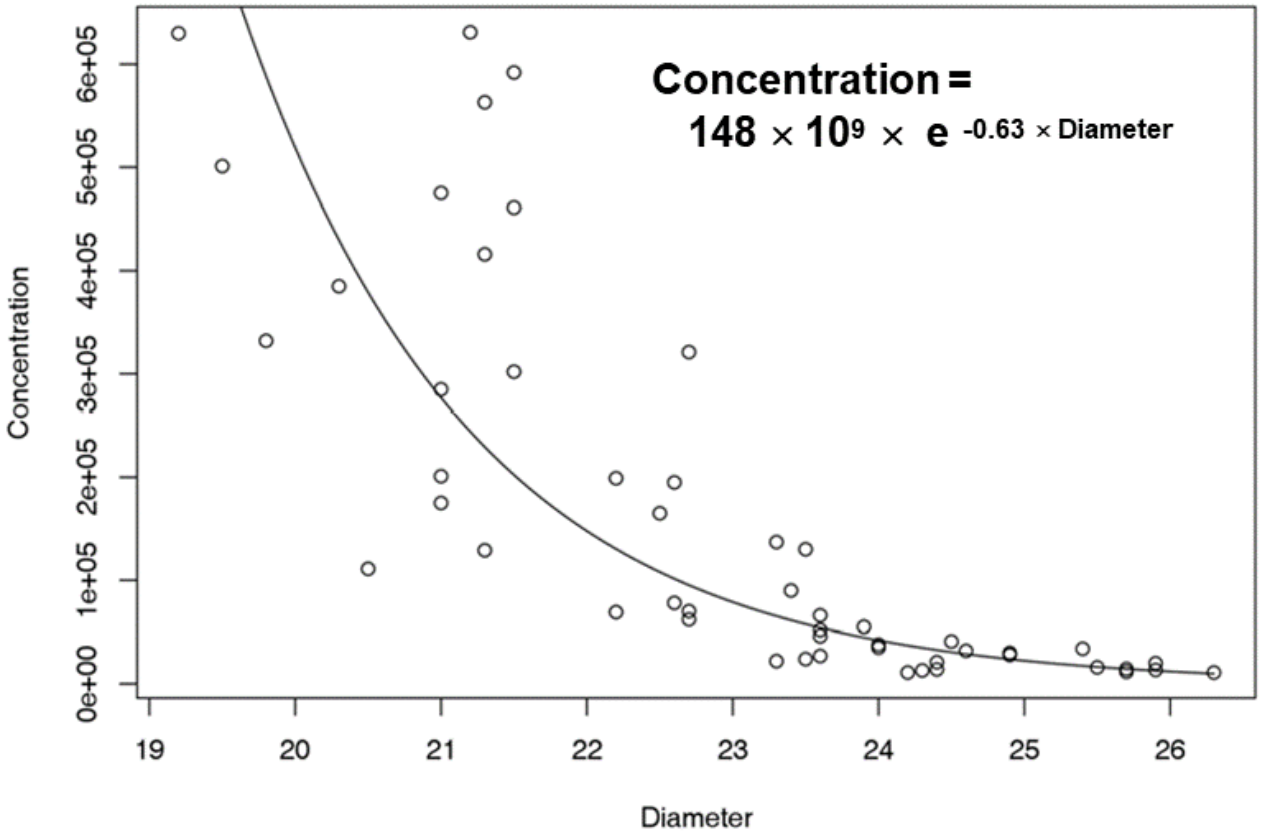
We have a function that links this prediction to our value of interest (concentration):

log / exponential

This allows us to make predictions for the concentration:

$$\text{Concentration} = 148 \times 10^{9} \times e^{-0.63 \times \text{Diameter}}$$

Predicting Concentration from diameter



The Hellung data in R

```
> hellung
    glucose       conc   diameter
1           1  631000          21.2
2           1  592000          21.5
3           1  563000          21.3
4           1  475000          21.0
5           1  461000          21.5
[...]
33          2  630000          19.2
34          2  501000          19.5
35          2  332000          19.8
36          2  285000          21.0
37          2  201000          21.0
```

hellung package:ISwR R Documentation

Growth of Tetrahymena cells

Description:

The 'hellung' data frame has 51 rows and 3 columns. diameter and concentration of _Tetrahymena_ cells with and without glucose added to growth medium.

Format:

This data frame contains the following columns:

'glucose' a numeric vector code, 1: yes, 2: no.

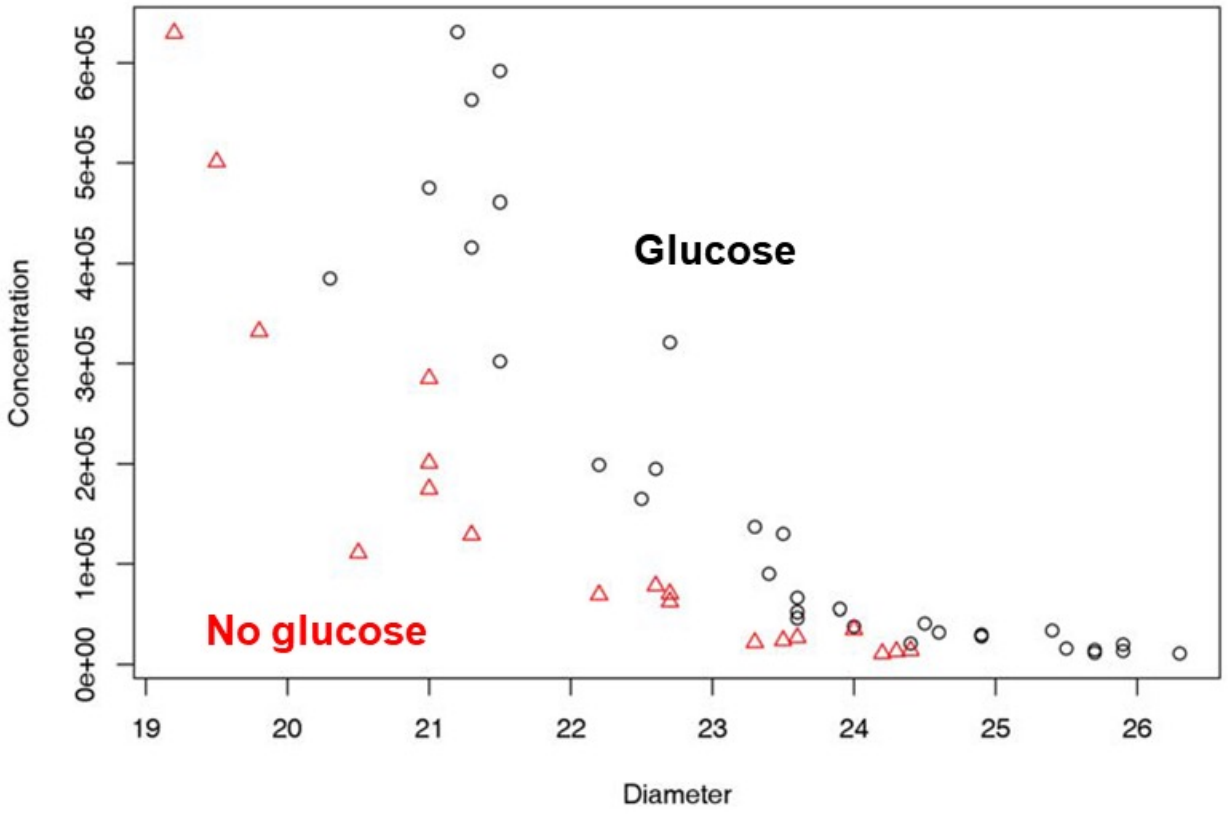
'conc' a numeric vector, cell concentration (counts/ml).

'diameter' a numeric vector, cell diameter (micrometre).

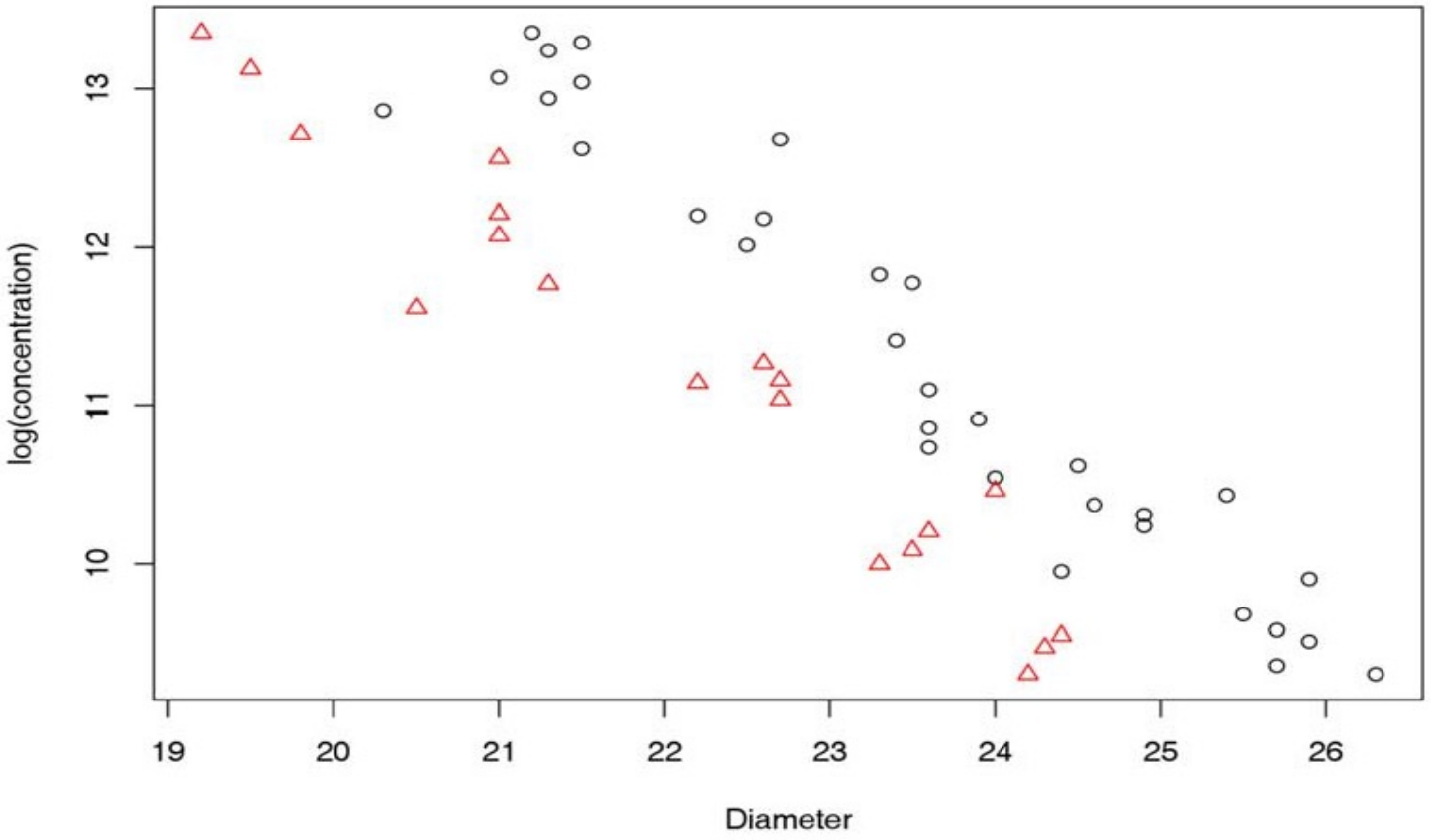
Source:

D. Kronborg and L.T. Skovgaard (1990), _Regressionsanalyse_, Table 1.1, FADLs Forlag (in Danish).

Concentration according to Diameter and Glucose



Log(concentration) according to diameter and glucose



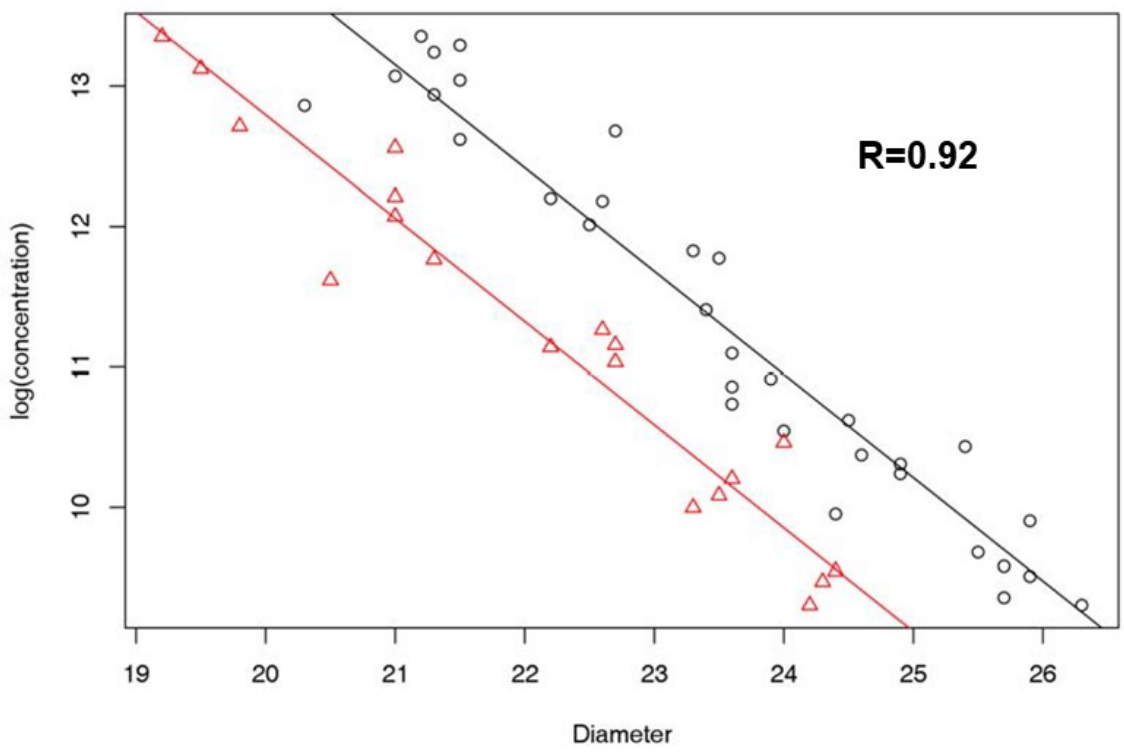
Reminder: using categorical variables as explanatory variables

We would like to use categorical variables in a linear model, as in:

Concentration $= b_0 + b_1$ Diameter $+ b_2$ « Glucose » + error

Intuitively, we want to estimate a « No glucose » and a « Glucose » effect.

Prediction of log Concentration according to Diameter and Glucose



Prediction of Concentration according to Diameter and Glucose