

Swiss Institute of
Bioinformatics

SIB Swiss Institute of Bioinformatics

Advanced statistics: Statistical modeling 2022, 22-25 August 2021

Rachel Marcone (Jeitziner) and Mauro Delorenzi

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

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**)

Statistical models

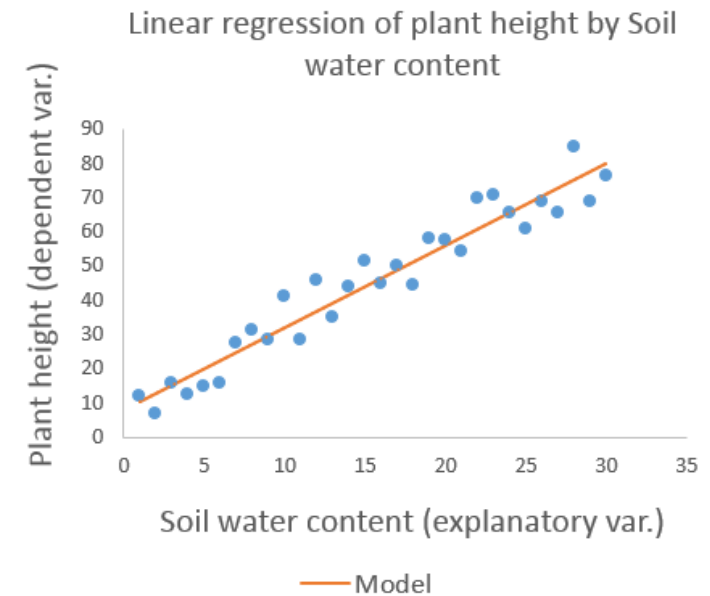
What is a statistical model ?

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** $\xrightarrow{\text{model}}$ **answers** + **measures of uncertainty**

(from Terry Speed)

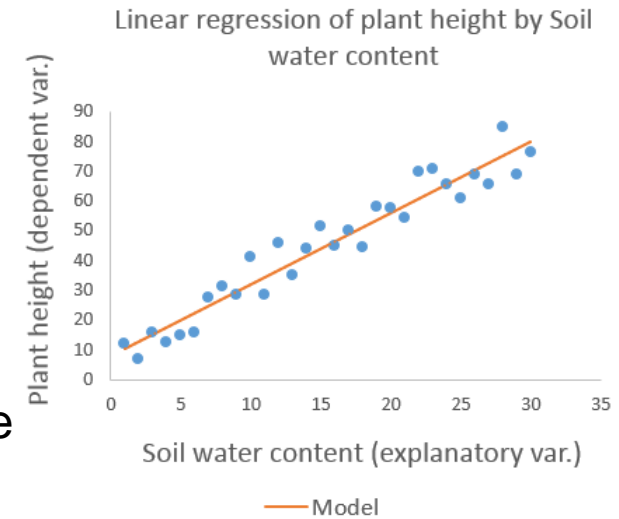
$$\text{Height} = \text{intercept} + \text{slope} * \text{soil water content}$$

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



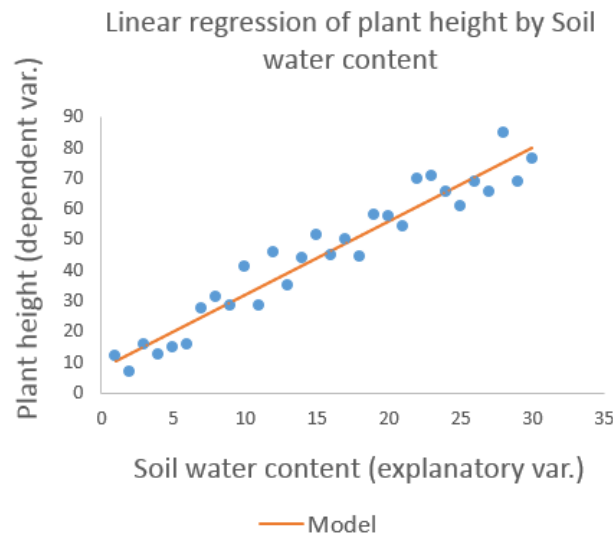
What is a model parameter ?

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)

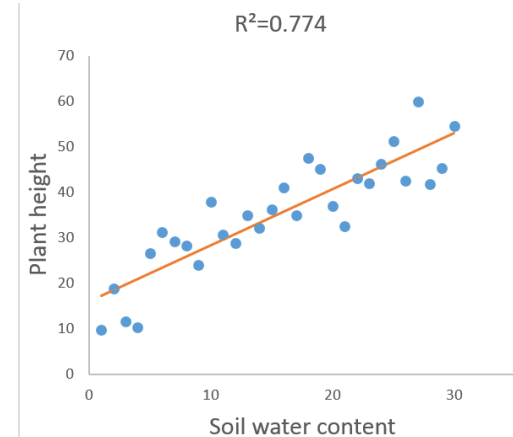
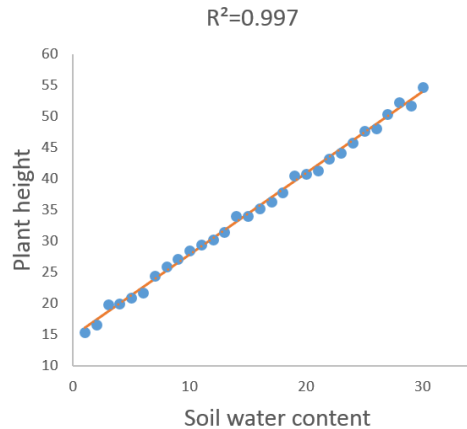
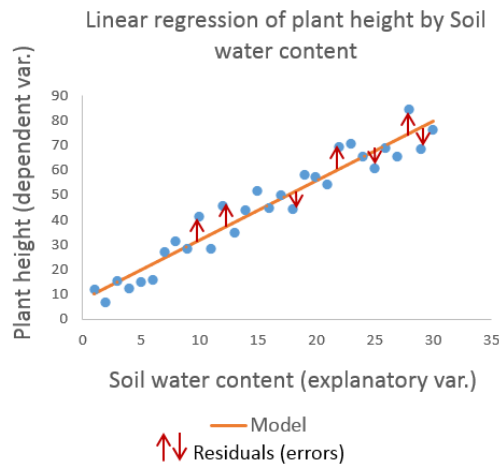


$$\text{Height} = \text{intercept} + \text{slope} * \text{soil water content}$$

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

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

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

Modeling overview revised

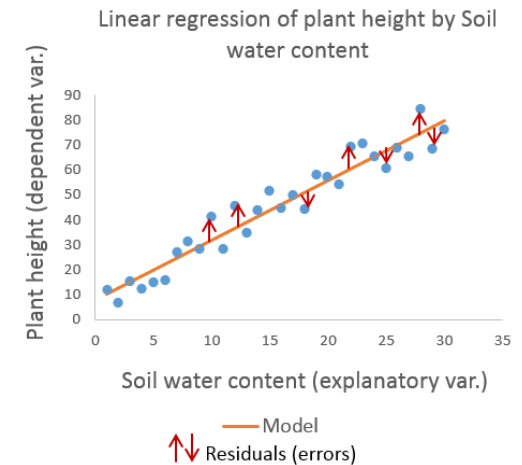
Given a response $g(Y)$ or Y
that (might) depend on a variable X

$X \rightarrow ? \rightarrow$ individual values Y_i

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

$X \rightarrow E(Y)$ the expected value of $Y \rightarrow$ individual values Y_i

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



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 this 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
```

Model formulas in R

The generic form is **response ~ predictors**

The predictors can be **numeric** or **factor**

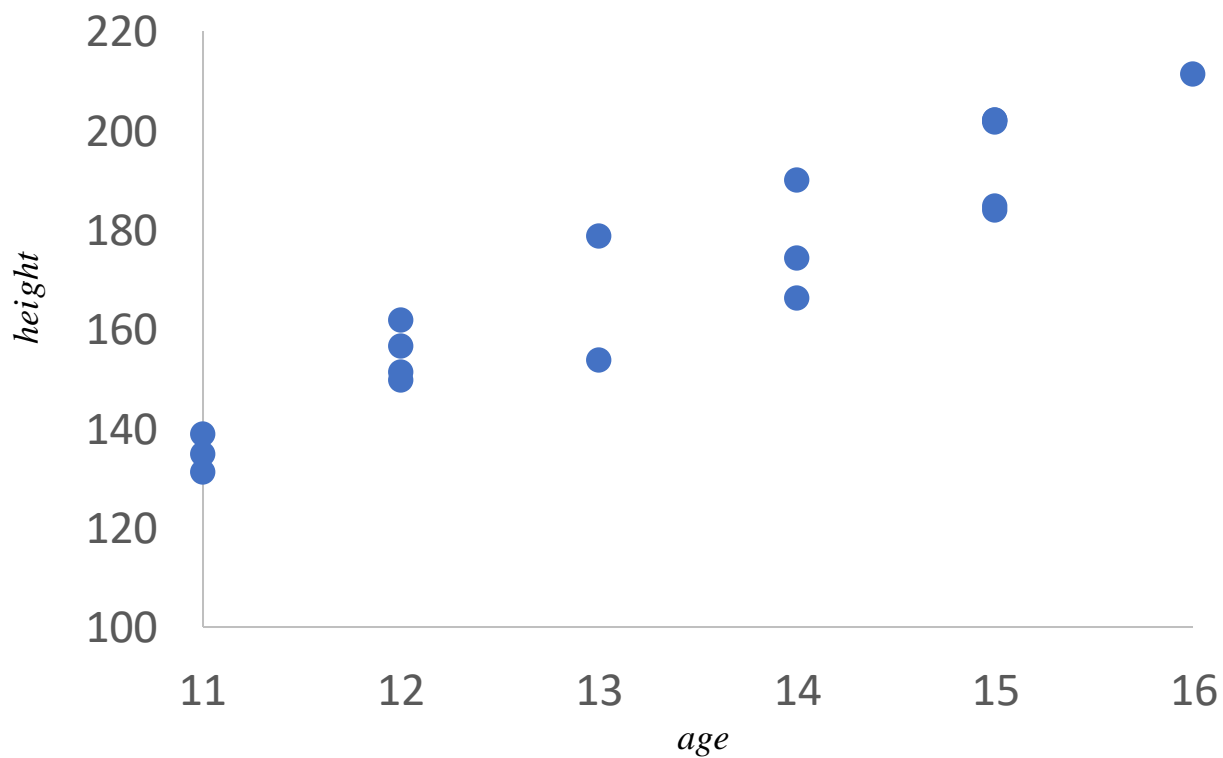
Other symbols to create formulas with **combinations of variables** (e.g. **interactions**)

- +** to **add** more variables ($a + b$)
- to **leave** out variables ($a*b - a:b$ is the same as $a + b$)
- :** to introduce **interactions** between two terms ($a:b$)
- *** to include **both interactions and the terms** ($a*b$ is the same as $a + b + a:b$)
- $\wedge n$** to **add variables** to the **power of n**
- l()** treats what's in () as a **mathematical expression** ($a + b$ versus $l(a + b)$)

Linear models

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

Example: scatterplot of age vs height in teenagers



(Simple) Linear Regression

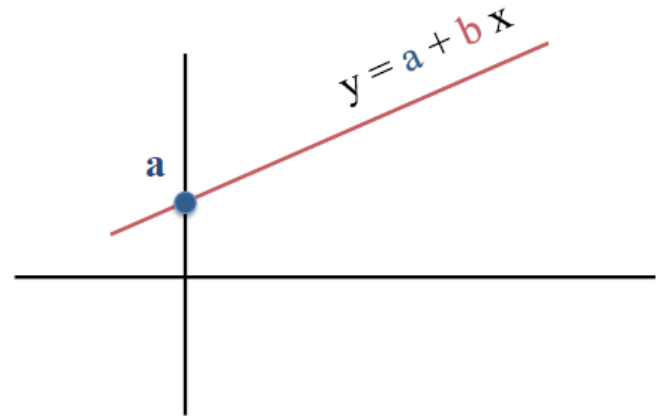
Simple linear regression refers to drawing a (particular, special) line through a scatterplot

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

$$y = a + b x$$

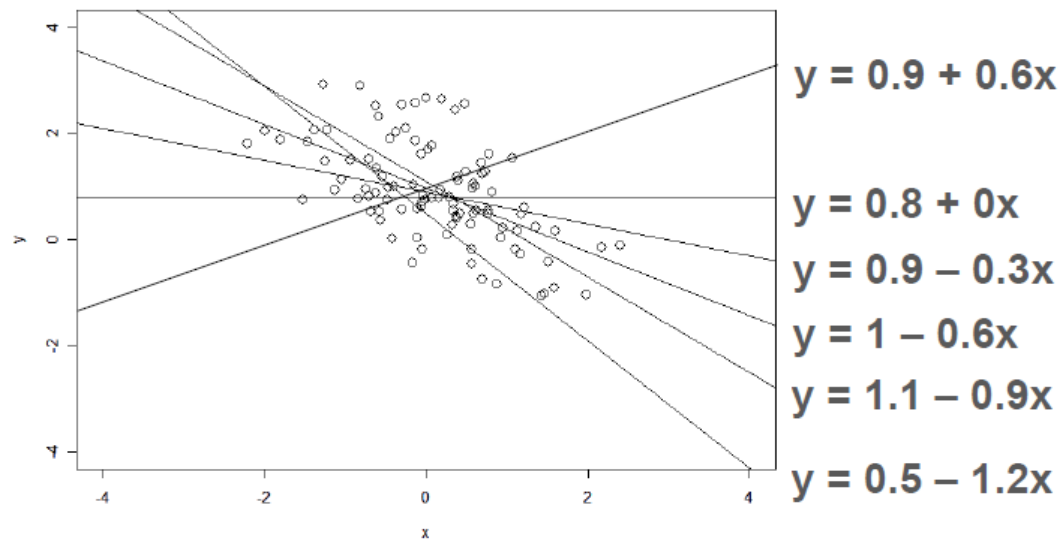
where a is called the **intercept** and b is the **slope**.



(Simple) Linear Regression

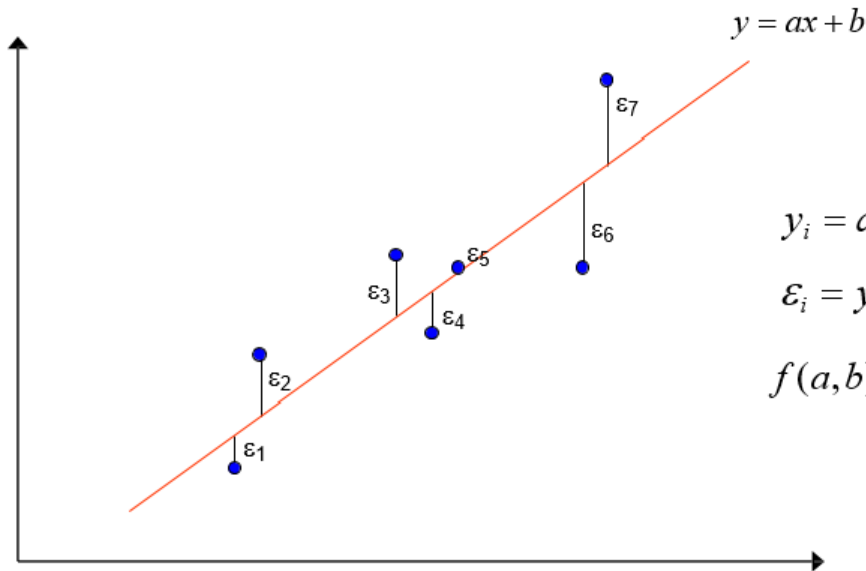
What is the “best” line which fits this data ?

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



Linear regression: least-squares fitting

Least-square fitting



Regression line
such that:

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

minimum

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

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

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

$$\frac{\partial f(a, b)}{\partial a} = 0$$

$$\frac{\partial f(a, b)}{\partial b} = 0$$

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

Linear regression: least-squares fitting

Formalization and extension of linear regression

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

$i = 1, \dots, n$

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)

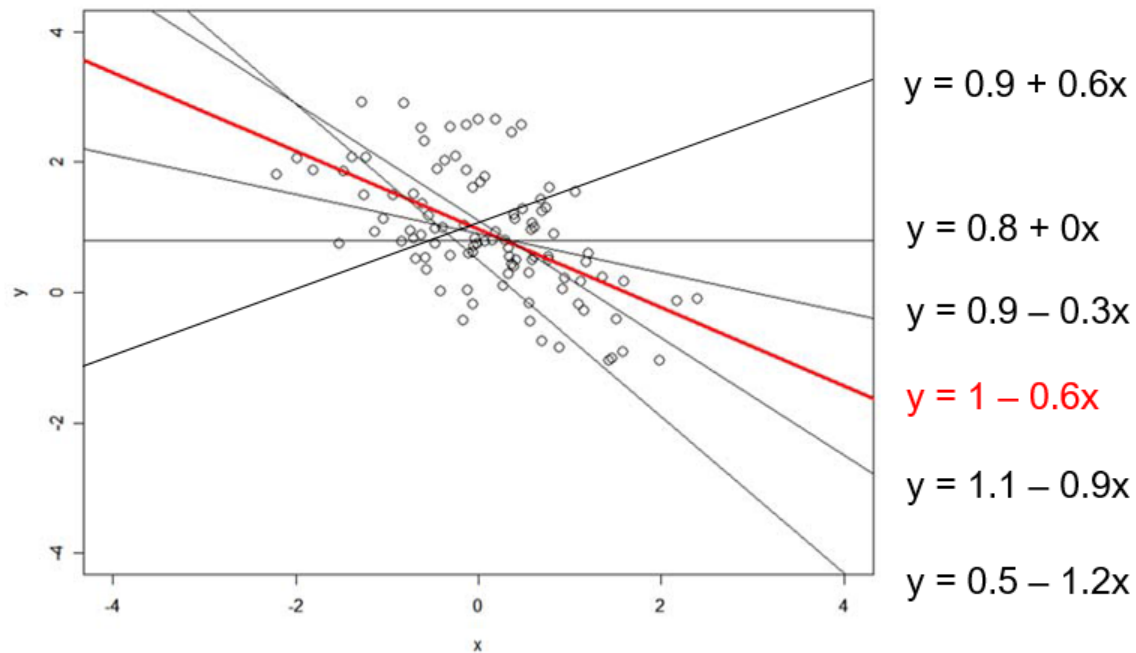
Minimizing $\sum_i \varepsilon_i^2$ yields b_0 and b_1 estimators of β_0 and β_1

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

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Linear regression: least-squares fitting

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



(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)

(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 is calculated so that the sum (and mean) of the residuals is 0 (« in average, the model is correct »).

Ideally, we want the regression to include all the predictable variance, so that the distribution of the residuals is random and does not depend on X or on the predicted Y.

Linear models (general case)

p parameter linear model

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

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

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

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

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

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

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

$$E\{\boldsymbol{\varepsilon}\} = \mathbf{0}$$

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

Linear models: parameter estimation

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

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

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

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

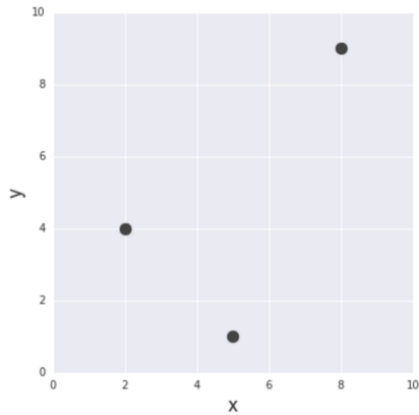
Linear models: linearity

Linearity is about the model parameters

$$\left. \begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ Y_i &= \beta \sin X_i + \varepsilon_i \end{aligned} \right\} \text{Linear in } \beta\text{s}$$

$$\left. \begin{aligned} Y_i &= \beta_0 + \log(\beta_1 X_{i1} + \beta_2 X_{i2}) + \beta_3 X_{i3} + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 \exp(\beta_2 X_i + \beta_3) + \varepsilon_i \end{aligned} \right\} \text{Not linear in } \beta\text{s}$$

Linear models: linearity

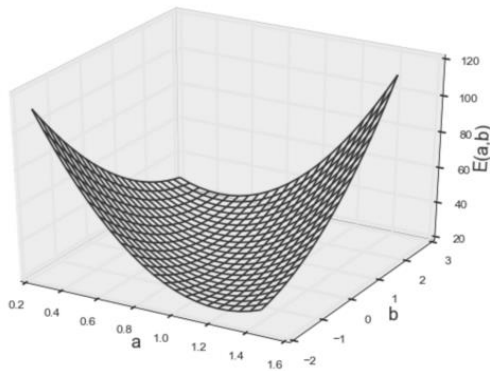


$$E(a, b) = \sum_{i=1}^3 (f(x_i) - y_i)^2$$

$$= \sum_{i=1}^3 (ax_i + b - y_i)^2$$

$$= (2a + b - 4)^2 + (5a + b - 1)^2 + (8a + b - 9)^2$$

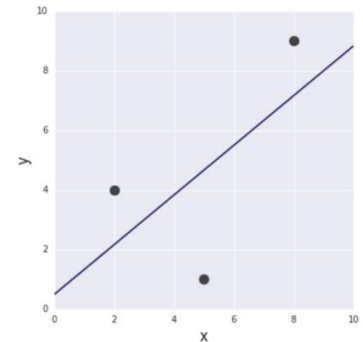
$$= 93a^2 + 3b^2 + 30ab - 170a - 28b + 98$$



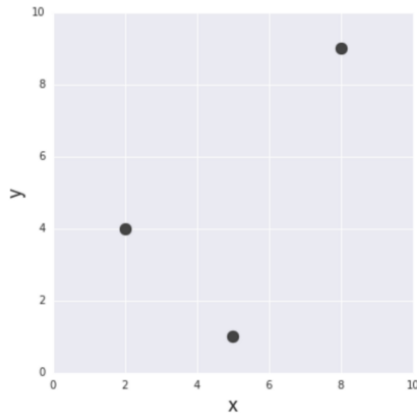
$$\frac{\partial}{\partial a} E(a, b) = 186a + 30b - 170 = 0$$

$$\frac{\partial}{\partial b} E(a, b) = 6b + 30a - 28 = 0$$

$$f(x) = \frac{5}{6}x + \frac{1}{2}$$



Linear models: linearity

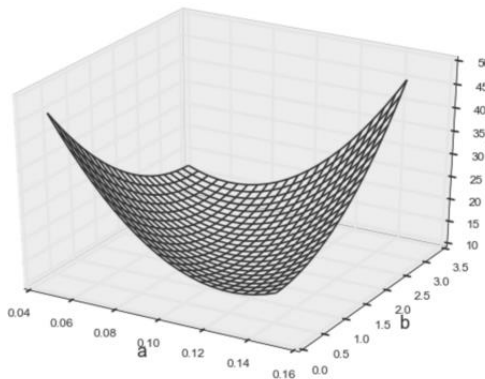


$$E(a, b) = \sum_{i=1}^3 (f(x_i) - y_i)^2$$

$$= \sum_{i=1}^3 (ax_i^2 + b - y_i)^2$$

$$= (4a + b - 4)^2 + (25a + b - 1)^2 + (64a + b - 9)^2$$

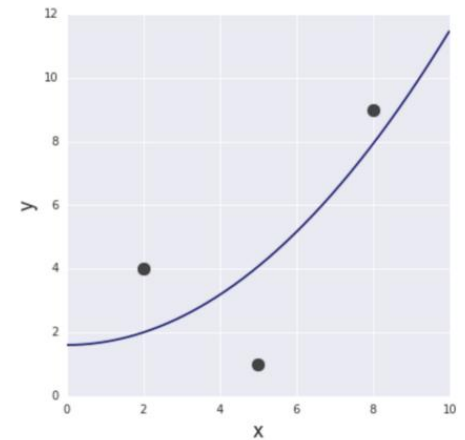
$$= 4737a^2 + 3b^2 + 186ab - 1234a - 28b + 98$$



$$\frac{\partial}{\partial a} E(a, b) = 9474a + 186b - 1234 = 0$$

$$\frac{\partial}{\partial b} E(a, b) = 6b + 186a - 28 = 0$$

$$f(x) = \frac{61}{618}x^2 + \frac{331}{206}$$



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

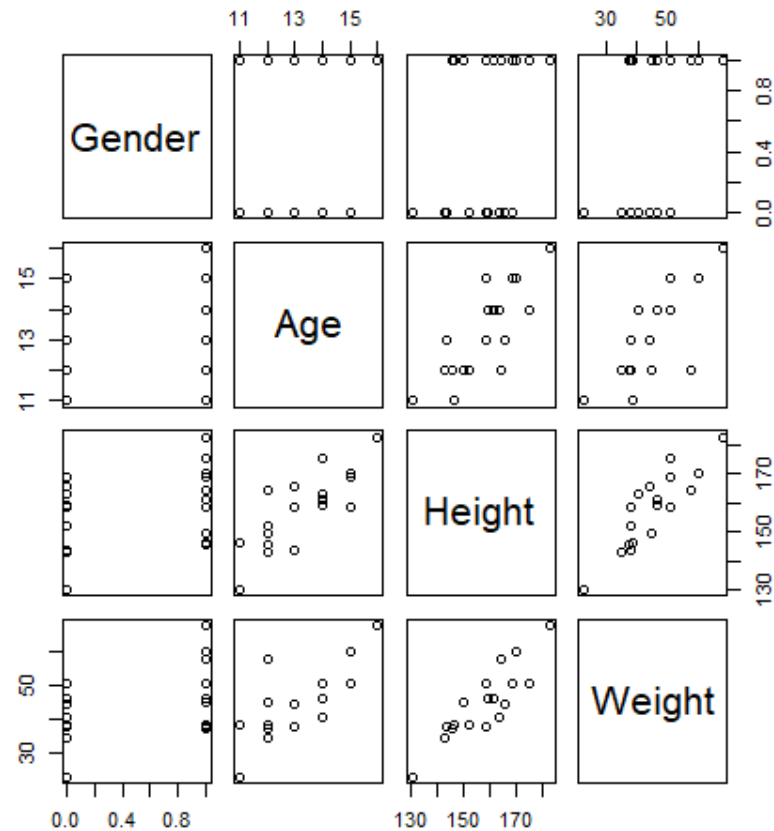
The CLASS dataset from SAS

```
> 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

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

Call:

```
lm(formula = Height ~ Age, data = class)
```

Coefficients:

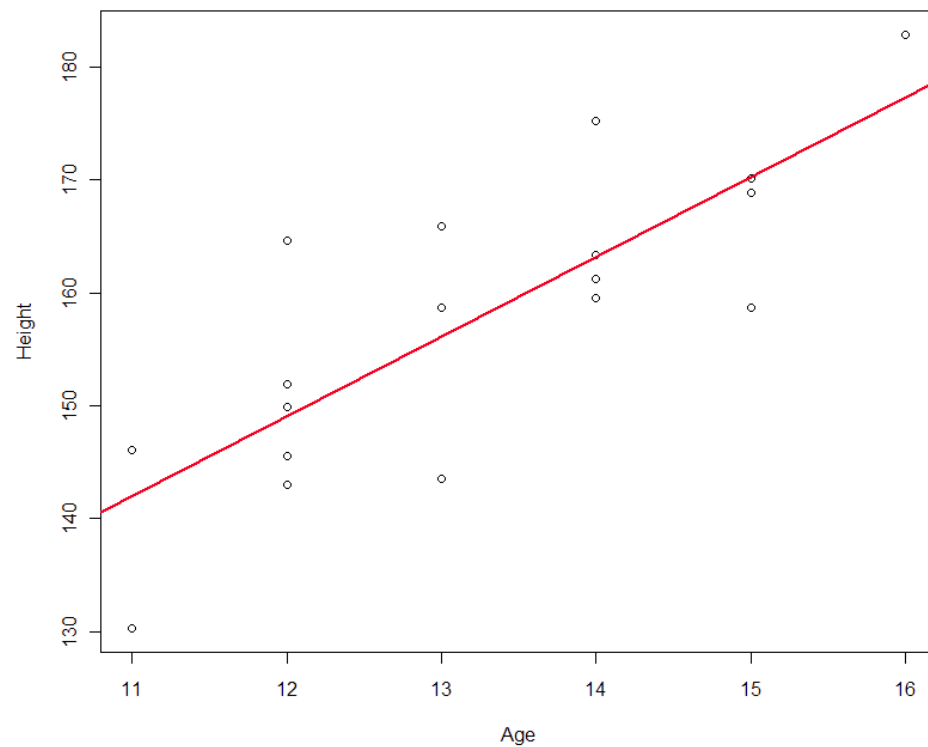
(Intercept)	Age
64.07	7.08

Model: Height = 64.07 + 7.08 x Age

The CLASS dataset from SAS

```
plot( Age, Height )
```

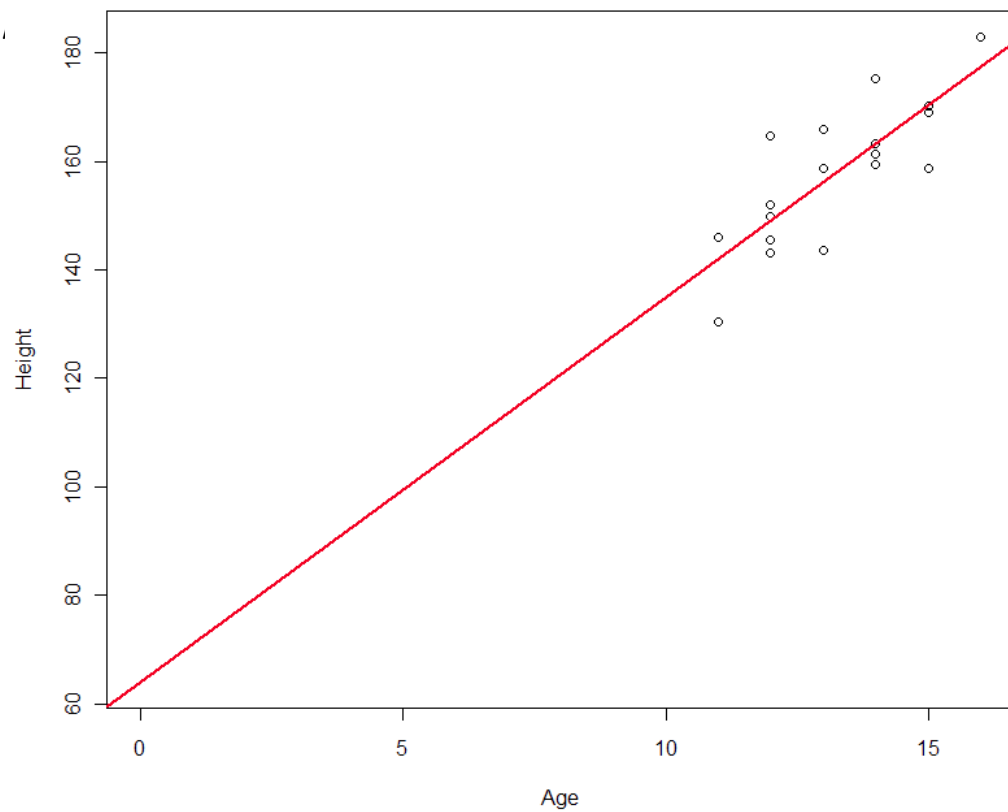
```
abline(model, col="red", lwd=2)
```



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

```
> 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

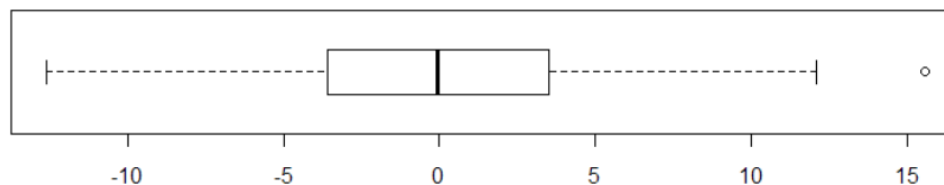
The CLASS dataset from SAS

**Five-number summary of the residuals
(but no mean – why ?), equivalent to**

```
> fivenum( residuals( model ) )  
      8      11      17      4      7  
-12.590 -3.573 -0.078  3.490 15.571
```

or, graphically, using a boxplot:

```
> boxplot( residuals ( model), horizontal=T)
```



The CLASS dataset from SAS

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

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

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

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

** It is not interesting for the intercept, but usually interesting for the slope.

Estimate and Std. Error are used for hypothesis testing

t-value = Estimate / Std. Error (of this estimate), should follow a t-distribution

(assumes that the residuals 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 (which we would usually like to be small)

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, and the model explains all the variance).

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

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

```
> summary(model)$r.squared [1] 0.6584257  
> cor(Age, Height)^2 [1] 0.6584257
```

The Adjusted R-squared is similar to R-squared, but it takes into account the number of variables in the model (we will come back to this later).

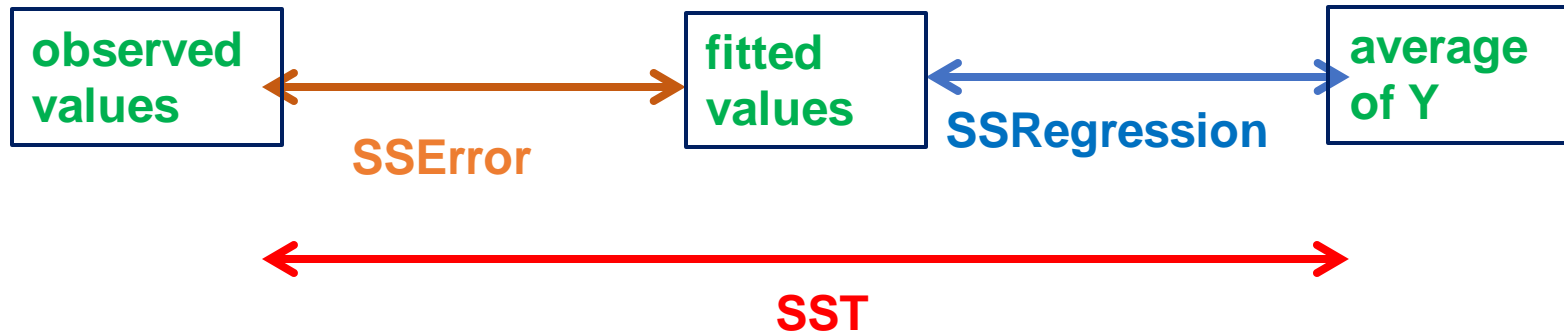
$$SST = SSR + SSE$$

Total sum of squares = regression SS + residual SS

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

$$R^2 = SSR / SST$$

HEURISTIC REPRESENTATION



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

```
> 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 \pm 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 ; \pm 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

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:

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

SST SSR SSE

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

Test Distributions

t-distribution

Student's *t*-distribution (or simply the ***t*-distribution**) is any member of a family of continuous [probability distributions](#) that arise when estimating the [mean](#) of a [normally distributed population](#) in situations where the [sample size](#) is small and the population's [standard deviation](#) is unknown. It was developed by English statistician [William Sealy Gosset](#) under the pseudonym "Student".

[\[https://en.wikipedia.org/wiki/Student's_t-distribution\]](https://en.wikipedia.org/wiki/Student's_t-distribution)

Test Distributions

F-distribution

In [probability theory](#) and [statistics](#), the **F-distribution** or **F-ratio**, also known as **Snedecor's F distribution** or the **Fisher–Snedecor distribution** (after [Ronald Fisher](#) and [George W. Snedecor](#)) is a [continuous probability distribution](#) that arises frequently as the [null distribution](#) of a [test statistic](#), most notably in the [analysis of variance](#) (ANOVA)

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of

$$X = \frac{S_1/d_1}{S_2/d_2}$$

where S_1 and S_2 are independent [random variables](#) with [chi-square distributions](#) with respective degrees of freedom d_1 and d_2 .

[\[https://en.wikipedia.org/wiki/F-distribution\]](https://en.wikipedia.org/wiki/F-distribution)

Test Distributions

Chi square distribution

In [probability theory](#) and [statistics](#), the **chi-squared distribution** (also **chi-square** or χ^2 -**distribution**) with k [degrees of freedom](#) is the distribution of a sum of the squares of k [independent standard normal](#) random variables.

If Z_1, \dots, Z_k are [independent, standard normal](#) random variables, then the sum of their squares,

$$Q = \sum_{i=1}^k Z_i^2,$$

is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2.$$

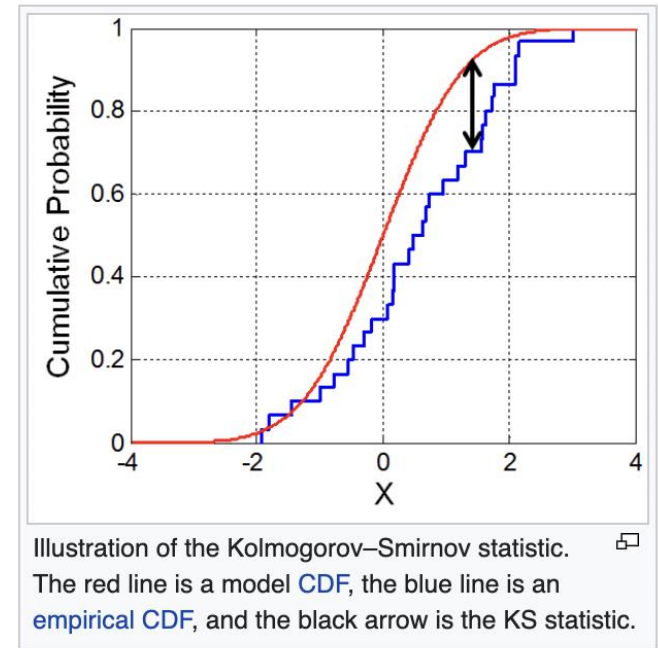
The chi-squared distribution has one parameter: a positive integer k that specifies the number of [degrees of freedom](#) (the number of random variables being summed, Z_i s).

[\[https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test\]](https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test)

KS Test

Chi square distribution

The **Kolmogorov–Smirnov test** (K-S test or KS test) is a [nonparametric test](#) of the equality of one-dimensional [probability distributions](#) that can be used to compare a [sample](#) with a reference probability distribution (one-sample K–S test), or ...



The Kolmogorov–Smirnov [statistic](#) for a given [cumulative distribution function](#) $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|$$

The [null distribution](#) of this statistic is calculated under the [null hypothesis](#) that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case).

[\[https://en.wikipedia.org/wiki/Chi-squared_distribution\]](https://en.wikipedia.org/wiki/Chi-squared_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 F-statistic allows us to test if the whole regression (adding all variables vs having only the intercept in) is significant.

With only one variable, it provides *exactly* the same result as the t-test for the significance of the coefficient of this variable.

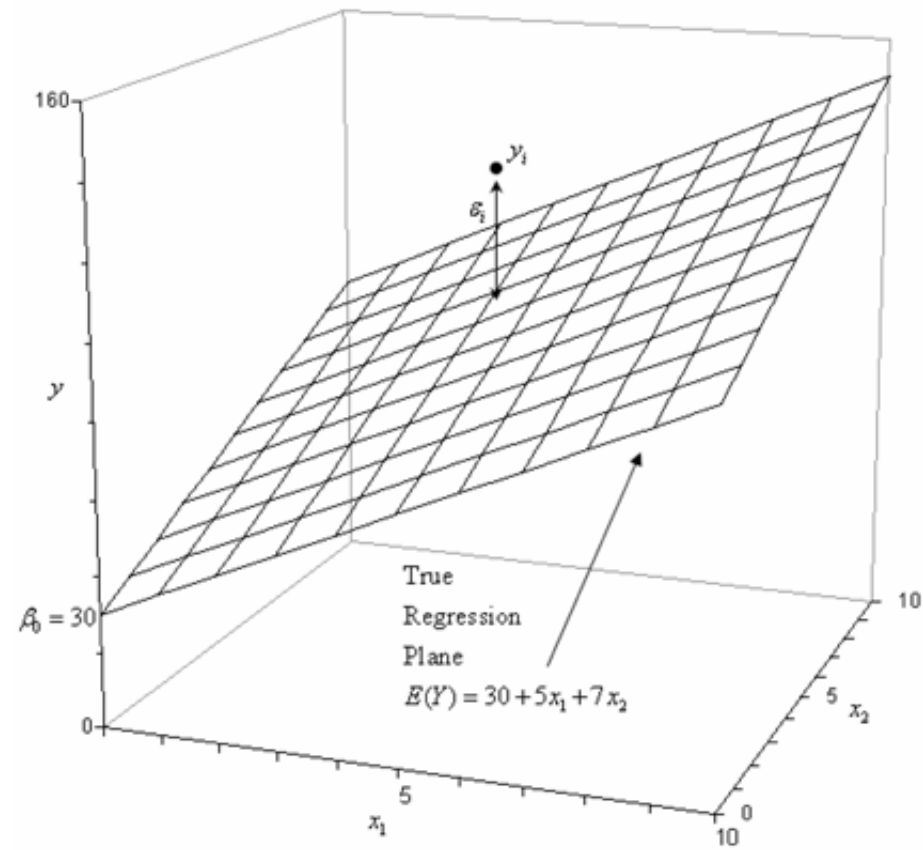
Coefficients:

	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

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

Multiple linear regression



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

Call:

```
lm(formula = Height ~ Age + Weight)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-9.20695	-3.30604	-0.04478	2.11432	10.41880

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

Residual standard error: 5.728 on 16 degrees of freedom

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

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

**This model allows us to determine the respective
contribution of each variable separately.**

Separately ??

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

This is similar to the simple regression case.

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 (see also the R²).

It is likely 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 .

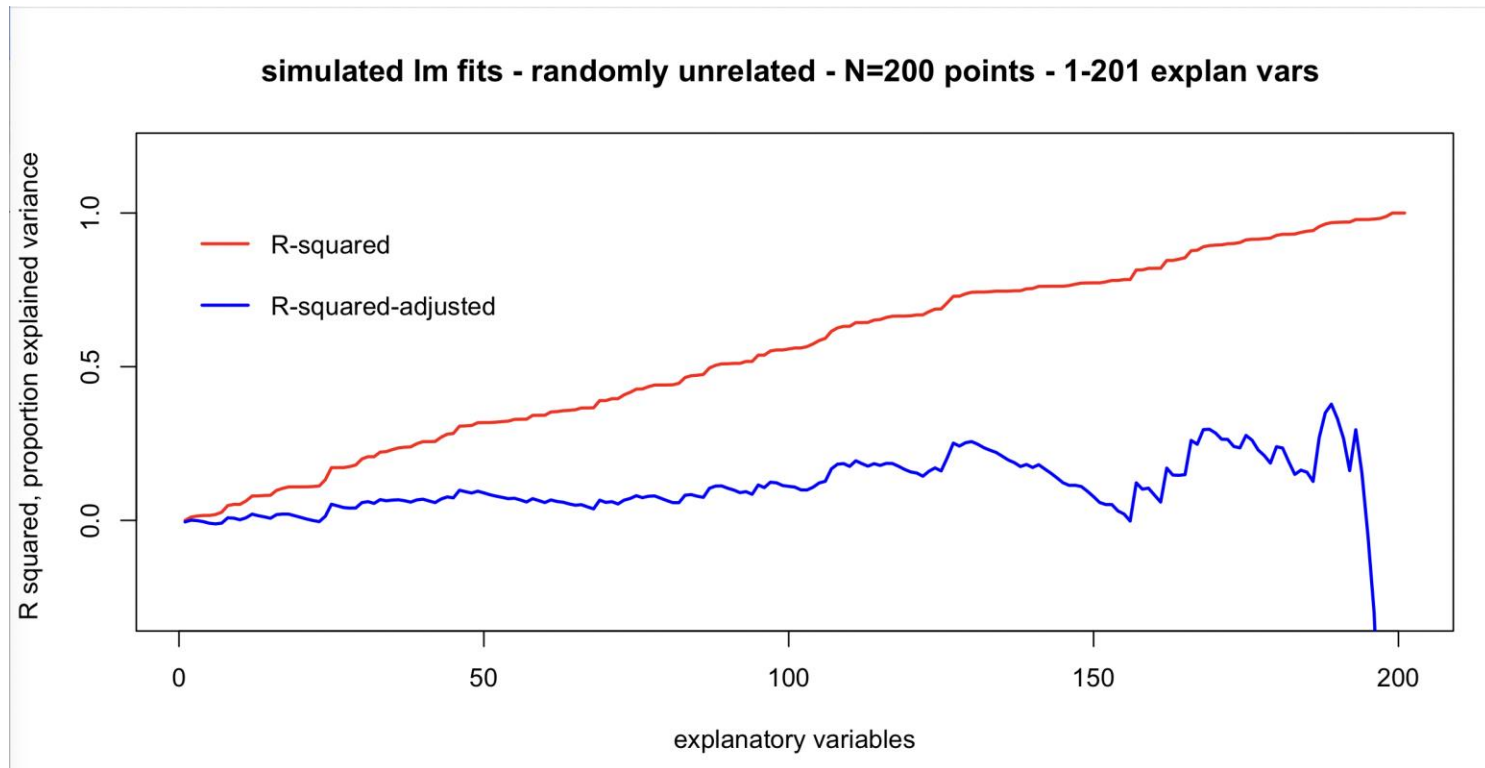
One multiple regression with n variables

```
y <- rnorm(10)  
x1 <- rnorm(10); x2 <- rnorm(10); ... ; x9 <-rnorm(10)  
summary(lm(y ~ x1)); summary(lm(y ~ x1+x2));...
```

```
1: Multiple R-squared: 0.142, Adjusted R-squared: 0.035  
2: Multiple R-squared: 0.517, Adjusted R-squared: 0.379  
3: Multiple R-squared: 0.557, Adjusted R-squared: 0.502  
4: Multiple R-squared: 0.558, Adjusted R-squared: 0.204  
5: Multiple R-squared: 0.795, Adjusted R-squared: 0.539  
6: Multiple R-squared: 0.832, Adjusted R-squared: 0.496  
7: Multiple R-squared: 0.984, Adjusted R-squared: 0.928  
8: Multiple R-squared: 0.985, Adjusted R-squared: 0.865  
9: Multiple R-squared: 1.000, Adjusted R-squared: NaN
```

Adjusted R squared simulation

New example



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:

$$\text{Height} = b_0 + b_1 \text{ Age} + b_2 \text{ « Gender »} + \text{error}$$

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.

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

and the model can be interpreted as follows:

- b_0 is the baseline for height among women
- 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

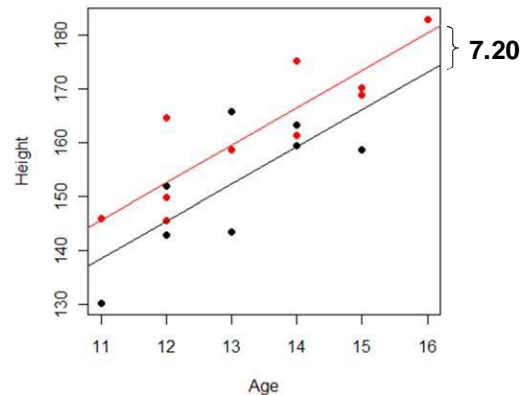
The factor
GenderM
corresponds to
the difference in
baseline for Males
compared to
females

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)



Categorical variables

We could also compute the difference in means between males and females directly:

```
> means <- tapply( data$Height, data$Gender, FUN=mean)
> means
      F      M
153.8958 162.3314
> diff(means)
      M
8.435622
```

This result is slightly different from the 7.20 cm difference found with the linear model.

Where does the difference come from ?

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.

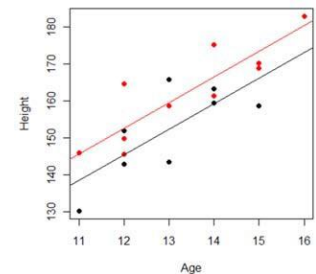
If this assumption is incorrect, it means that there is an **interaction** between the factors « age » and « gender », that is, the effect of age is different depending on the gender.

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

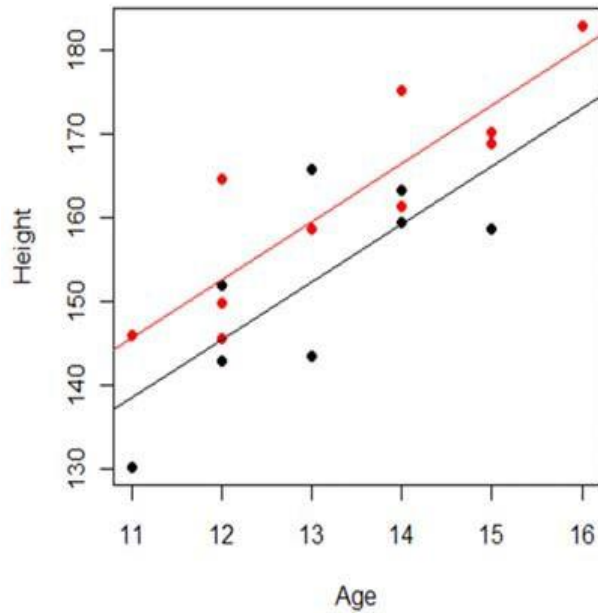
baseline for
height among
Female

difference in
baseline for Males
compared to females

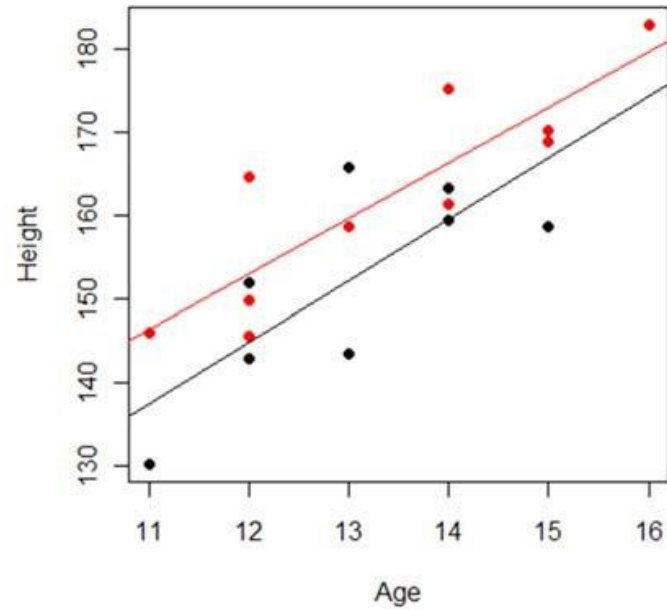
age effect only for
males

$\text{Height} = 56.26 + 7.38 \cdot \text{Age} + 17.13 \text{ (only for males)} - 0.75 \cdot \text{Age} \text{ (only for males)}$

Interaction



No interaction



With interaction

What if Males were the baseline ?

```
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
```

The two models are exactly the same; only the way we look at the coefficient changes.

```
Call:
lm(formula = Height ~ Age + Gender1)

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)  69.495     15.135   4.592  0.000301 ***
Age          6.928       1.117   6.202  1.27e-05 ***
Gender1F     -7.204       3.251  -2.216  0.041517 *
---
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
```

```
Gender1 <- relevel(Gender, ref="M")
```

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).

Linear models (matrix form)

Matrix form of linear models

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

is equivalent to

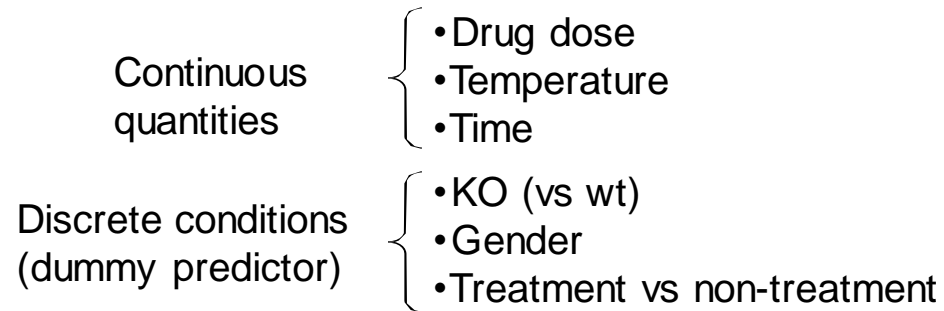
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\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

or

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

Continuous versus dummy predictors

X is the **design matrix**; a column of X_{ij} can be used to encode



Example of eye colors

Black	0	0
Blue	1	0
Green	1	1

Discrete conditions require “zeros and ones” coding.

Reference condition coded as zero, alternative coded as one.

Discrete conditions with N levels require N-1 columns with 0/1.

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 $E_i \sim \text{i.i.d } N(0, \text{sigma constant} | i)$

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 (spread, dispersion): maybe we are using a model that is not perfect, could improve ? (need another explanatory variable ?, a data transformation might help ?, maybe there are some outliers or "bad points" ?, or measures were not independent from each other ?
- If they are not normally distributed (residuals behave differently than "expected"): similar questions as above
- If they are not independent (hard to test additionally): similar questions
- If there are Outliers: which points / why ? Repeat / eliminate ? , and similar questions as above; and could consider using "robust regression" methods (use of modified least squares criteria)

Residuals

Type of *Residuals*:

- Raw residuals $R_i = \text{Observed} - \text{Fitted (Expected)} = Y - Y_i$
- Rescaled (specifically to each data point) to have expected sd = 1 :
Studentized Residuals (standardized R.)
 - Should follow about a t-distribution resp. approx. $N(0,1)$ (under assumptions)

Basic model checking

Examination of *Influence*:

- Are there “overly influential points” ?, which / why ? Bad “design” ? Repeat / eliminate ? Detection of *influential observations*: *Hat matrix*

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 i th observation to the j th fitted value

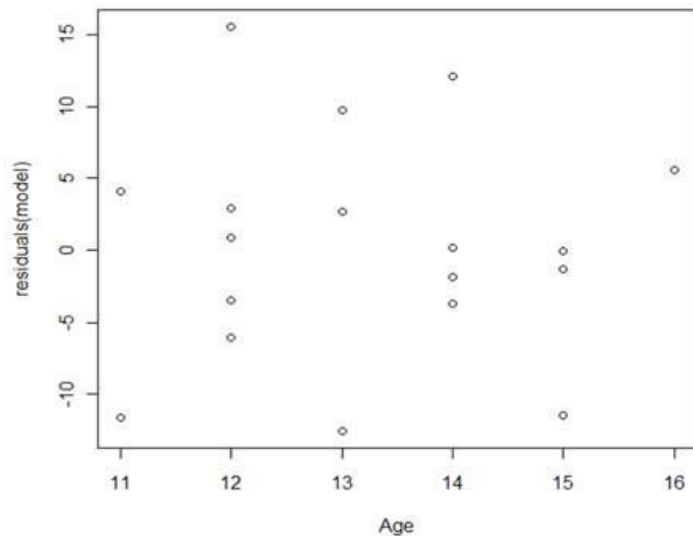
Most informative are the diagonal values of the hat matrix

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

Average value of $h = \text{nb of predictors } p / \text{nb of points } n = p/n$

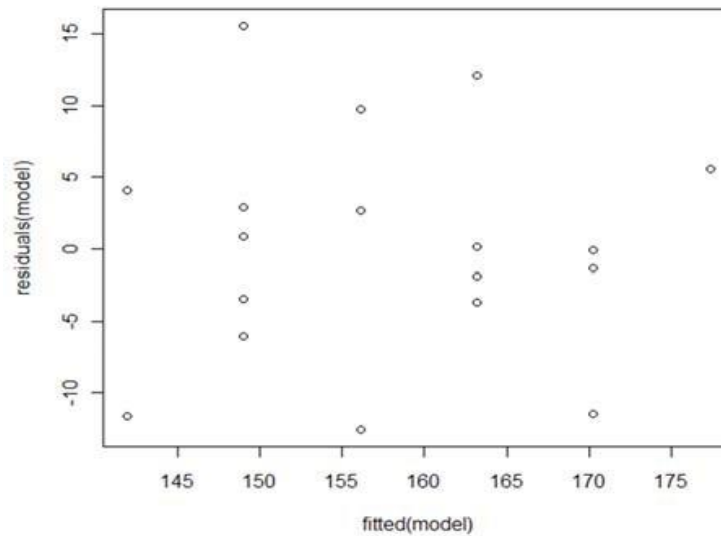
Cutoff typically $2 \cdot p/n$ to $3 \cdot p/n$: points beyond are considered worthy of a careful examination

Residuals



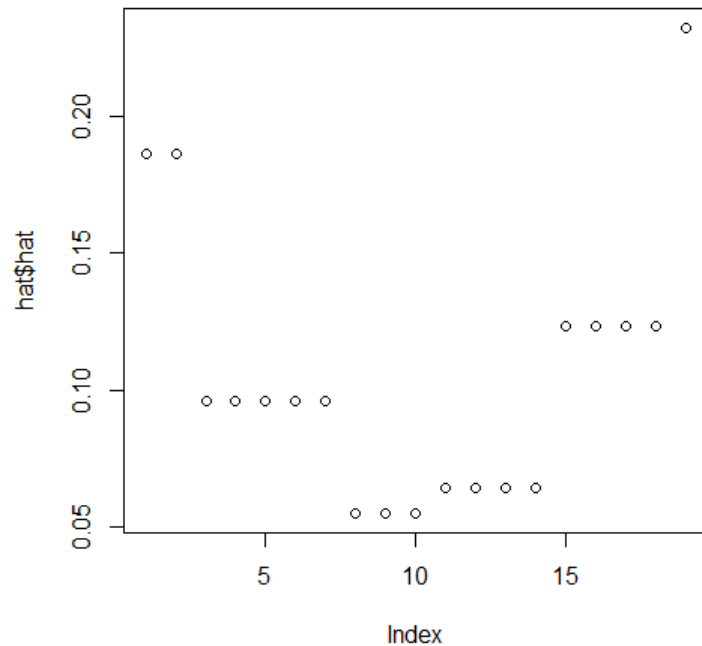
```
plot( Age, residuals(model) )
```

Works only for simple regression (only one variable on x axis)



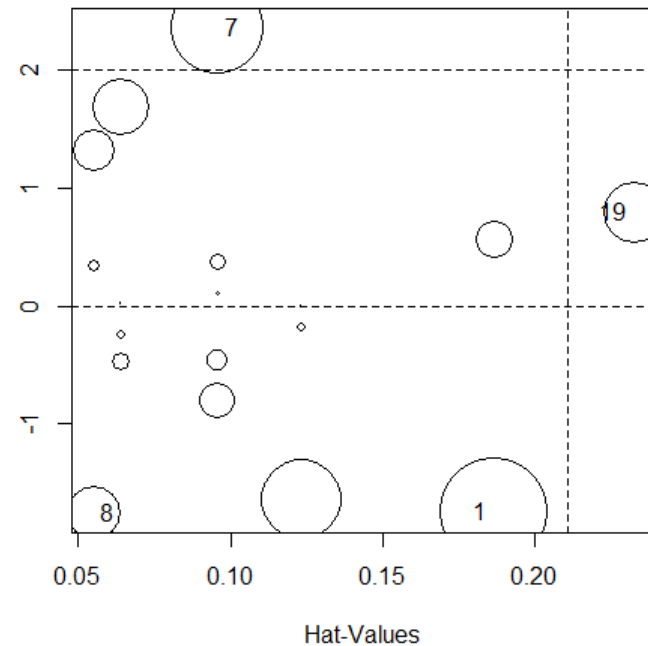
```
plot( fitted(model), residuals(model) )
```

Works also for multiple regression



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

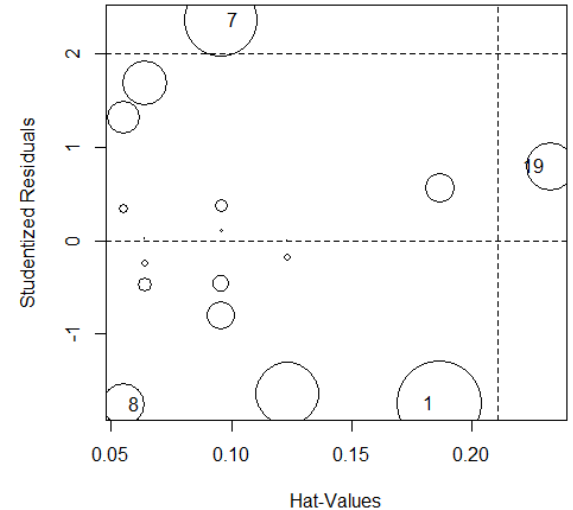


```
hat <- lm.influence(model)
plot(hat$hat)
```

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

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.

CONFIDENCE INTERVALS FOR THE FITTED PARAMETER VALUES

Usual method

Estimate \pm 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 ; \pm 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 \pm 2.610 ; [4.469, 9.689]

> confint(mm)

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

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 $CI \rightarrow 0$ as $n \rightarrow \infty$

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 $\rightarrow 0$ as $n \rightarrow \infty$

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 $1 - \alpha$ 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 $1 - \alpha$ 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 $\rightarrow 0$ as $n \rightarrow \infty$

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

CONFIDENCE INTERVALS FOR SINGLE VALUES, CURVE AND PREDICTIONS

Analogy

Data $\sim N(\mu, \sigma)$

Estimate μ with a $\hat{\mu}$

confidence interval for $\hat{\mu}$ to include real μ :

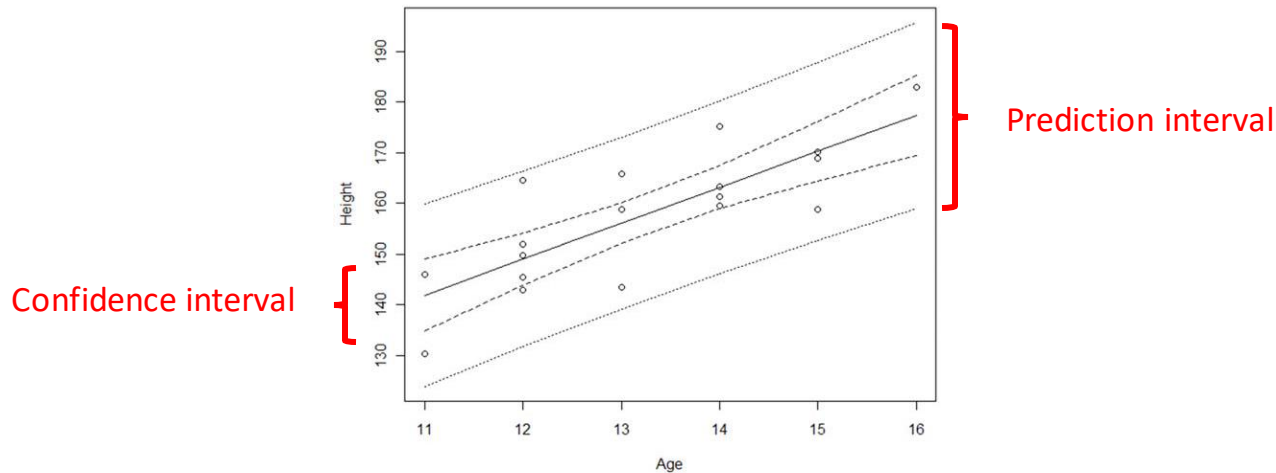
this gets smaller with increasing sample size (standard errors)

and is smaller than σ (usually $\sim \sigma / \sqrt{n}$) }

New data are $\sim N(\hat{\mu}, \sigma)$

scattering width is due to (imprecision in $\hat{\mu}$) + σ

Confidence bands



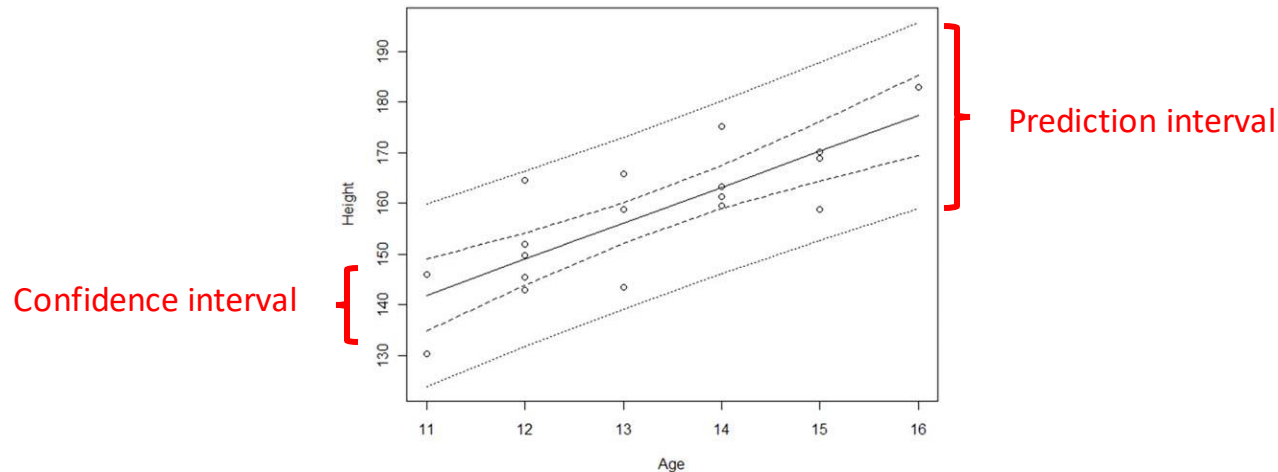
Narrow bands:

Wide bands:

describe the uncertainty about the regression line
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 bands



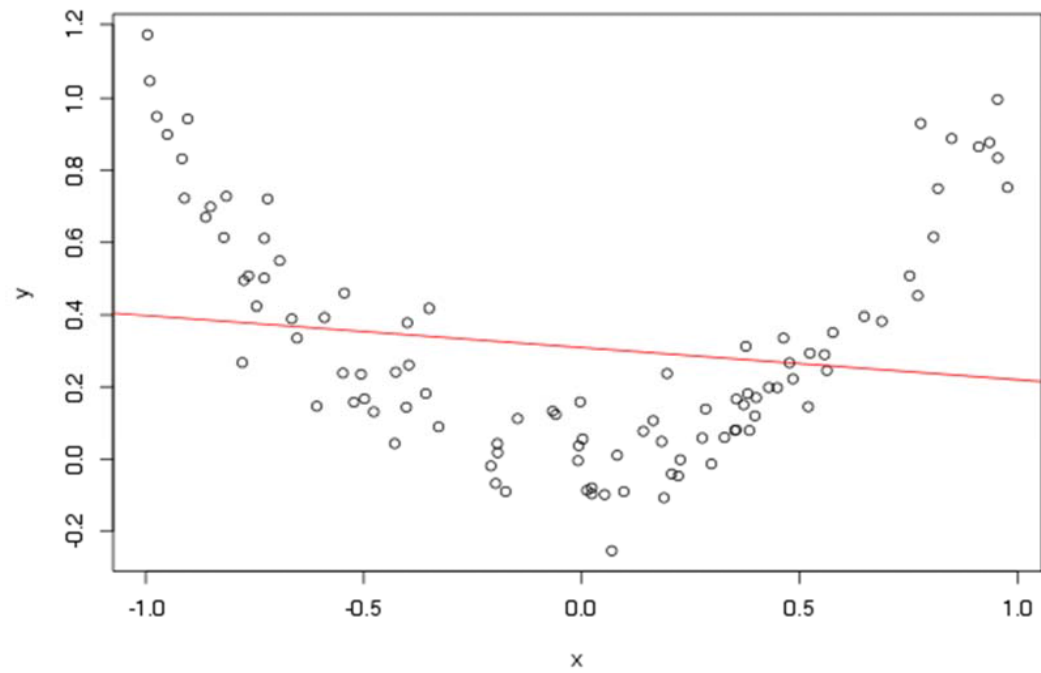
Narrow bands:

Wide bands:

describe the uncertainty about the regression line
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 ?

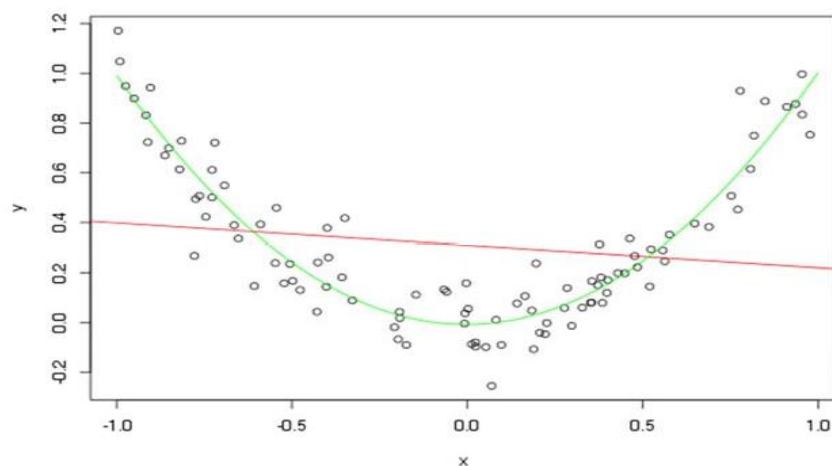


What if the data is not linear ?

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.

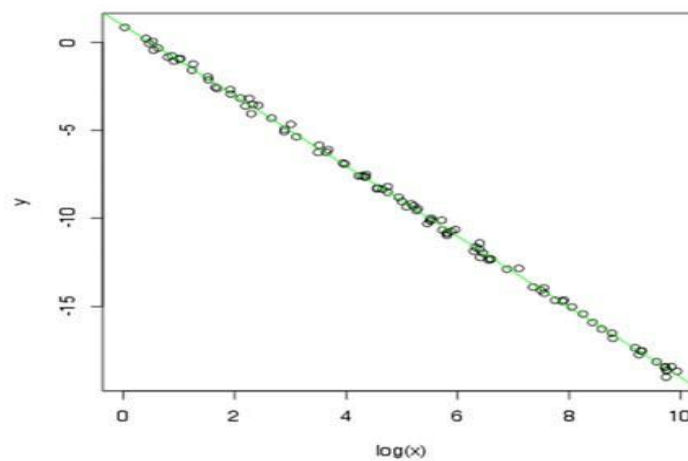
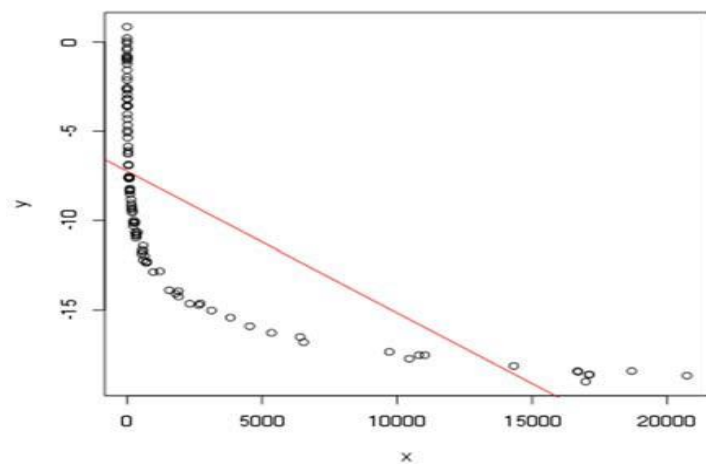


What if the data is not linear ?

Consider transforming the data (log)

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

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



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)
```

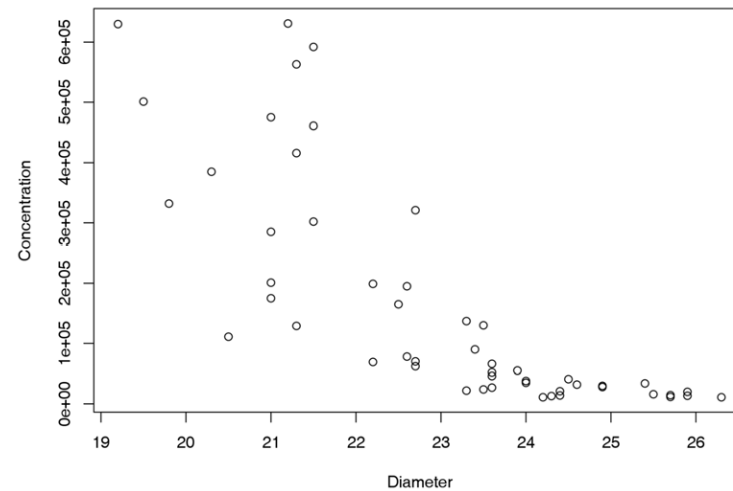
**Can we predict the concentration of cells
using the diameter and the
presence/absence of glucose ?**

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
[...]
```

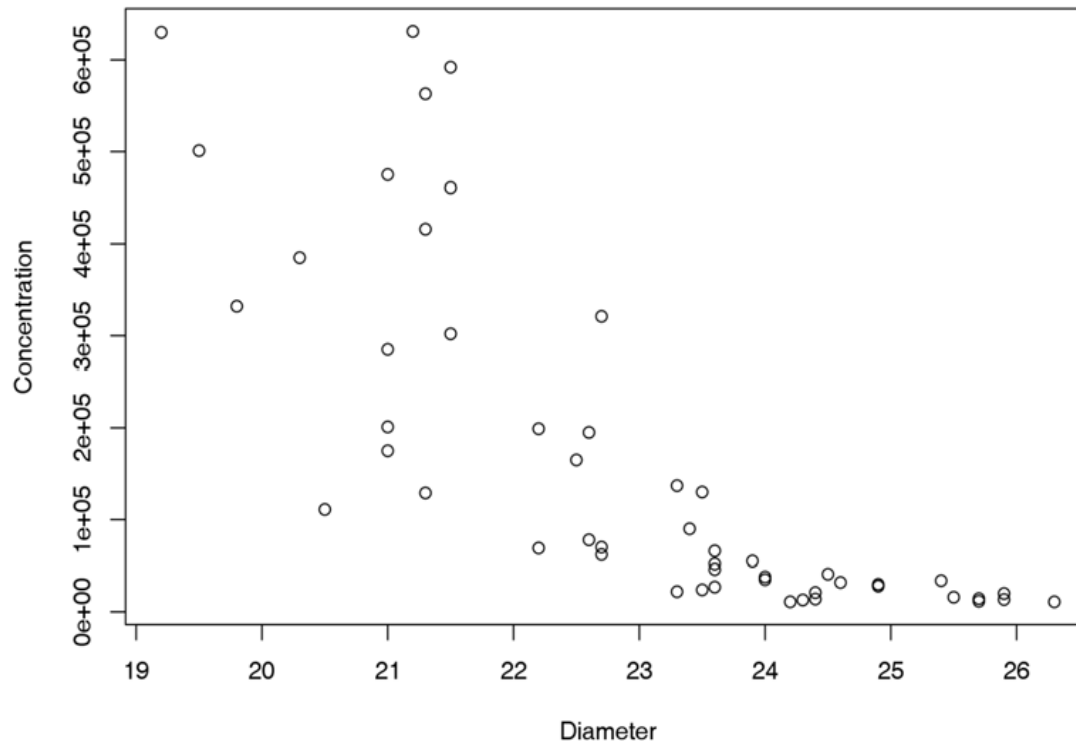
33	2	630000	19.2
34	2	501000	19.5
35	2	332000	19.8
36	2	285000	21.0
37	2	201000	21.0

Hellung dataset: Diameter vs Concentration

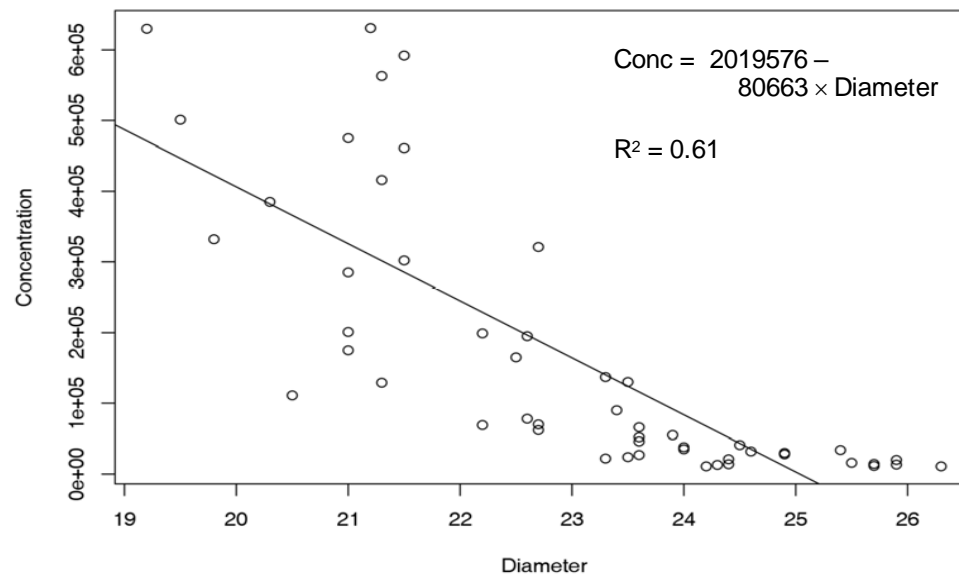


```
> plot(hellung$diameter, hellung$conc,  
       xlab="Diameter",  
       ylab="Concentration")
```

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



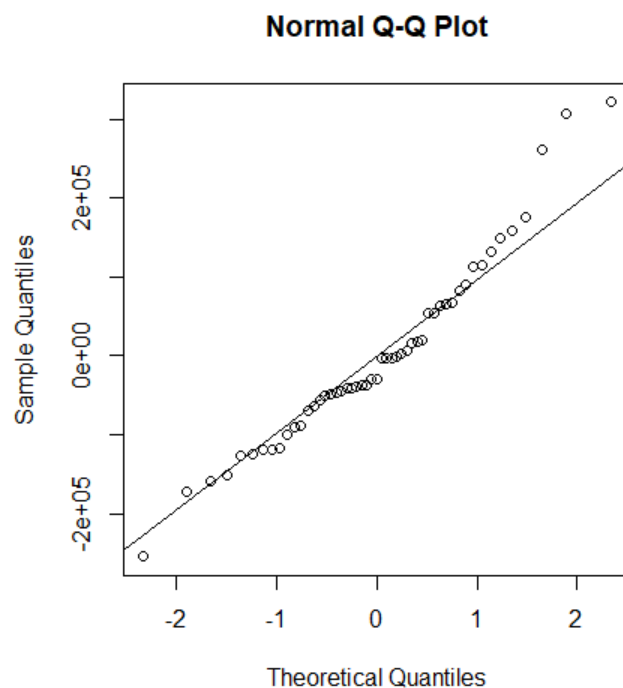
Linear model predicting Concentration from Diameter



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

Do the residuals follow a normal distribution ?

```
> qqnorm(residuals(model))  
> qqline(residuals(model))
```



```
> ks.test(residuals(model), "pnorm")
```

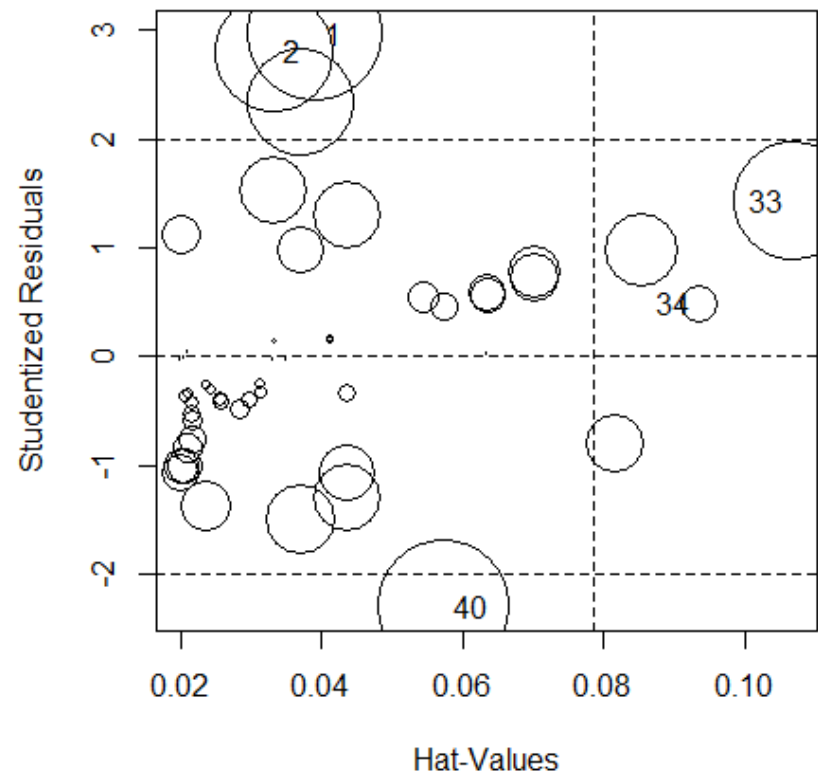
One-sample Kolmogorov-Smirnov test

```
data: residuals(model)  
D = 0.58824, p-value = 6.661e-16  
alternative hypothesis: two-sided
```

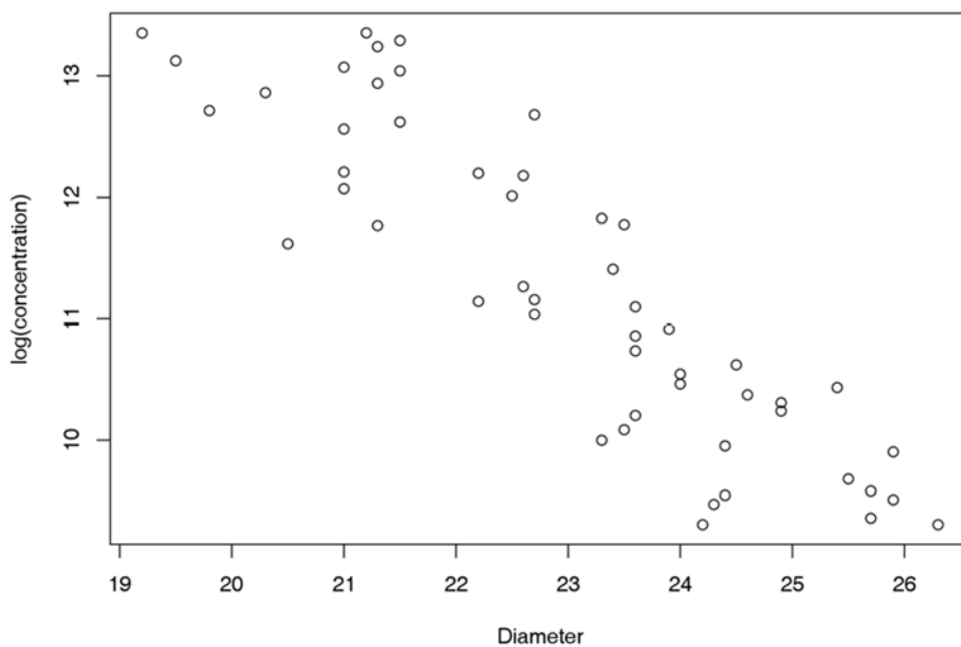
Residuals and hat values

```
> 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

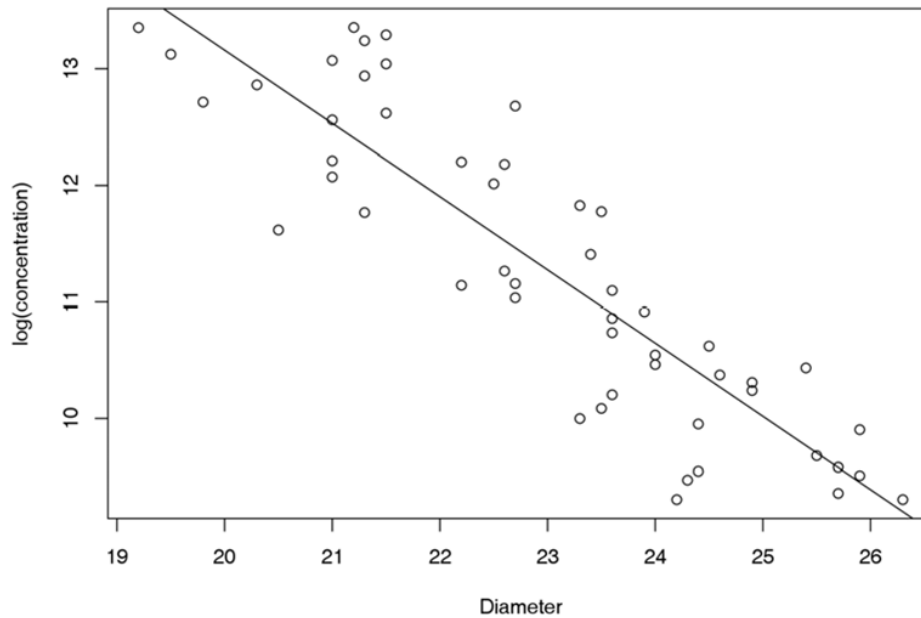


Transforming the data to improve the fit



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

Linear model predicting log(Concentration) from Diameter



$$\log(\text{conc}) = 25.7 - 0.62 \times \text{Diameter}$$

```
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 ***

```
---
```

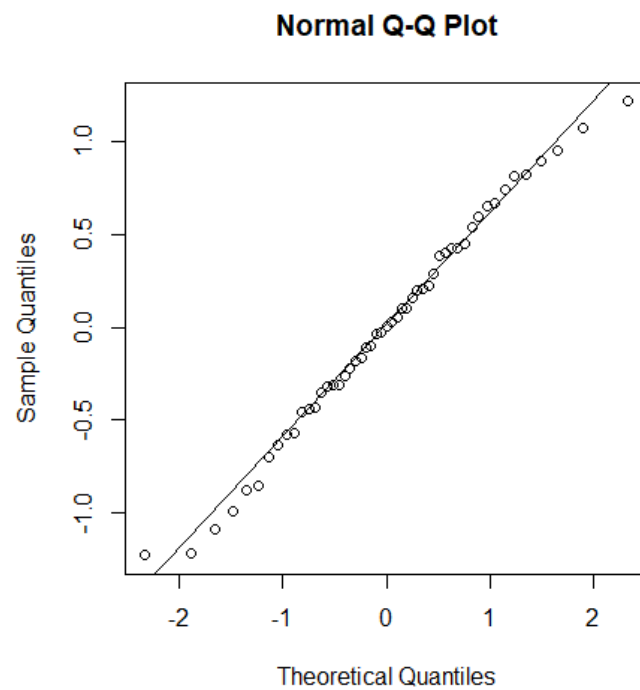
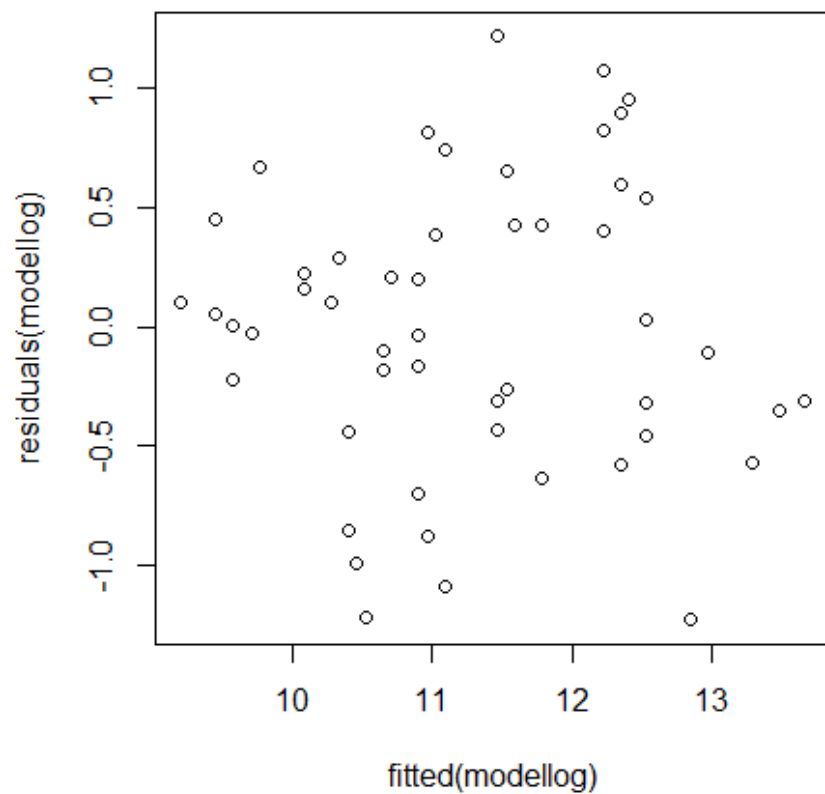
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6105 on 49 degrees of freedom
```

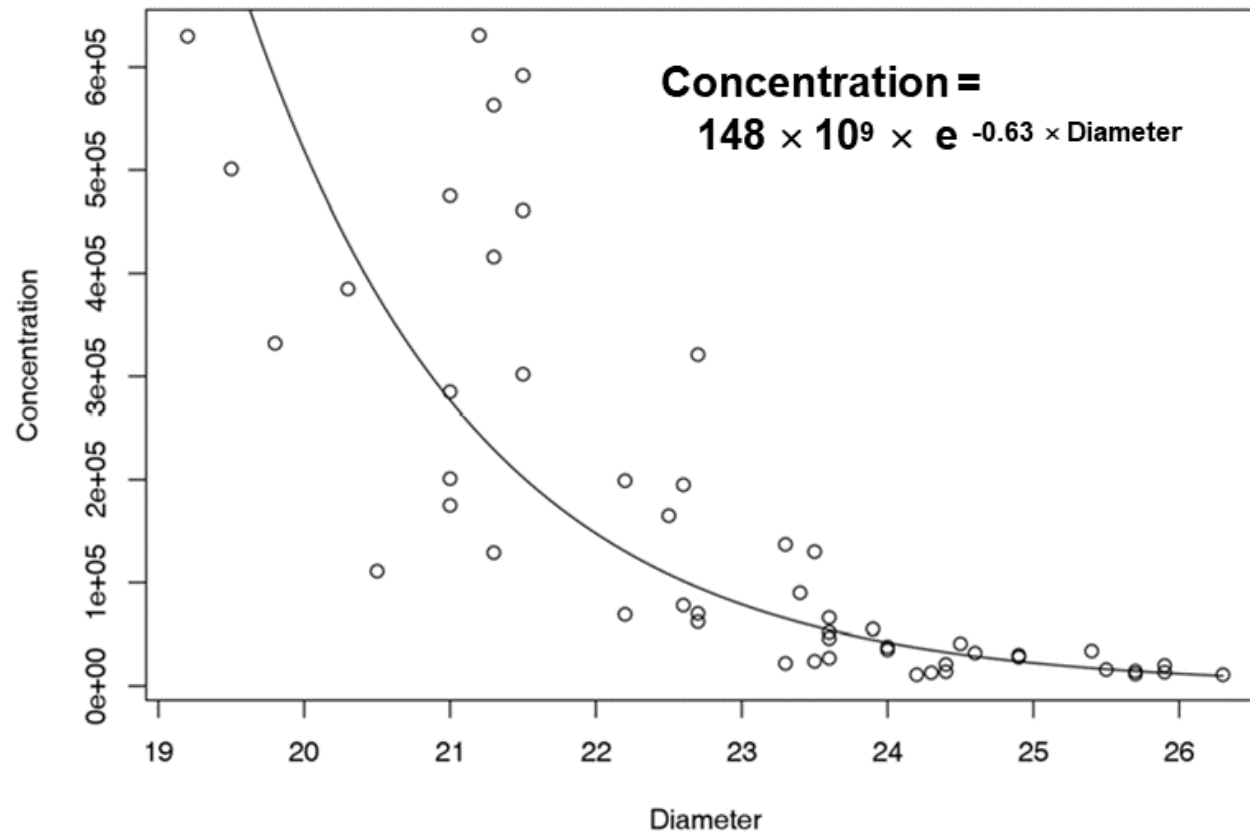
```
Multiple R-squared: 0.7817,    Adjusted R-squared: 0.7772
```

```
F-statistic: 175.4 on 1 and 49 DF,  p-value: < 2.2e-16
```

Diagnostic plots



Predicting Concentration from diameter



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:

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

The Hellung data in R

hellung

package:ISwR

R Documentation

```
> hellung
```

Growth of Tetrahymena cells

	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
----	---	--------	------

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.

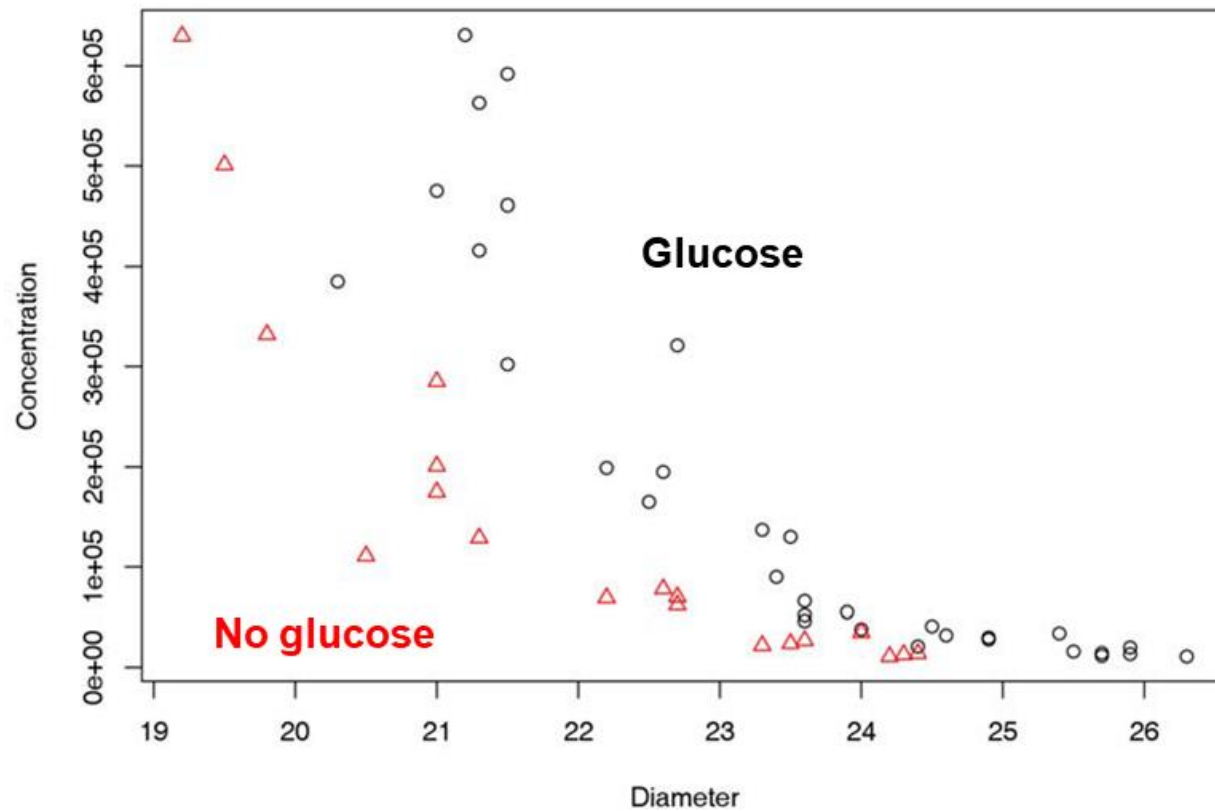
'conc' a numeric vector, cell concentration (counts/ml).

'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



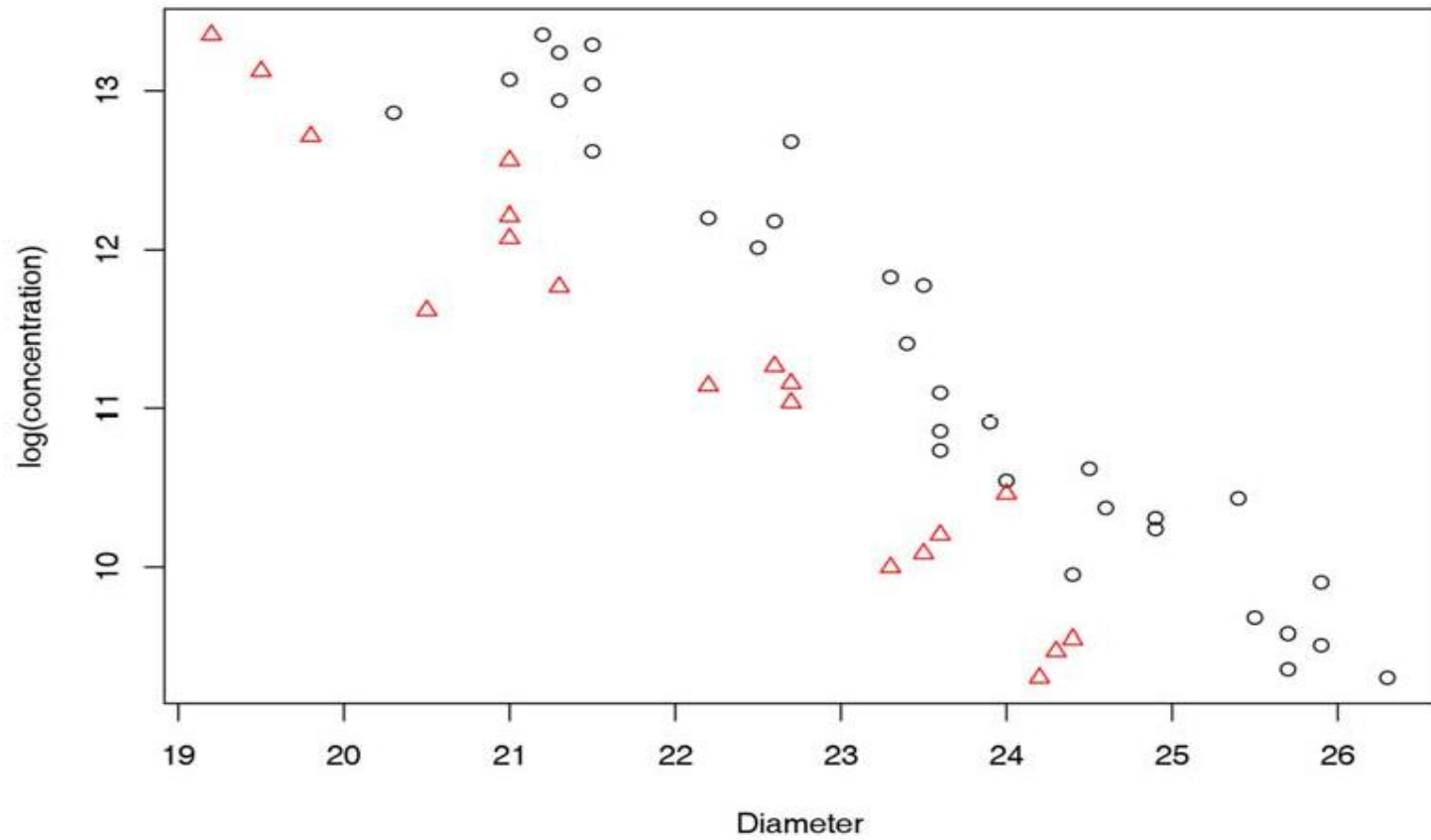
Reminder: using categorical variables as explanatory variables

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

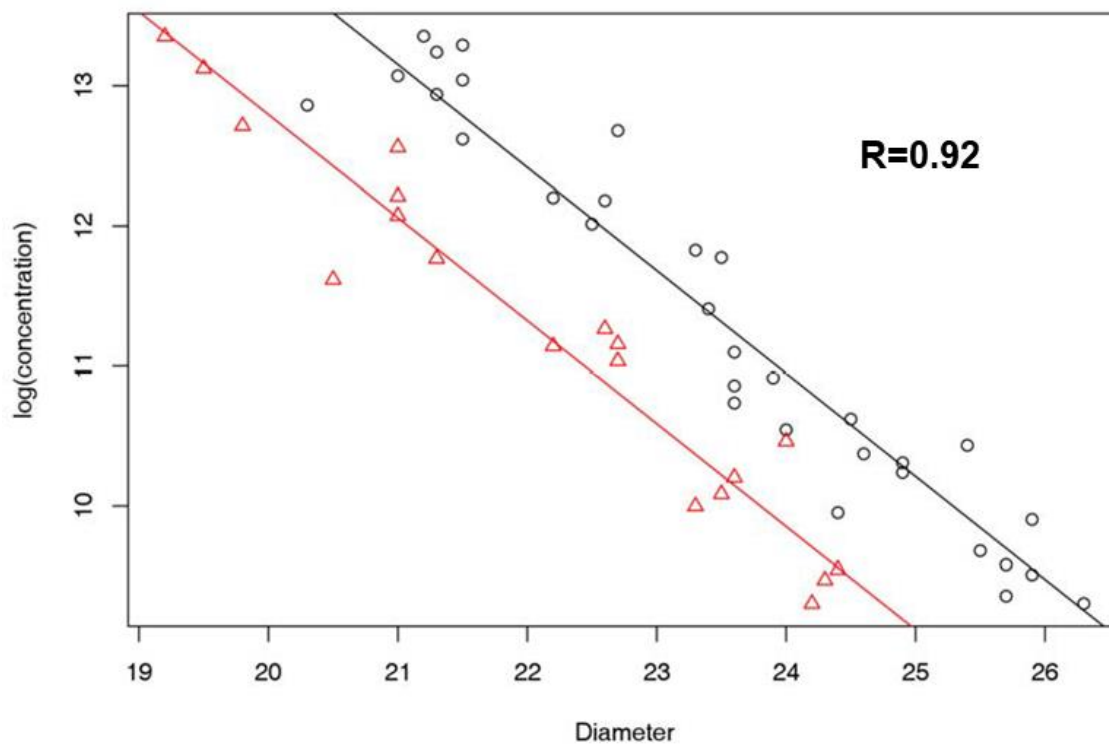
$$\text{Concentration} = b_0 + b_1 \text{ Diameter} + b_2 \text{ « Glucose »} + \text{error}$$

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

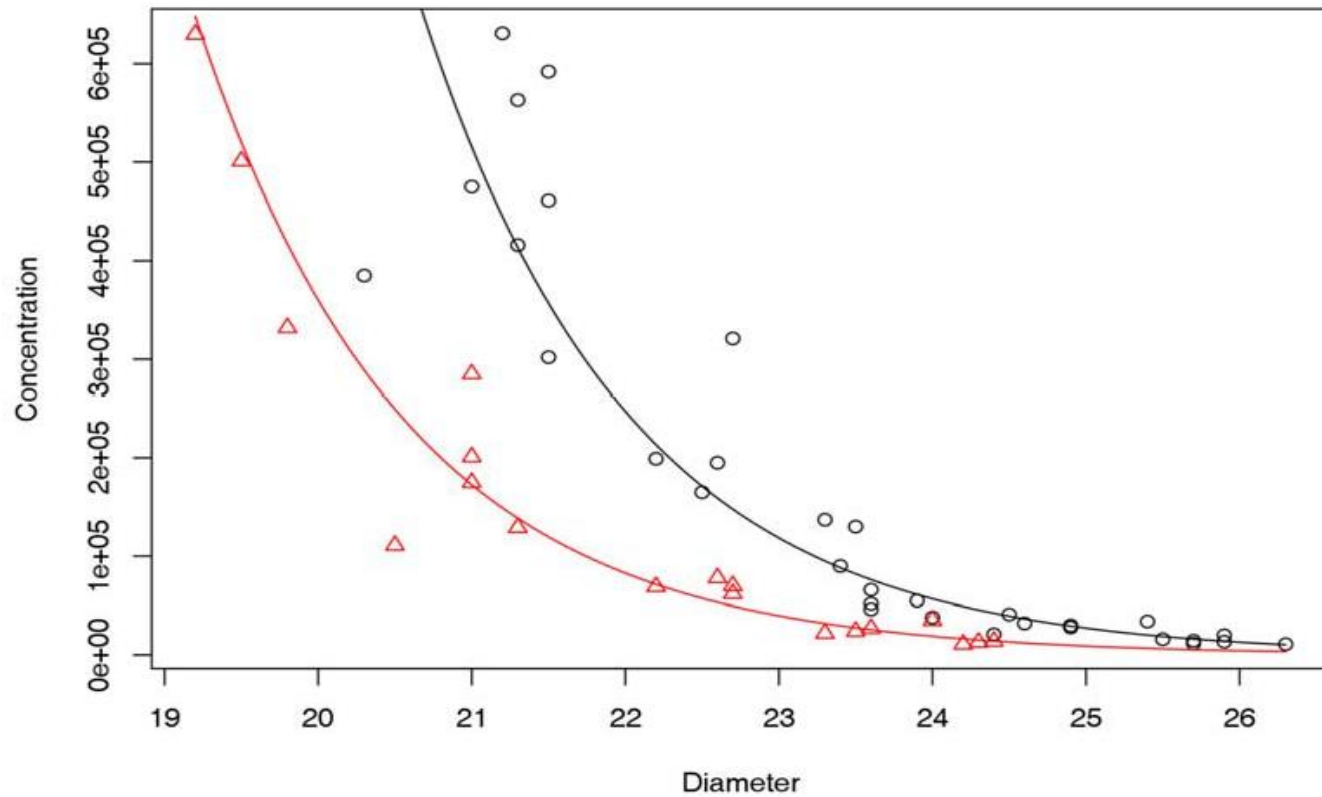
Log(concentration) according to diameter and glucose



Prediction of log Concentration according to Diameter and Glucose



Prediction of Concentration according to Diameter and Glucose



Pitfalls in regression: Extrapolation

We don't know what the relationship between X and Y looks like outside the range of the data.

Extrapolating the model outside of this range is likely to give meaningless results.

