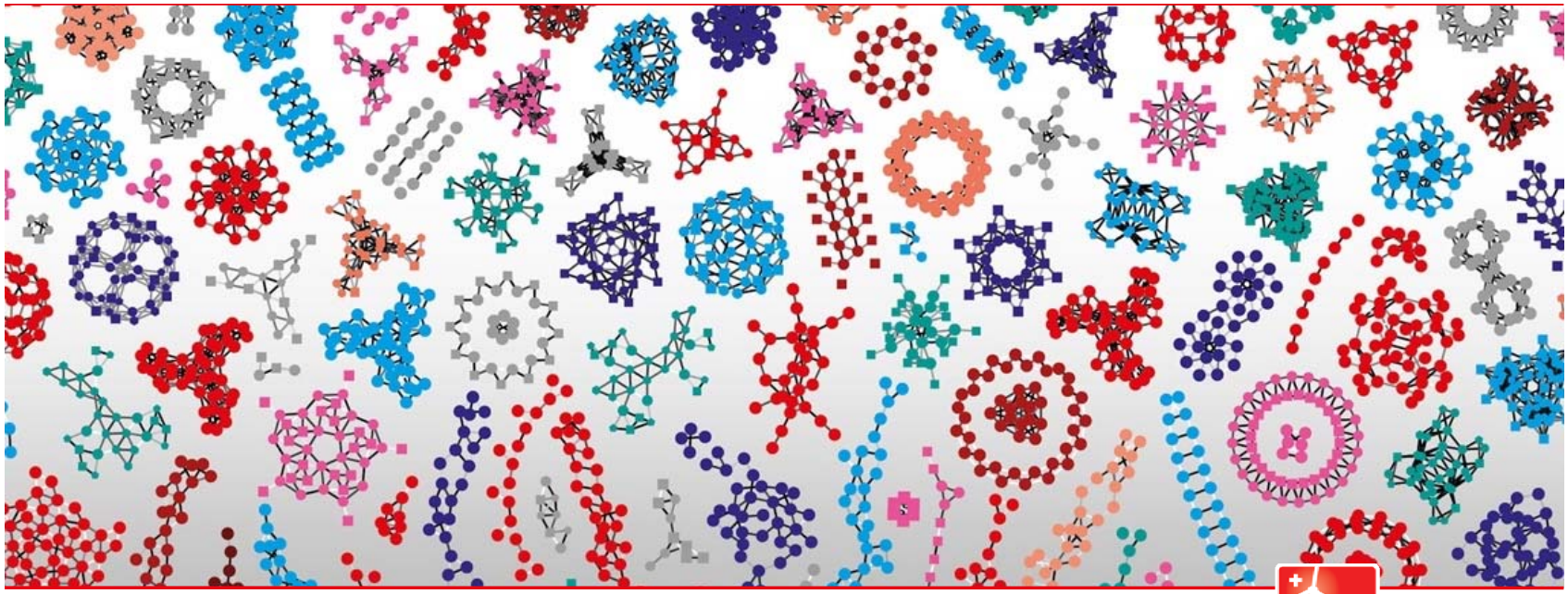


Swiss Institute of Bioinformatics

SIB Swiss Institute of Bioinformatics

Advanced statistics: Statistical modeling, 16-19 August 2021

Isabelle Dupanloup and Rachel Jeitziner

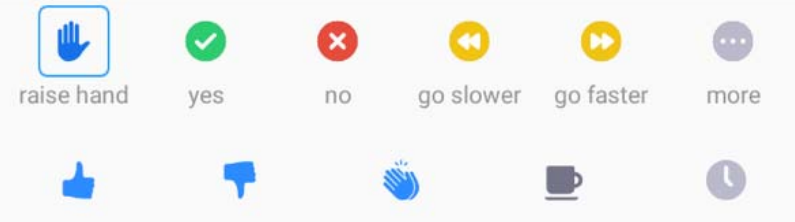
Course etiquette

Video on when possible

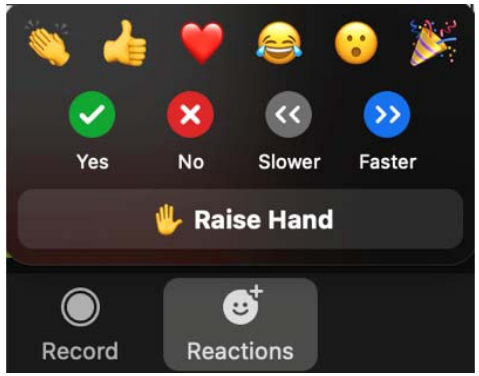
Mute when not speaking

Asking question

During lectures, you are encouraged to raise your hand if you have questions (if in-person), or use the “raise hand” Zoom functionality (if online). Find the buttons in the participants list (‘Participants’ button):



Alternatively, (depending on your zoom version or OS) use the ‘Reactions’ button:



Code of conduct

SIB abides by the ELIXIR Code of Conduct. We are all thus expected to abide by the same code. In summary:

We **value** each other's perspectives providing a safe environment for people to be themselves.

We will **maintain** high ethical standards across all ELIXIR events.

We **adopt** a zero-tolerance approach to harassment and discrimination in any form.

We will **apply** honesty and integrity in the dealing of any transgressions against the Code.

We are **committed** to making ELIXIR events a collaborative, supportive and enjoyable experience.

We will **ensure** that our environment allows everyone to feel respected and included.

https://elixir-europe.org/events/code-of-conduct

The Bioinformatics Core Facility at SIB

https://bcf.sib.swiss

- Teaching and training
- Biostatistics and bioinformatics support
- Collaboration

https://www.sib.swiss/mauro-delorenzi-frederic-schutz-group

Course material and credits

- Moodle: https://edu.sib.swiss/course/view.php?id=527
- Login: assm21
- Password: SIB_assm21

Please, give us feedback at the end of the course !

- Exam: exercises for credits (1 ECTS)
- Send answers to isabelle.dupanloup@sib.swiss

First, tell us about yourself !

- Background and research area
- What you expect from this course, experience with R

Photo by National Cancer Institute, Unsplash

Photo by Scott Graham, Unsplash

Advanced statistics: Statistical modeling

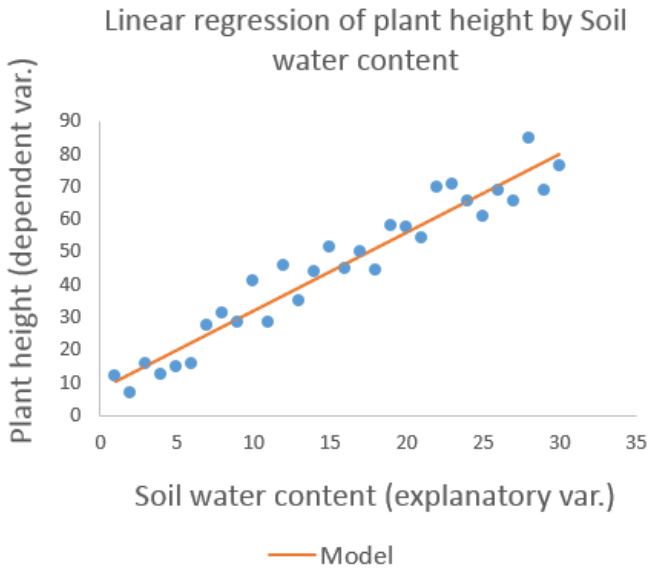
- Introductory statistics course: models and tools (such as linear regression) to analyze “simple” datasets (not appropriate for all types of data)
- Goal of the course: go beyond classical linear modelling
- Program of the course:
  - brief review of the basics of linear regression
  - explore extensions of linear models, such as polynomial regression, splines, local regression, and generalized additive models
  - logistic regression
  - mixed-effects linear models
  - application of mixed-effects linear models in analyzing longitudinal data

Statistical models

What is a statistical model ?

Statistical modeling: simplified, mathematically-formalized way to approximate reality (i.e. what generates your data) and optionally to make predictions from this approximation

Statistical model: mathematical equation that is used



What is a statistical model ?

A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data.
A statistical model represents, often in considerably idealized form, the data-generating process.

A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables.
A statistical model is "a formal representation of a theory".

https://en.wikipedia.org/wiki/Statistical_model

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***

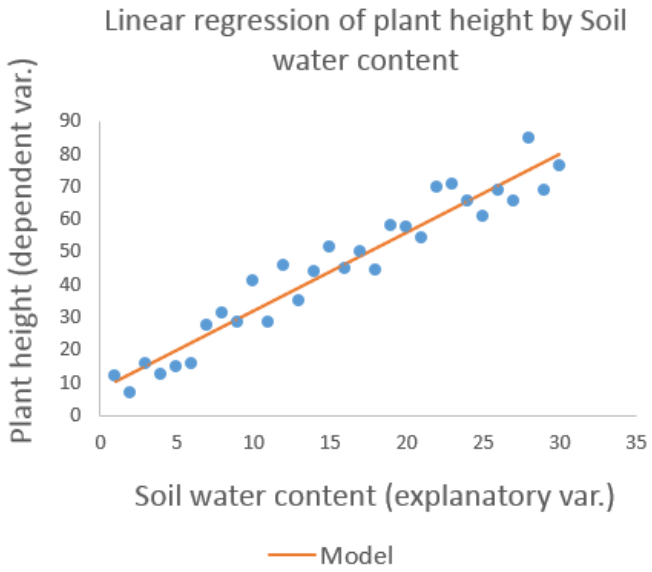
(from Terry Speed)

What are dependent and explanatory variables ?

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

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

Both variables may be quantitative or qualitative



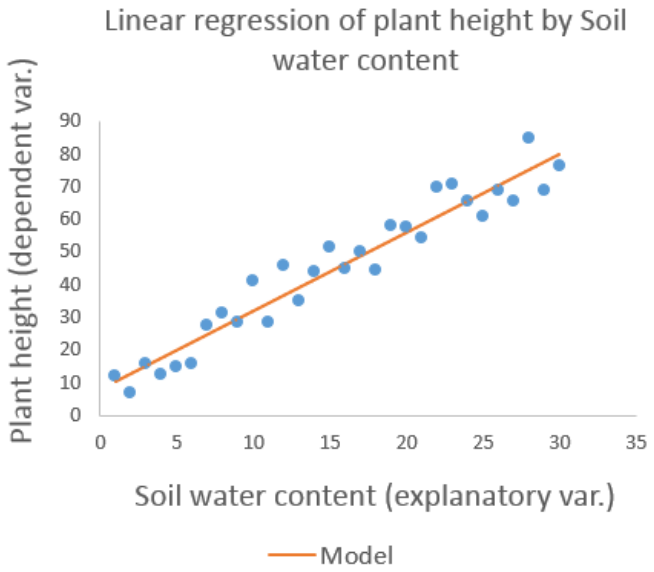
What is a model parameter ?

Statistical model: equation with quantities called model parameters

Statistical modeling

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

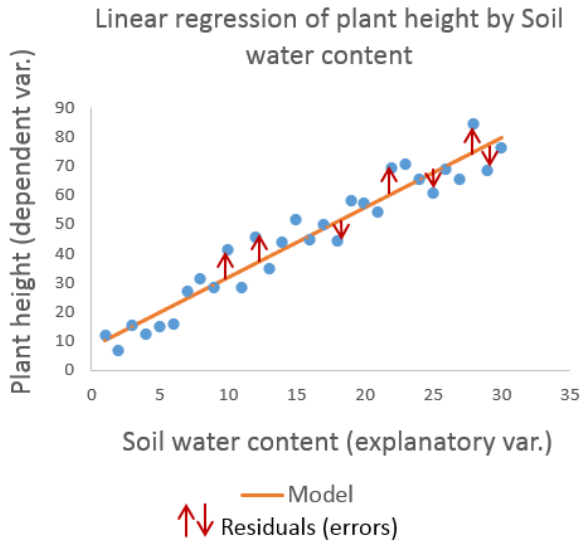
Height = **intercept** + **slope***soil water content

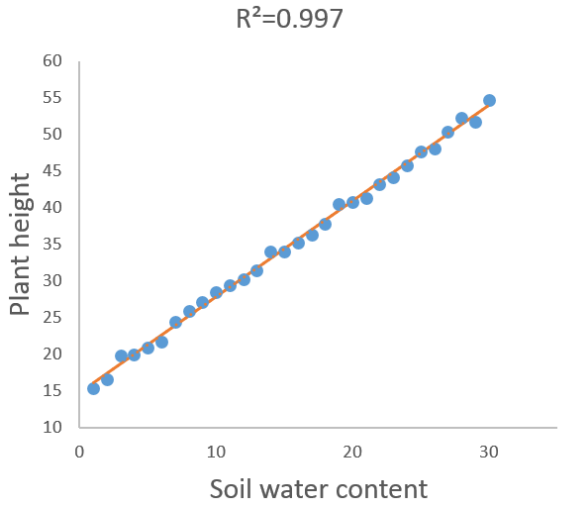


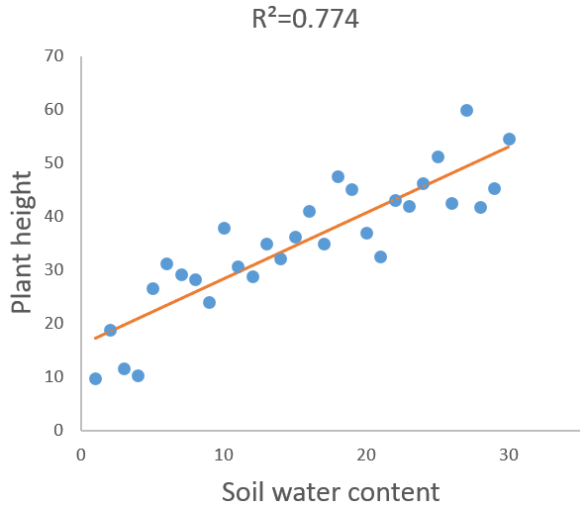
What is a model residual ?

Model residuals (or errors): distances between data points and the model equation with quantities called model parameters

Model residuals represent 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 models are of the form

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

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

A word of caution !

Modelling is not about just finding the right type of equation to describe the data, and finding the right algorithm to estimate the parameters of this equation !

We should not consider that the modeling problem consists only of simple pairs of data points (e.g. response and explanatory variables).

Other information of interest include for example how the data was collected, how it is structured, what we expect from the model (description ? prediction ?), and what other variables were not observed/measured.

We will not discuss this in detail, but we will touch on it briefly in some places.

Essentially, all models are wrong, but some are useful.

A word of caution !

The choice of a statistical model is not straightforward. It is erroneous to think that every data set has its own adapted model.

Every modelling tool answers specific questions.

The choice of a statistical model can also be guided by the shape of the relationships between the dependent and explanatory variables.

Once you choose the appropriate modelling tool, you should consider how many parameters you should include in the model. The more parameters, the better the fit of the model to the data. But overfitting to the data you sampled is a risk !

A NON-EXHAUSTIVE grid of statistical models

Dependent variable	Explanatory variable	Parametric model
1 quantitative variable	1 qualitative variable (2 levels)	T test
	1 qualitative variable (k levels)	One-way ANOVA
	Several qualitative variable with several levels	Multi-way ANOVA
	1 quantitative variable	Simple linear (or non-linear) regression
	Several quantitative variables	Multiple linear (or non-linear) regression
	Mixture of qualitative and quantitative variables	ANCOVA
Several quantitative variables	Qualitative and/or quantitative variables	MANOVA
1 qualitative variable	Qualitative and/or quantitative variables	Logistic regression
1 count variable	Qualitative and/or quantitative variables	Poisson regression

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

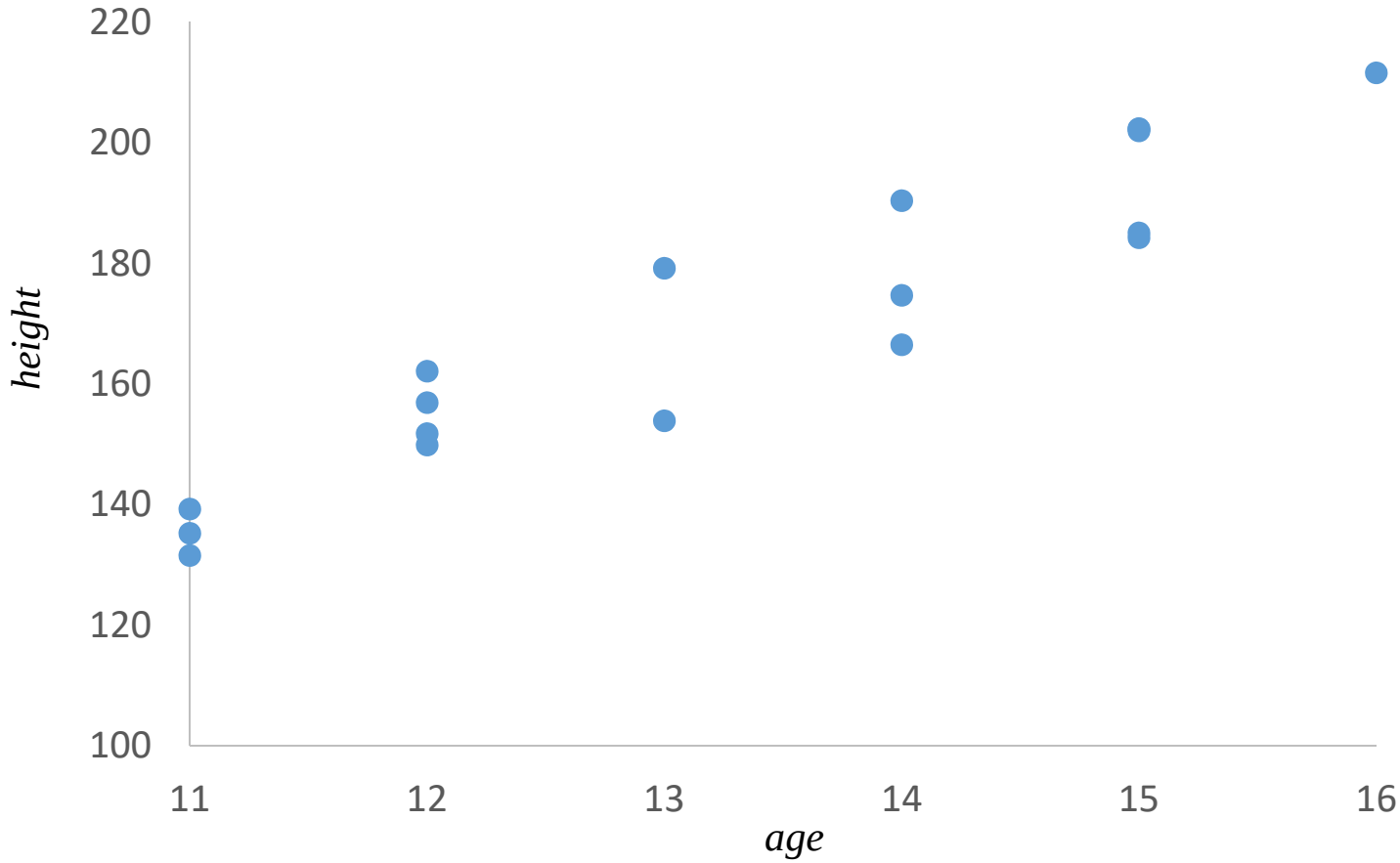
Other symbols to create formulas with combinations of variables (e.g. interactions)

- + 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)
- ^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

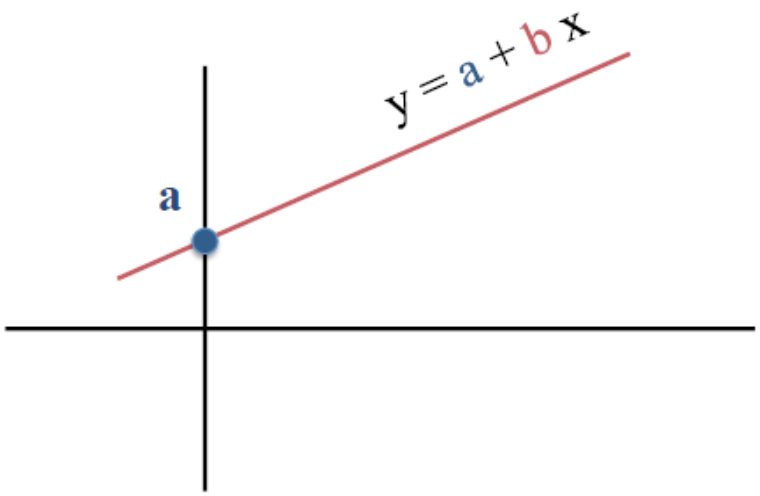
Simple linear regression refers to drawing a (particular, special) line through a scatterplot

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

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

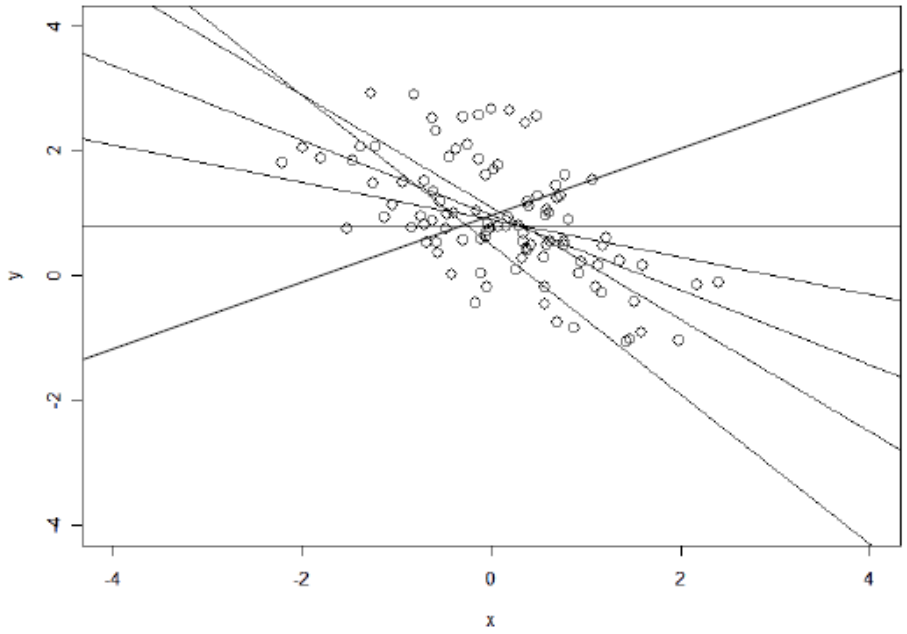
where a is called the intercept and b is the slope.



(Simple) Linear Regression

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

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



y = 0.9 + 0.6x

y = 0.8 + 0x

y = 0.9 – 0.3x

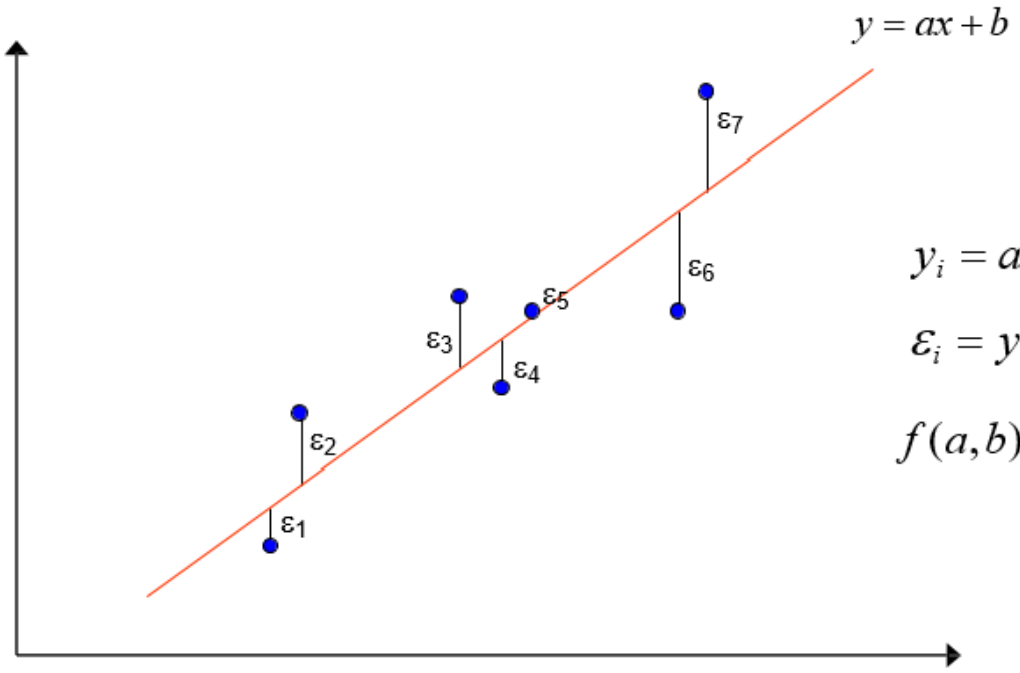
y = 1 – 0.6x

y = 1.1 – 0.9x

y = 0.5 – 1.2x

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

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)

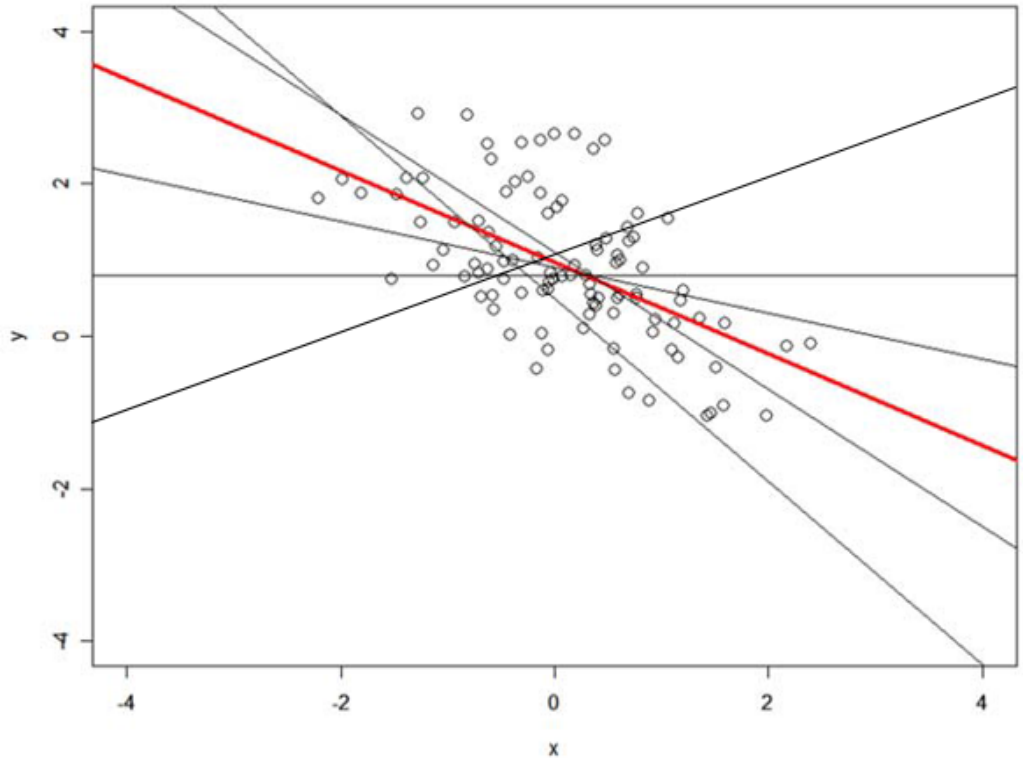
ε_i : error term $\sim N(0, \sigma^2)$ (estimated)

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

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)

(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

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

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: linearity

Linearity is about the model parameters

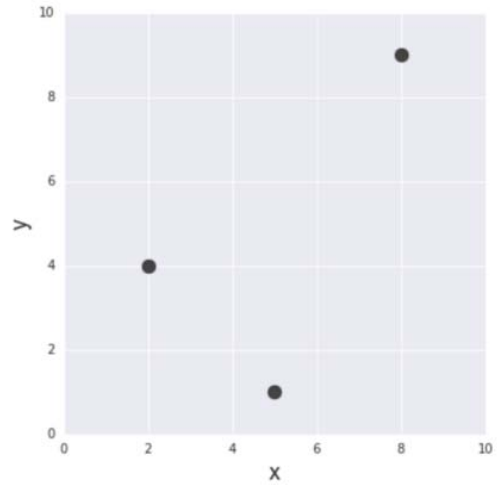
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$
$$Y_i = \beta \sin X_i + \varepsilon_i$$

} 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

Linear models: linearity

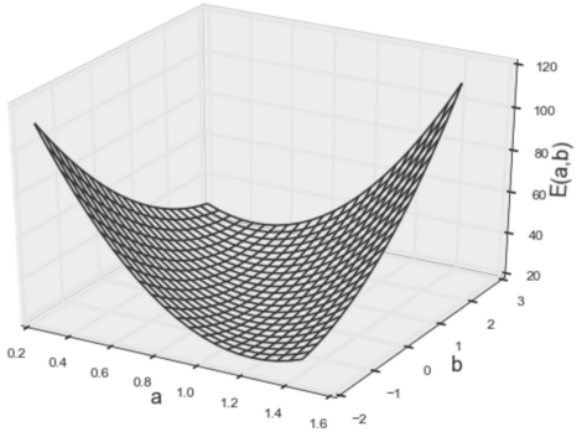


$$E(a,b) = \sum_{i=1}^{3} (f(x_i) - y_i)^2$$

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

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

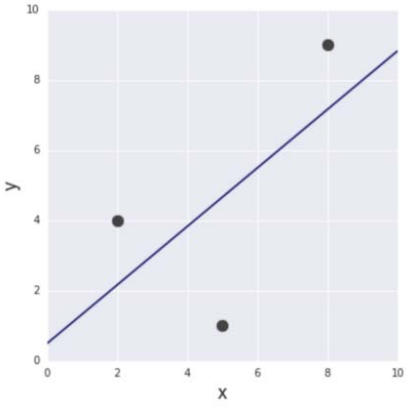
$$= 93a^2 + 3b^2 + 30ab - 170a - 28b + 98$$



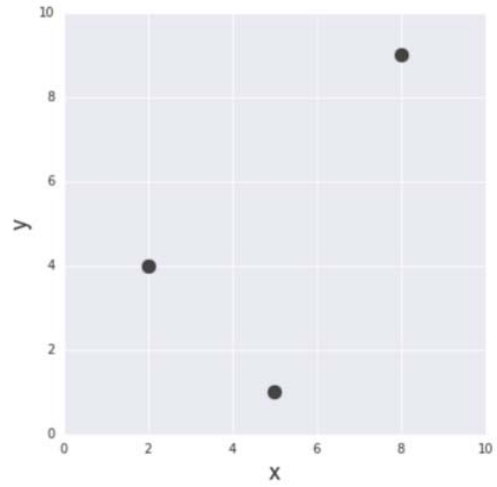
$$\frac{\partial}{\partial a} E(a,b) = 186a + 30b - 170 = 0$$

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

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



Linear models: linearity

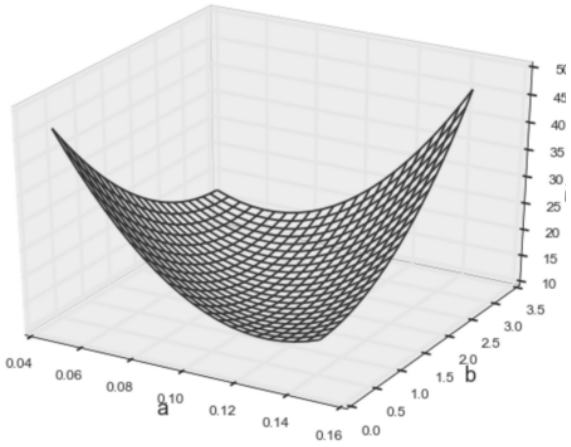


$$E(a,b) = \sum_{i=1}^{3} (f(x_i) - y_i)^2$$

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

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

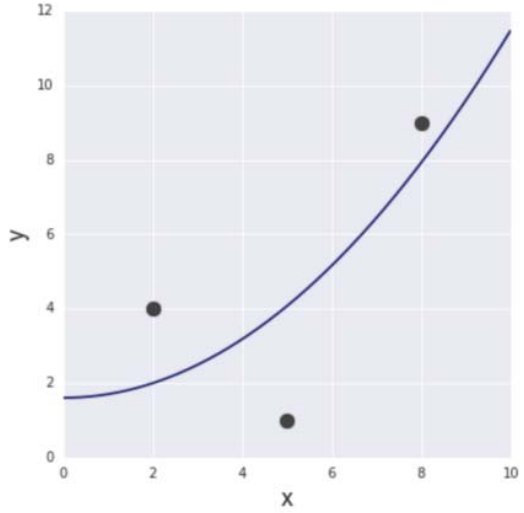
$$= 4737a^2 + 3b^2 + 186ab - 1234a - 28b + 98$$



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

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

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



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 ?"

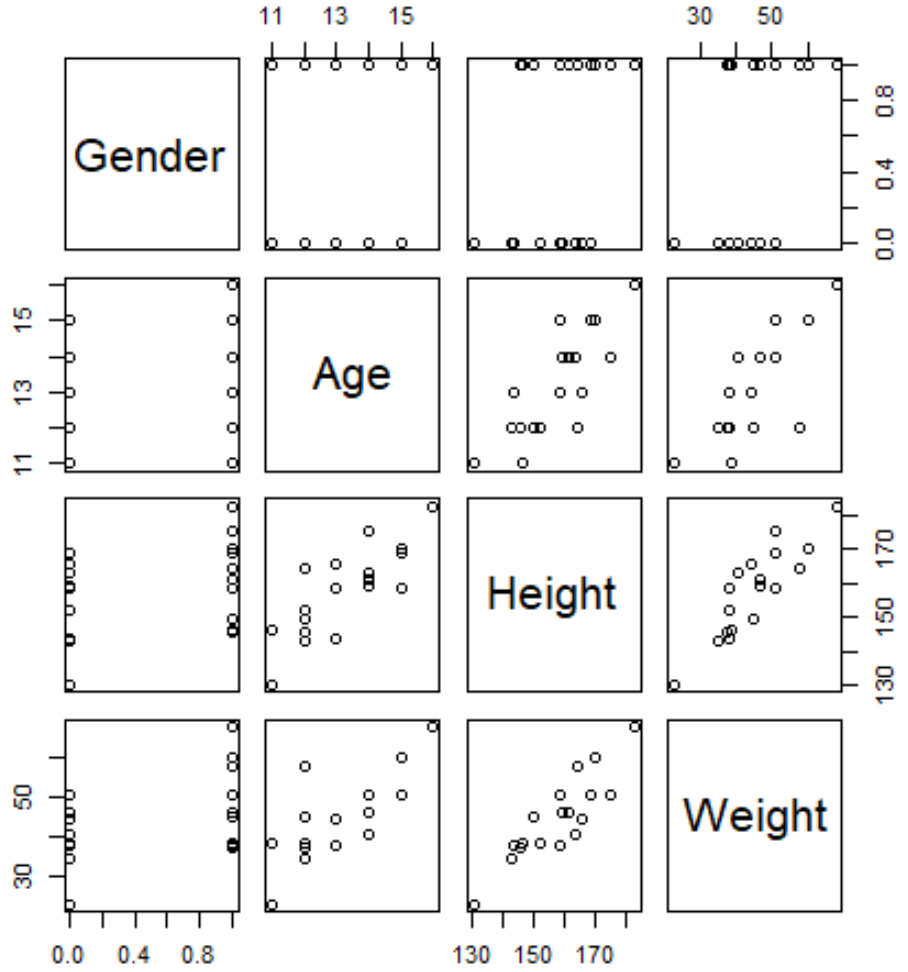
The CLASS dataset from SAS

```
> 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



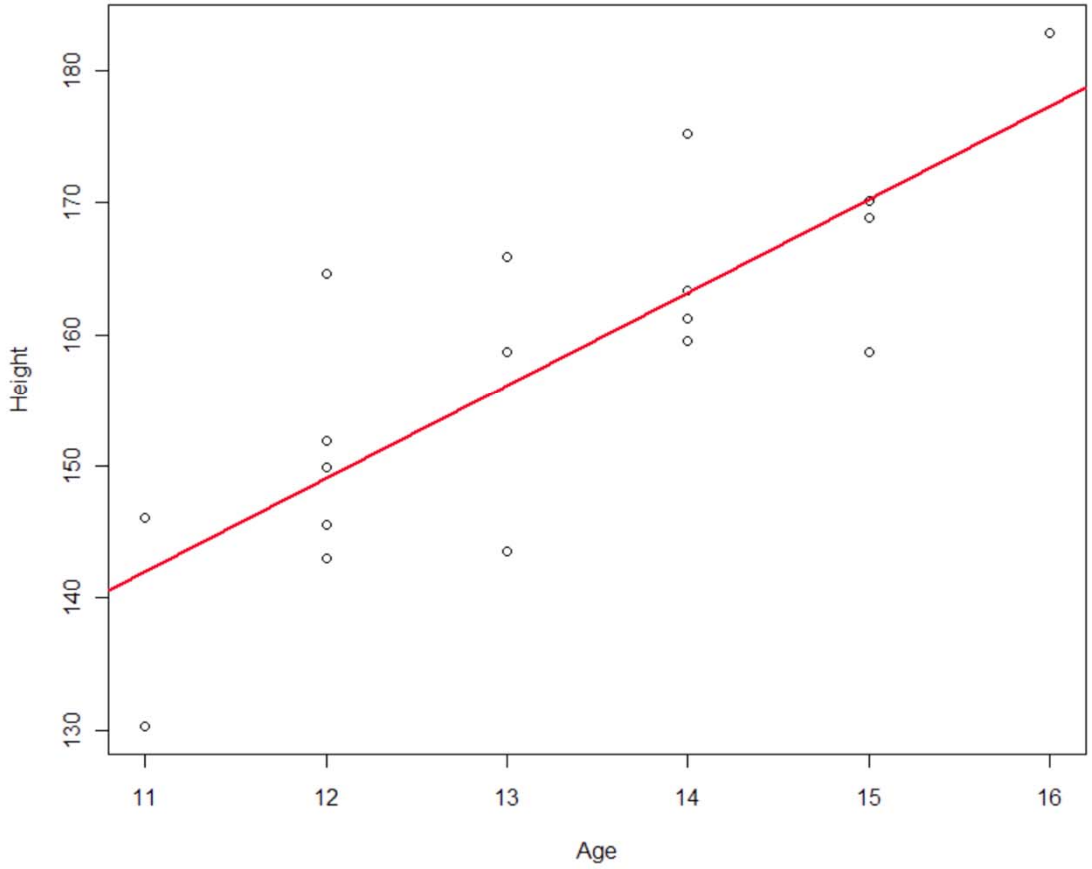
The CLASS dataset from SAS

```
> 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

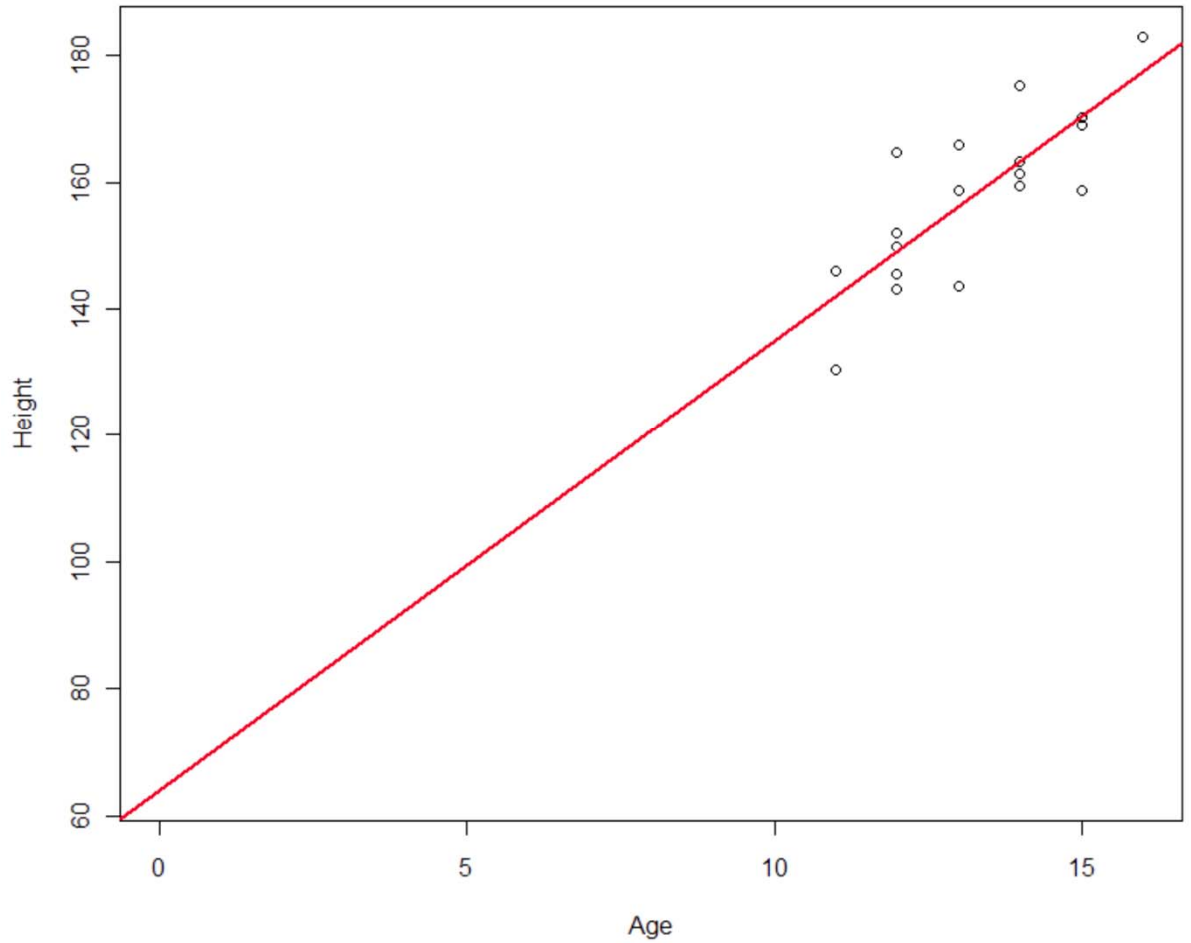
The CLASS dataset from SAS

```
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
```

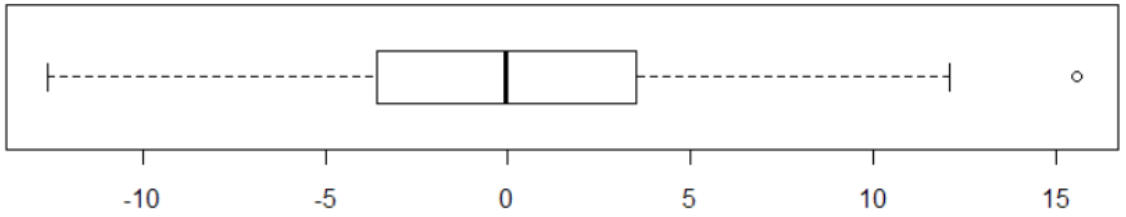
The CLASS dataset from SAS

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

This 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).

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

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

```
> 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).

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:

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

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	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$MST = \frac{SST}{n-1}$	

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
```

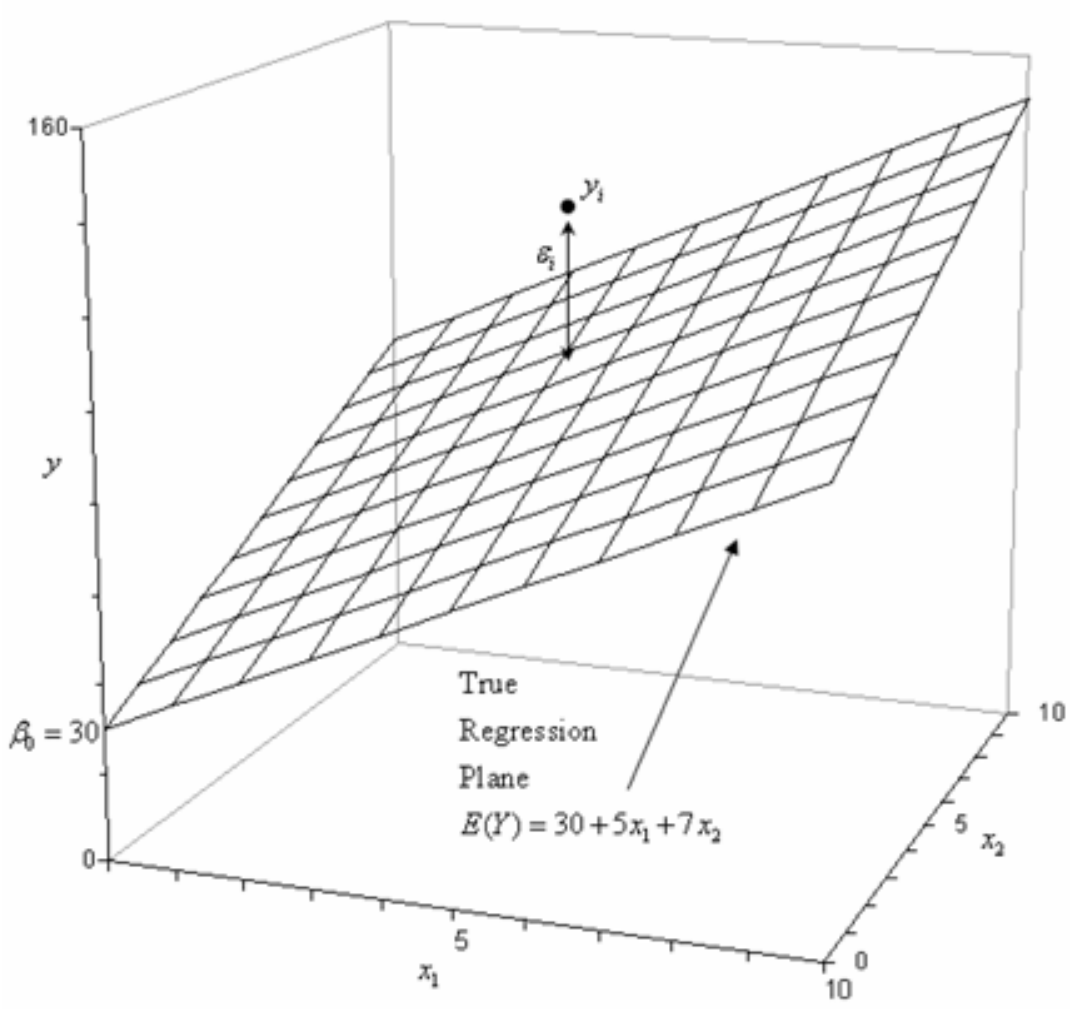
The F-statistic allows us to test if the whole regression (adding all variables *vs* having only the intercept in) is significant.

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.

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 below 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

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
```

Again, 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.

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

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

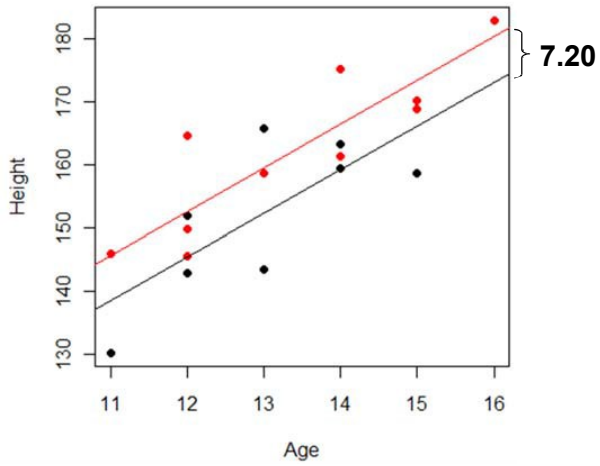
The factor GenderM corresponds to the difference in baseline for Males compared to females

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.

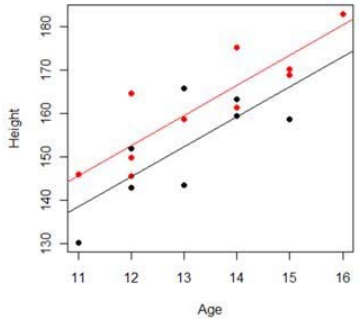
If this assumption is incorrect, it means that there is an ***interaction*** between the factors « age » and « gender », that is, the effect of age is different depending on the gender.

Interactions are modeled in R in the following way:

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

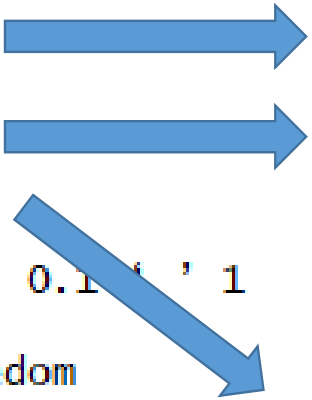
which is equivalent to

```
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
```



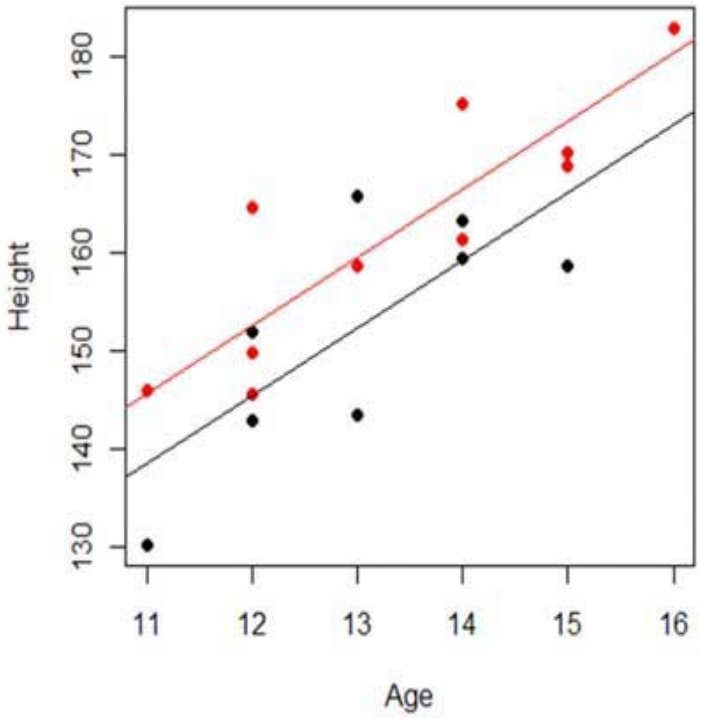
baseline for height among Female

difference in baseline for Males compared to females

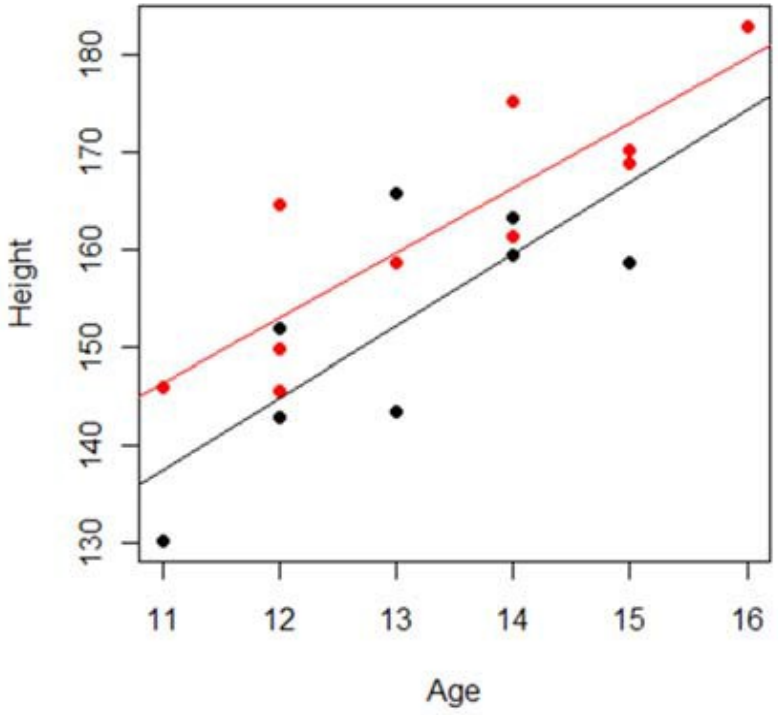
age effect only for males

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

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

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

Continuous versus dummy predictors

X is the **design matrix**; a column of Xij can be used to encode

Continuous quantities
- Drug dose
- Temperature
- Time

Discrete conditions (dummy predictor)
- KO (vs wt)
- Gender
- Treatment vs non-treatment

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.

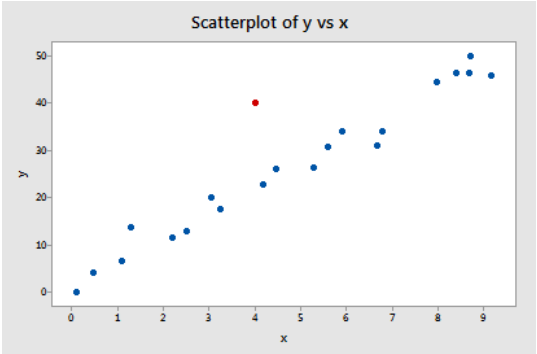
Diagnostic tools

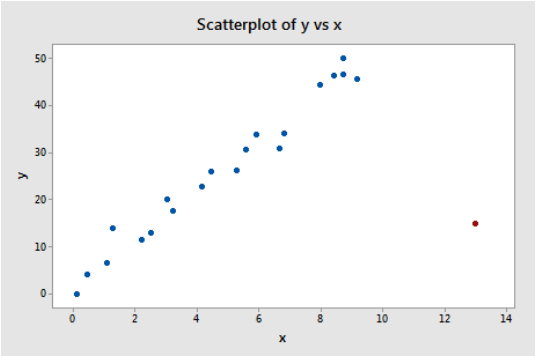
Basic model checking

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 !

Examination of *residuals*: (which should show no obvious trend, since any systematic effect in the residuals should ideally be captured by the model):

- ❑ Normality
- ❑ Non-constant variance
- ❑ Independence
- ❑ Curvature
- ❑ Outliers

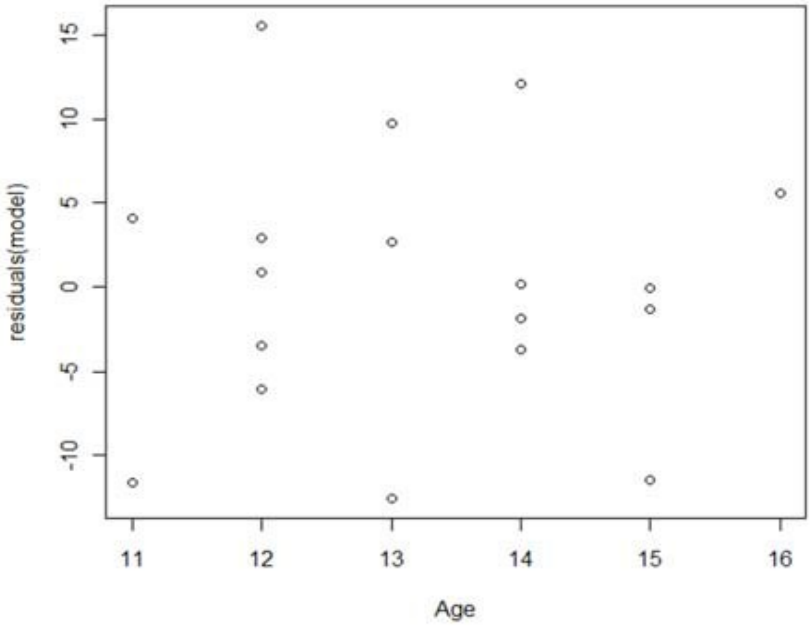


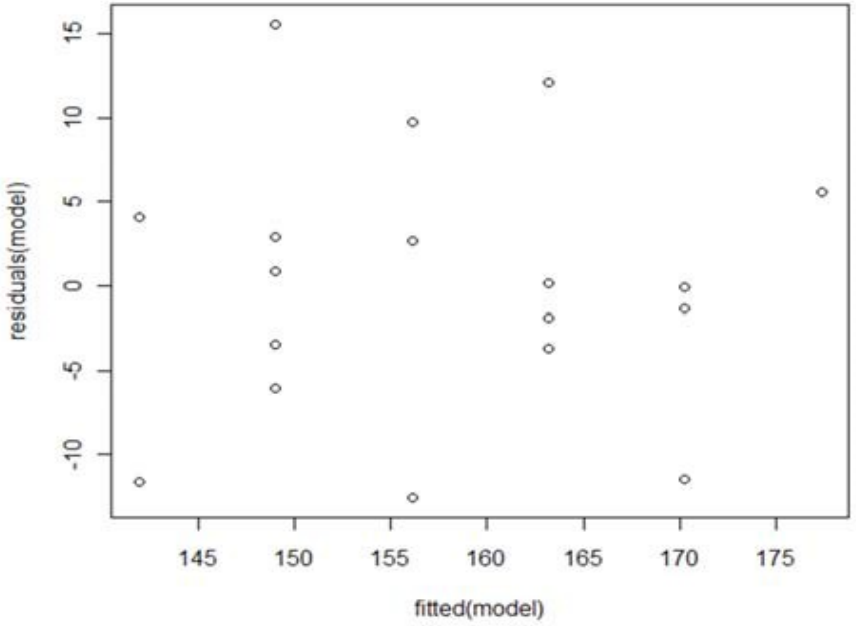


Detection of *influential observations*

- ❑ *Hat matrix*

Residuals






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

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

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

Works also for multiple regression

Hat values

High leverage ('influential') points are far from the center, and have potentially greater influence

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

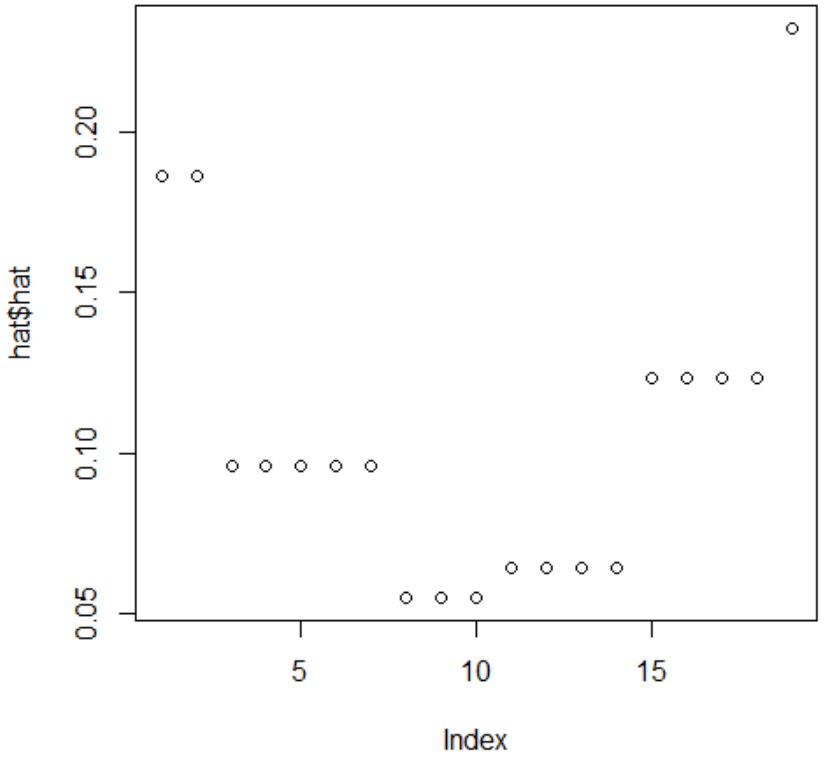
h_{ij}: contribution of the ith observation to the jth fitted value

h_i: contribution of the ith observation to the fitted values

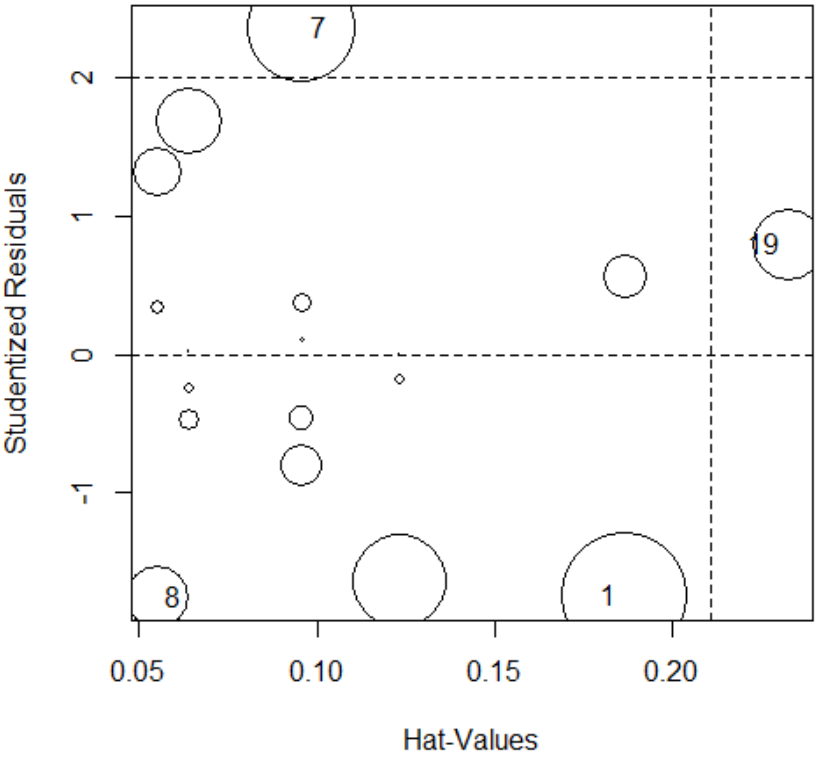
Average value of h = number of explanatory variables p/n

Cutoff typically 2p/n

Hat values

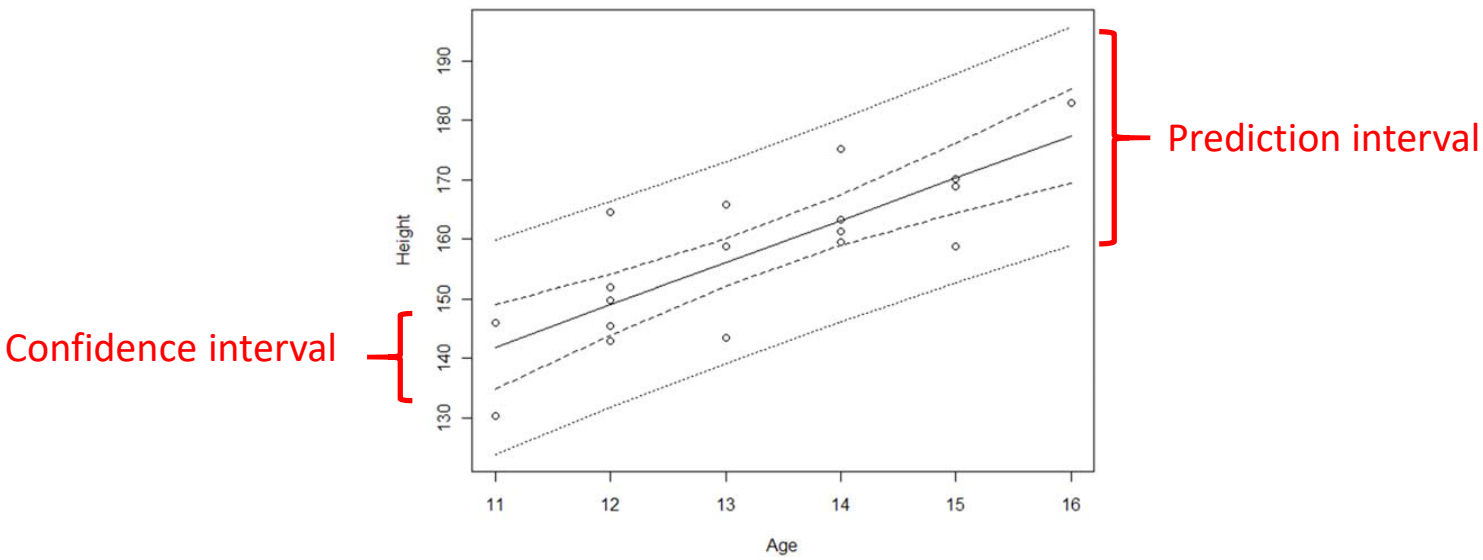


id	Name	Gender	Age	Height	Weight
1	JOYCE	F	11	130.302	22.877
2	THOMAS	M	11	146.050	38.505
3	JAMES	M	12	145.542	37.599
4	JANE	F	12	151.892	38.279
5	JOHN	M	12	149.860	45.074
6	LOUISE	F	12	143.002	34.881
7	ROBERT	M	12	164.592	57.984
8	ALICE	F	13	143.510	38.052
9	BARBARA	F	13	165.862	44.394
10	JEFFREY	M	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

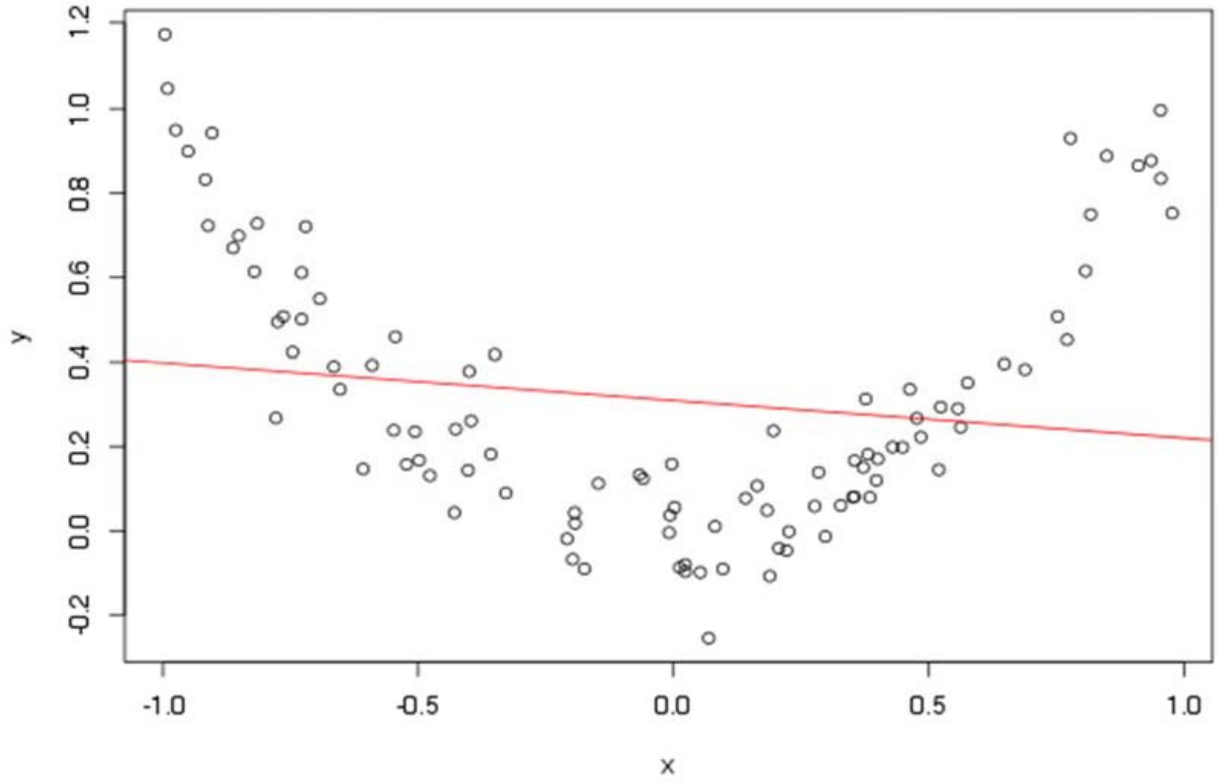


Narrow bands: describe the uncertainty about the regression line

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 ?

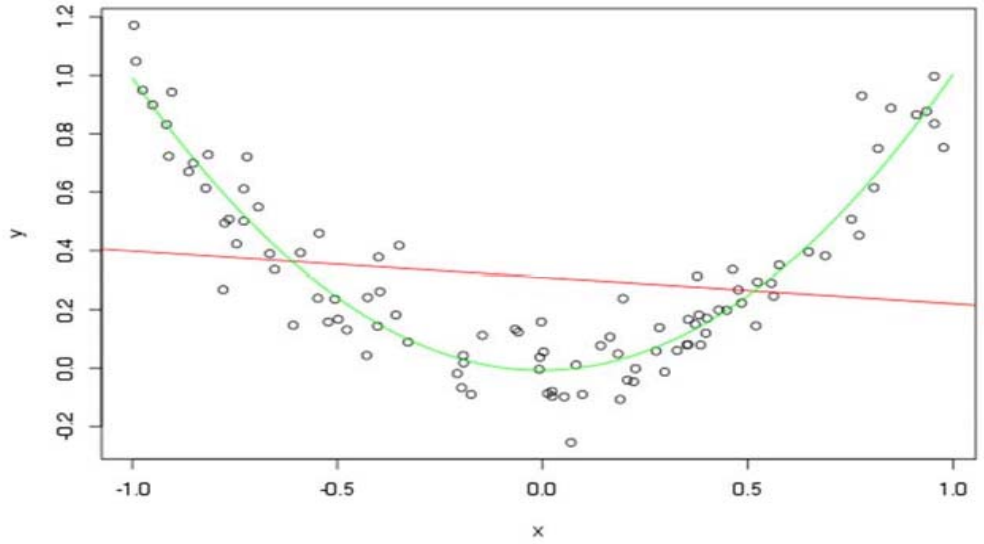


What if the data is not linear ?

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.

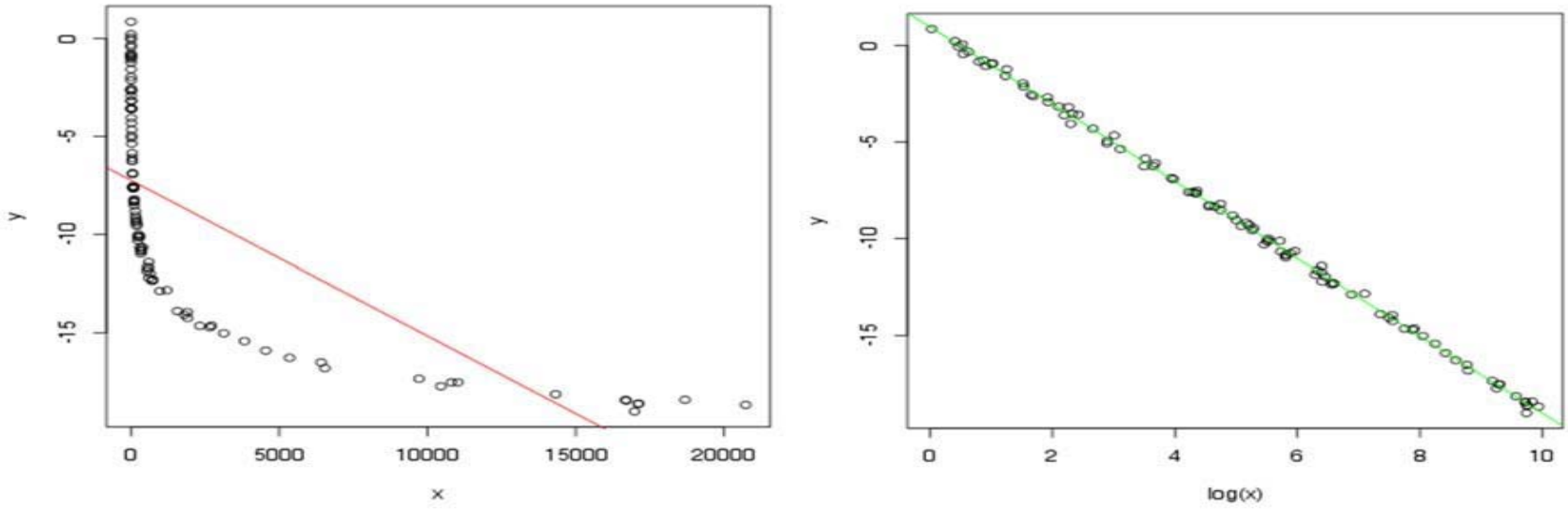


What if the data is not linear ?

Consider transforming the data (log)

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

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



Example: predicting cell concentration

The hellung dataset

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

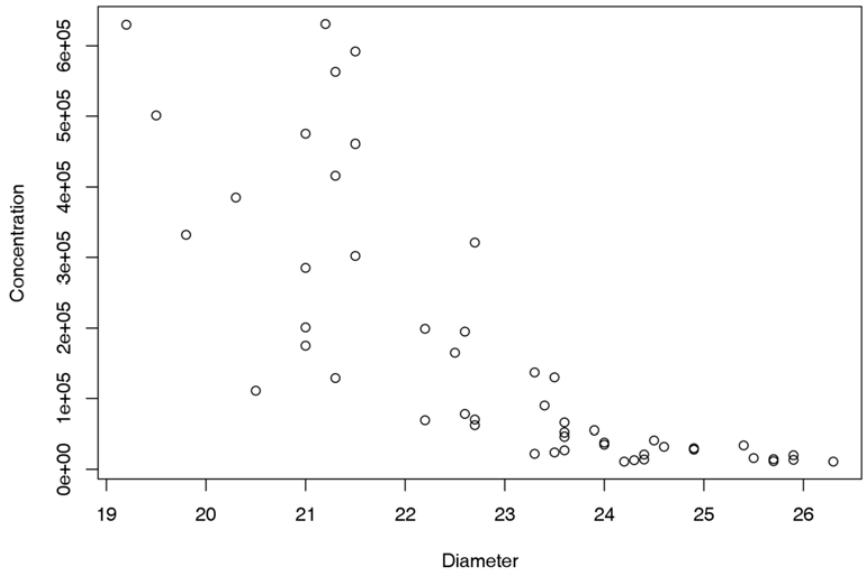
```
> library(ISwR); data(hellung)
```

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

The Hellung data in R

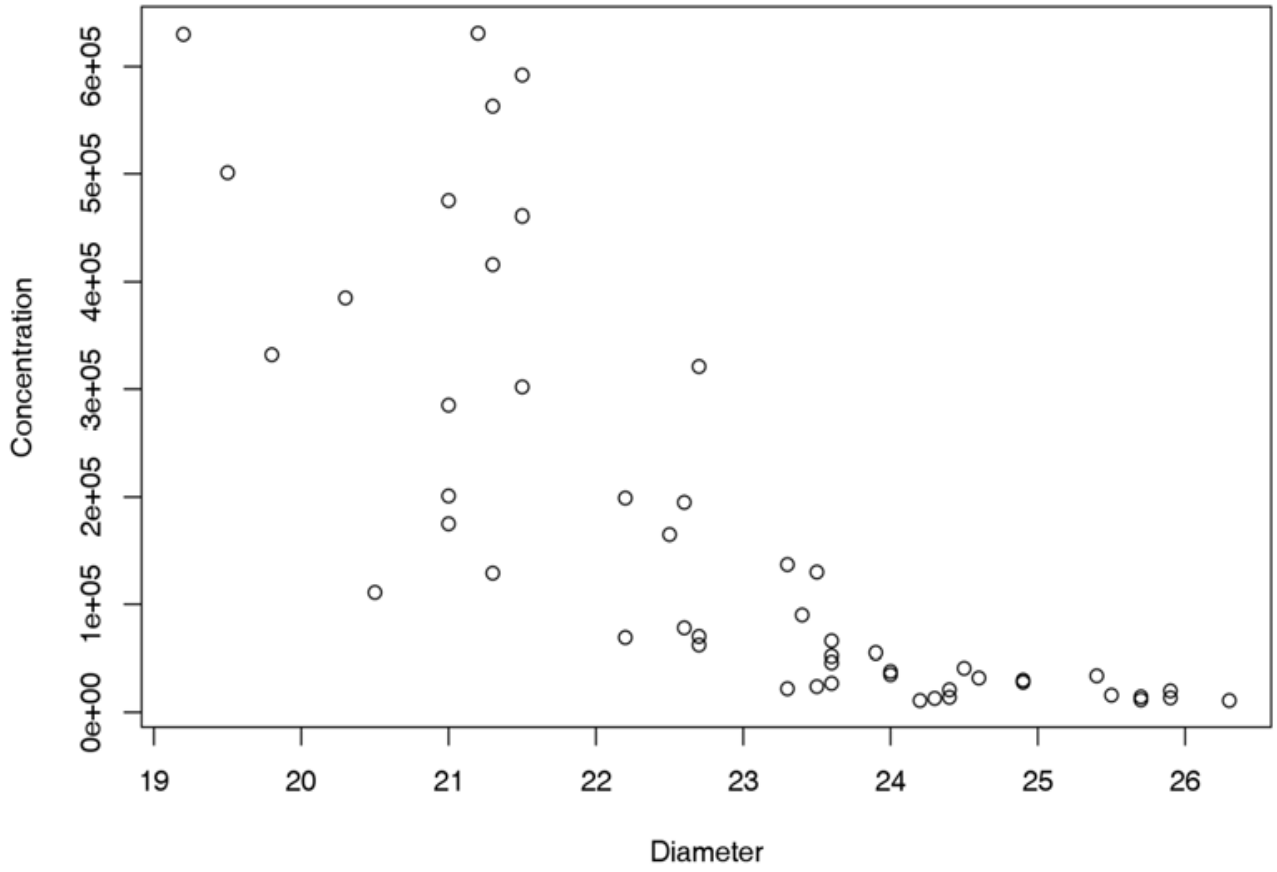
```
> hellung
   glucose   conc diameter
1        1 631000     21.2
2        1 592000     21.5
3        1 563000     21.3
4        1 475000     21.0
5        1 461000     21.5
[...]
33       2 630000     19.2
34       2 501000     19.5
35       2 332000     19.8
36       2 285000     21.0
37       2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

Hellung dataset: Diameter vs Concentration

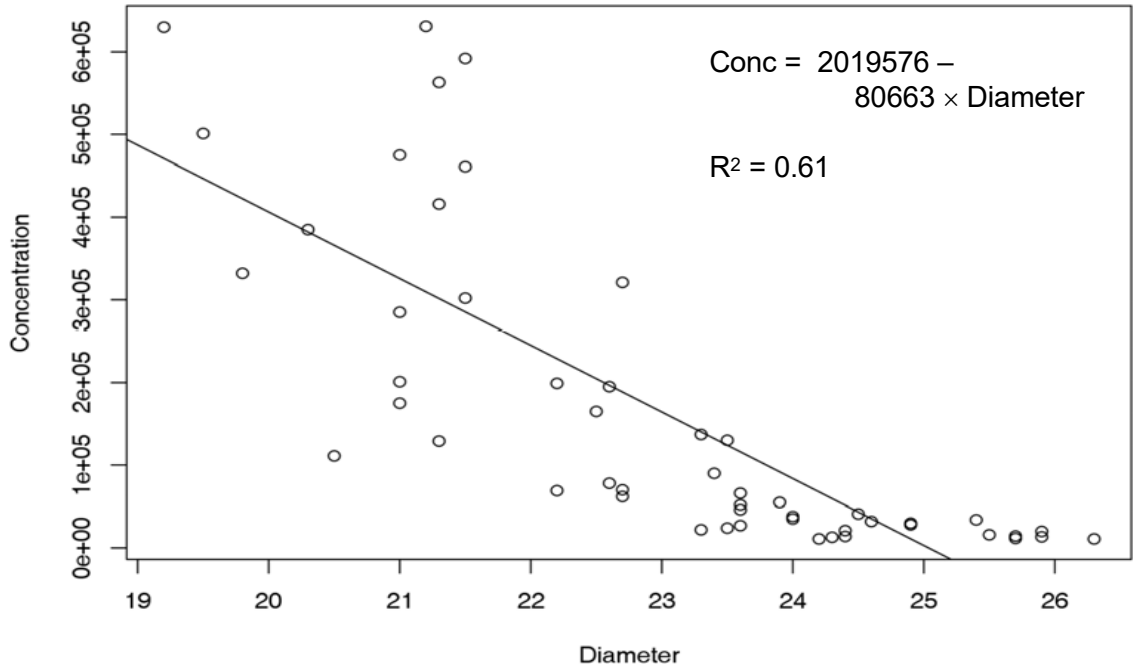



```
> plot(hellung$diameter, hellung$conc,
     xlab="Diameter", ylab="Concentration")
```

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



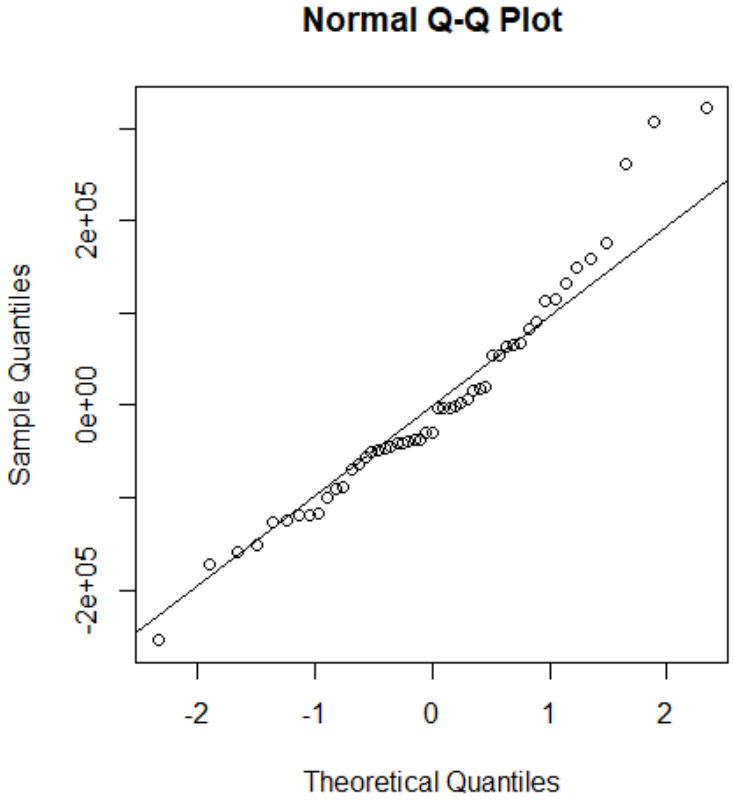
Linear model predicting Concentration from Diameter




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

Do the residuals follow a normal distribution ?

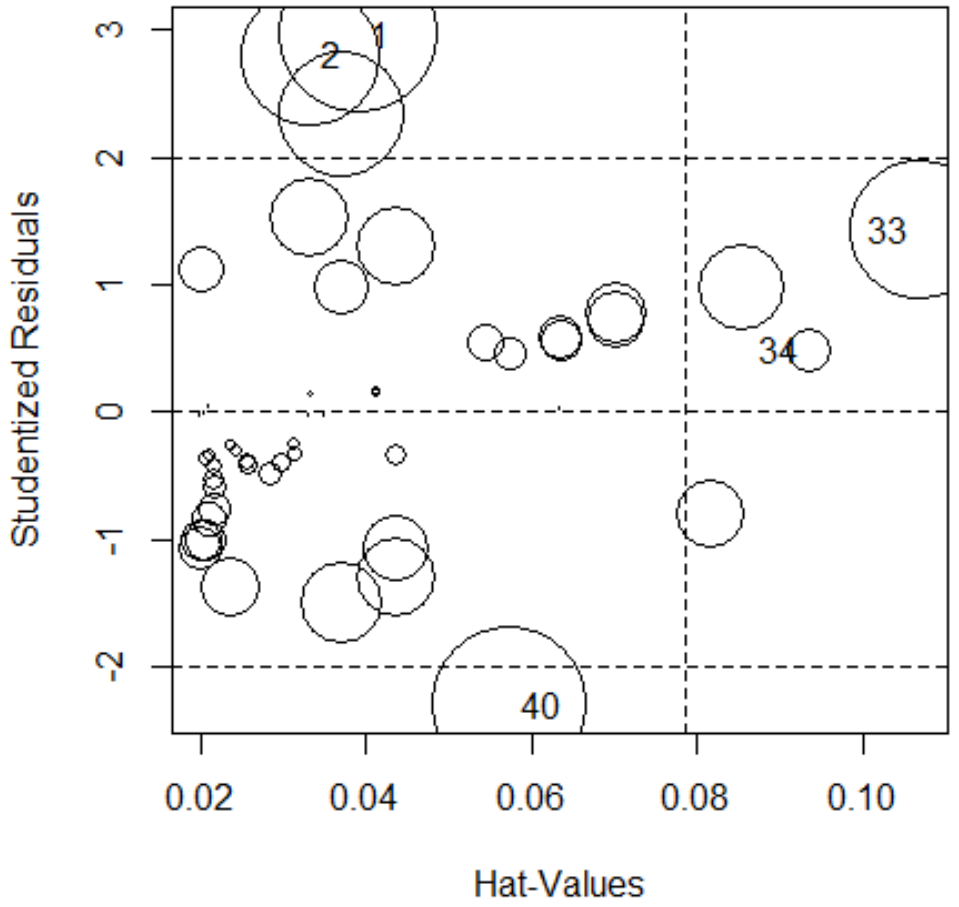
```
> qqnorm(residuals(model))
> qqline(residuals(model))
```



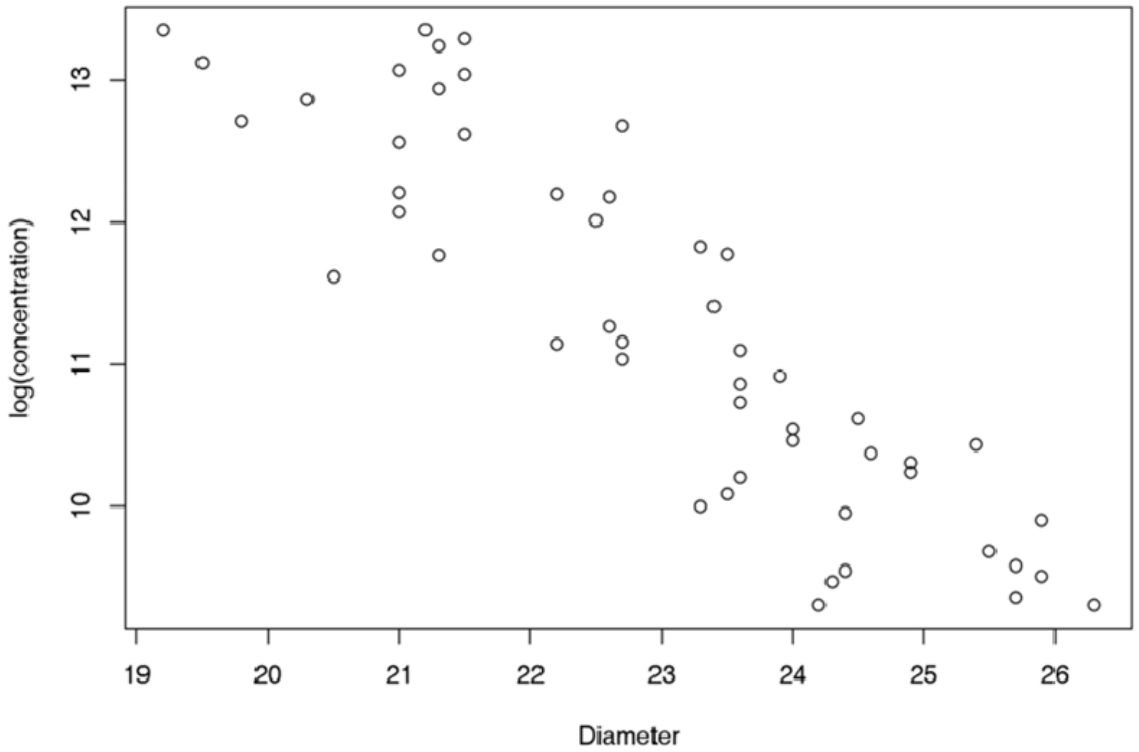

```
> ks.test(residuals(model), "pnorm")

        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
```

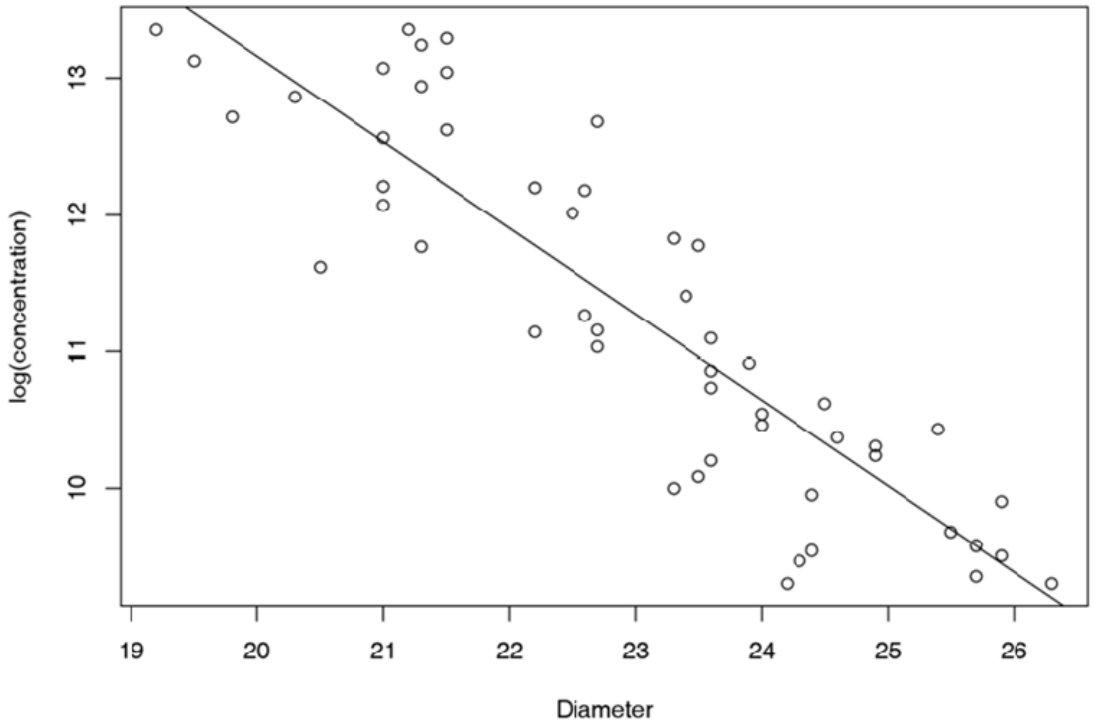


Transforming the data to improve the fit




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

Linear model predicting log(Concentration) from Diameter



log(conc) = 25.7 – 0.62 × Diameter

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

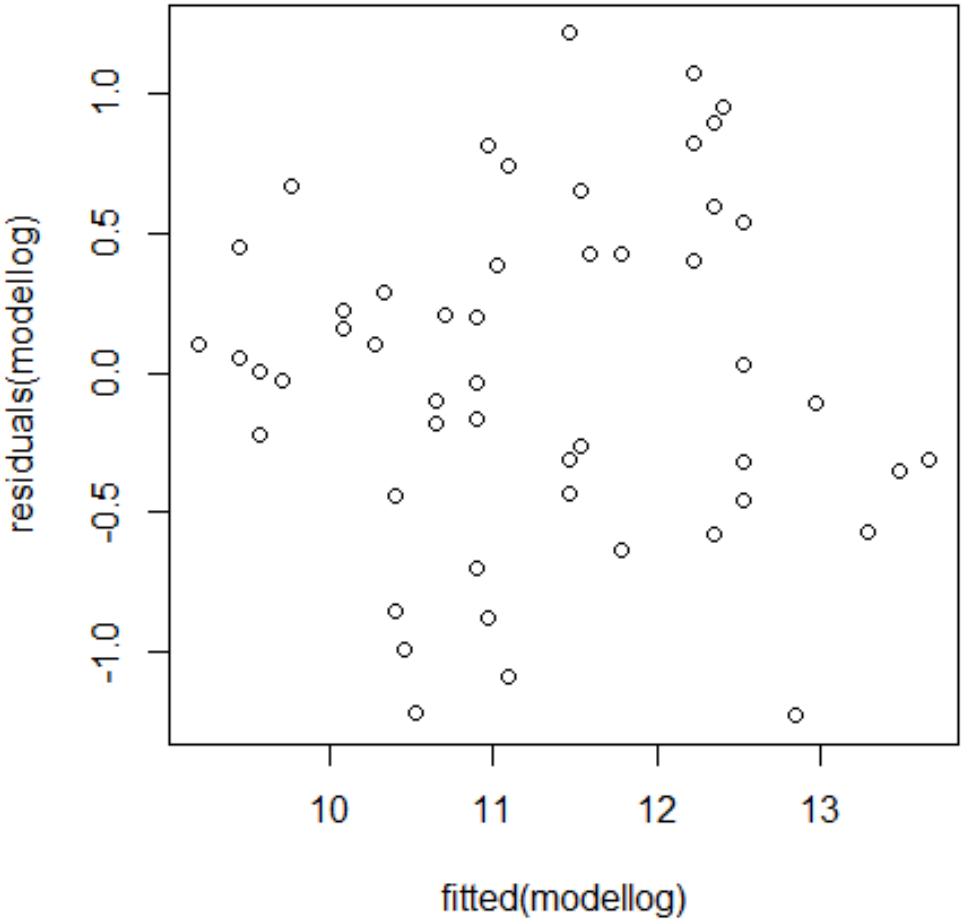
$R^2 = 0.78$

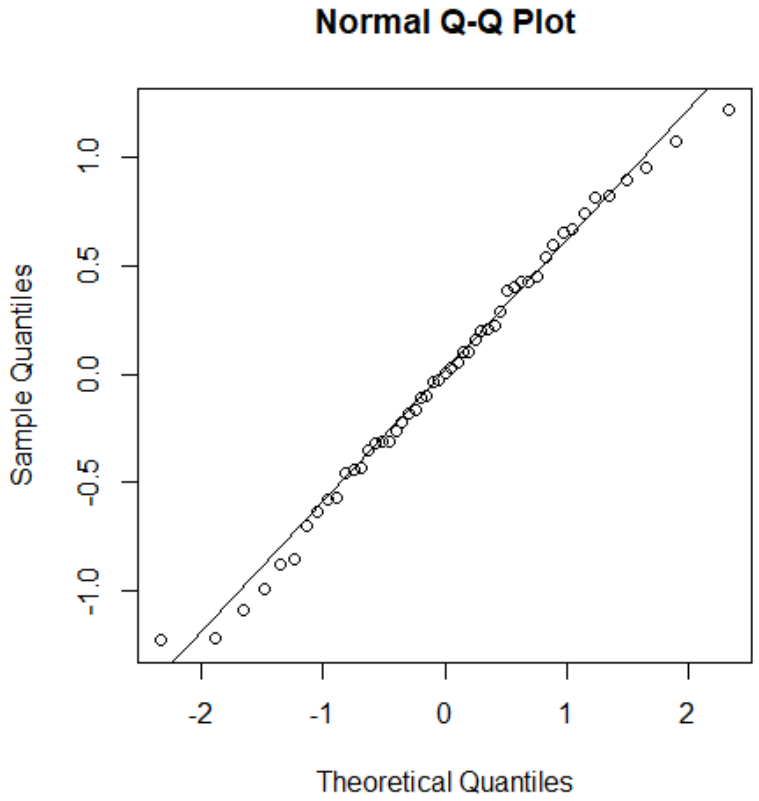
Details of the linear model

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

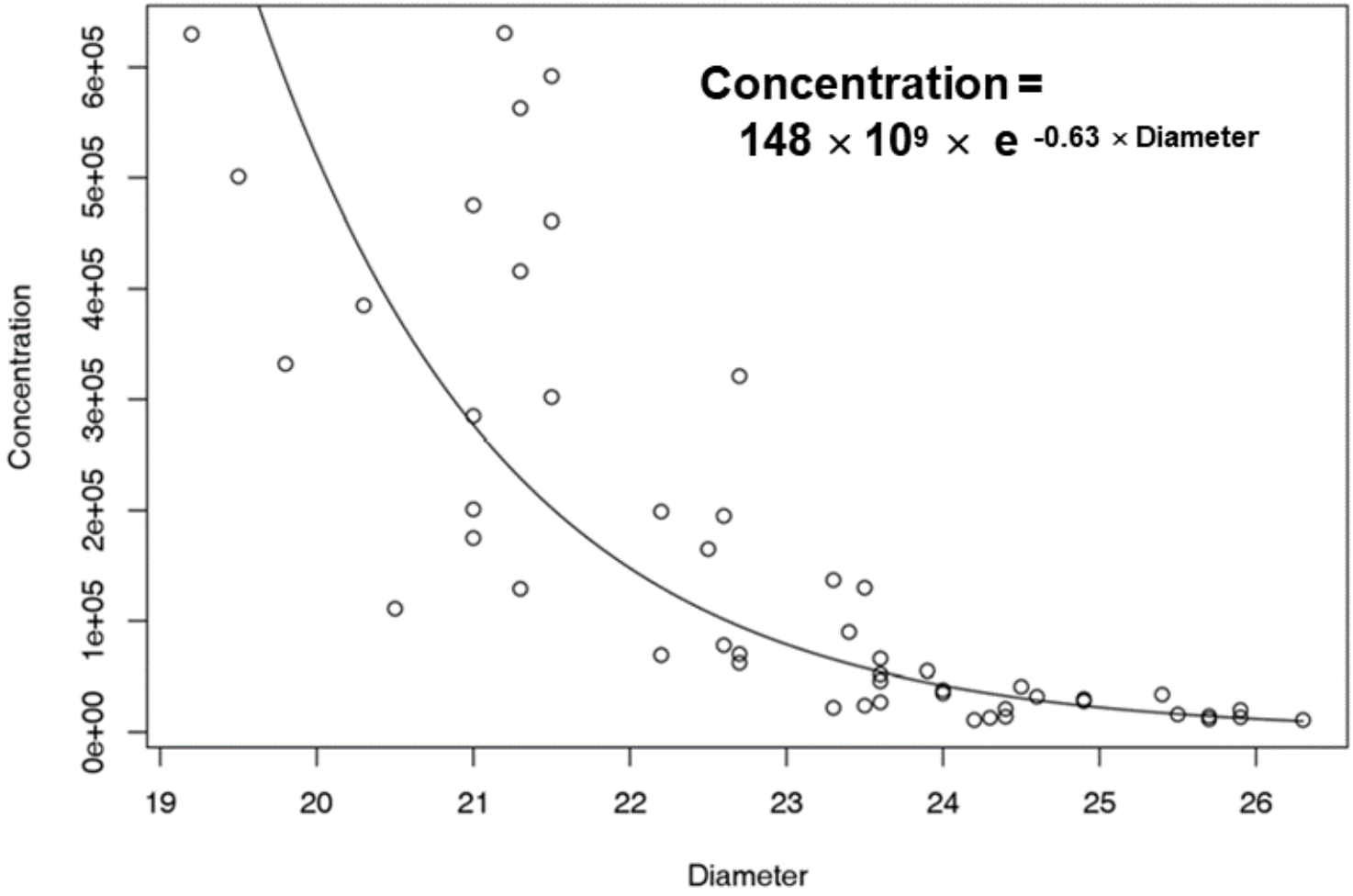
```
summary(modellog)

    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



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
   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.

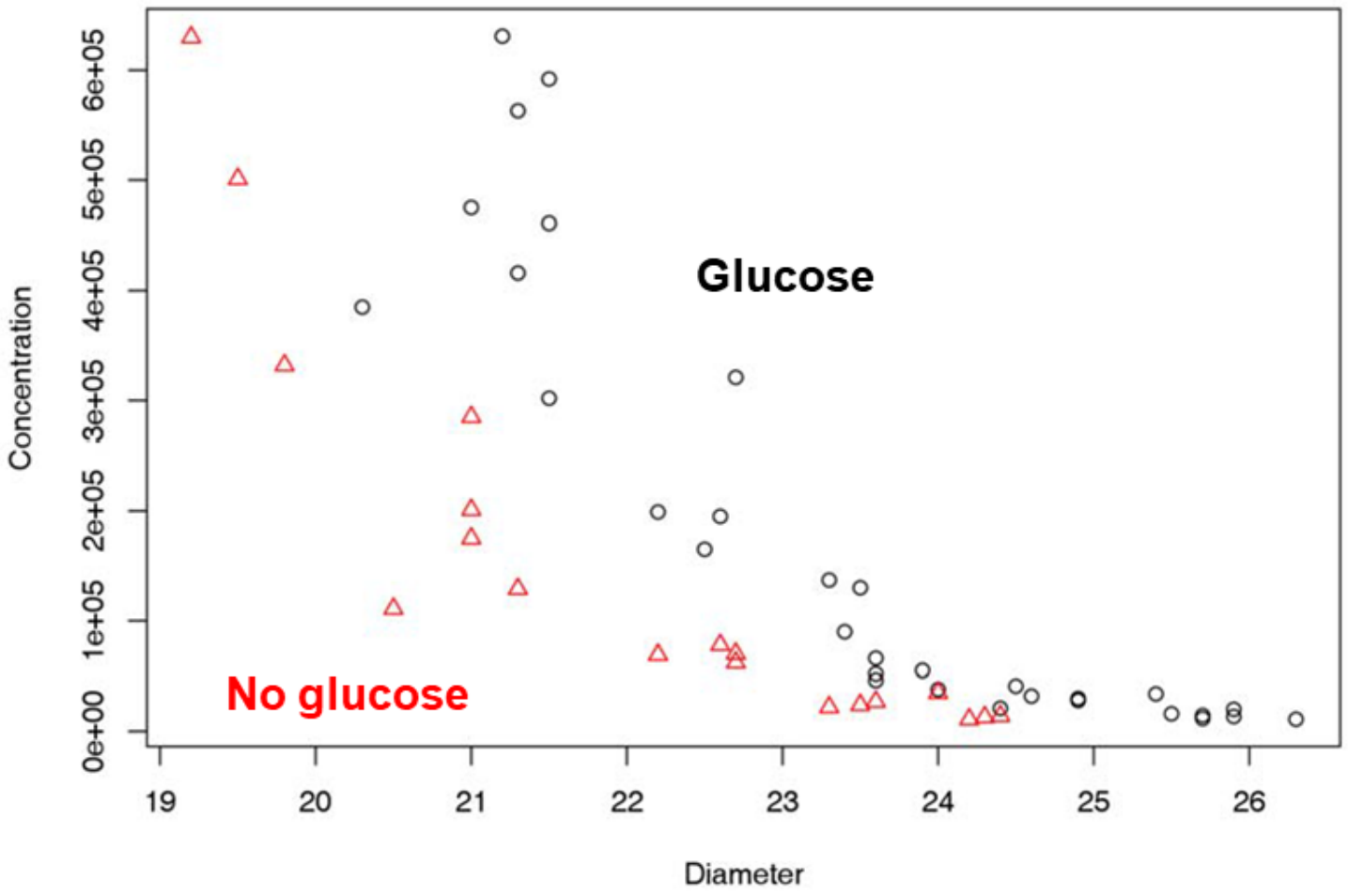
'conc' a numeric vector, cell concentration (counts/ml).

'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



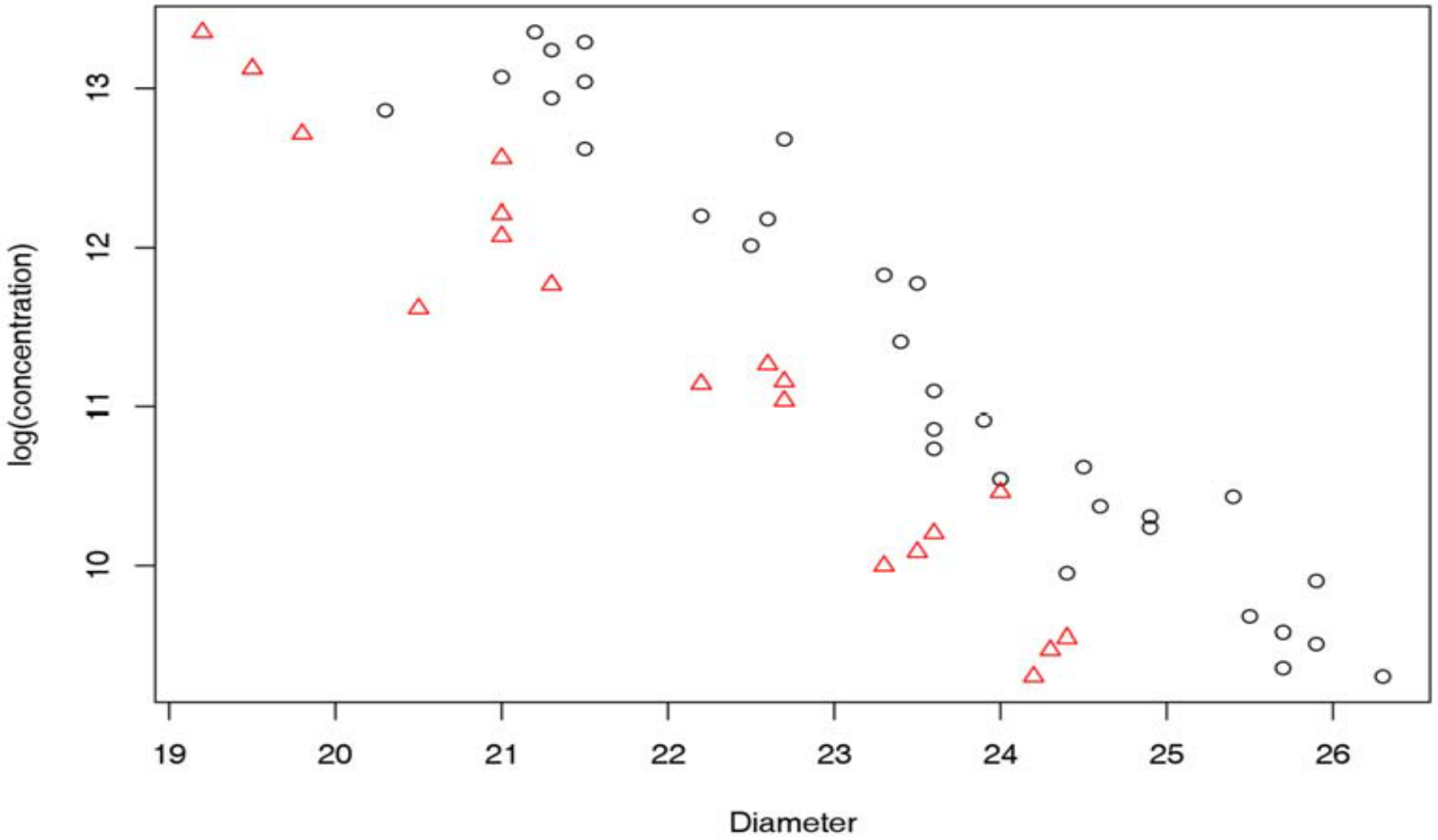
Reminder: using categorical variables as explanatory variables

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

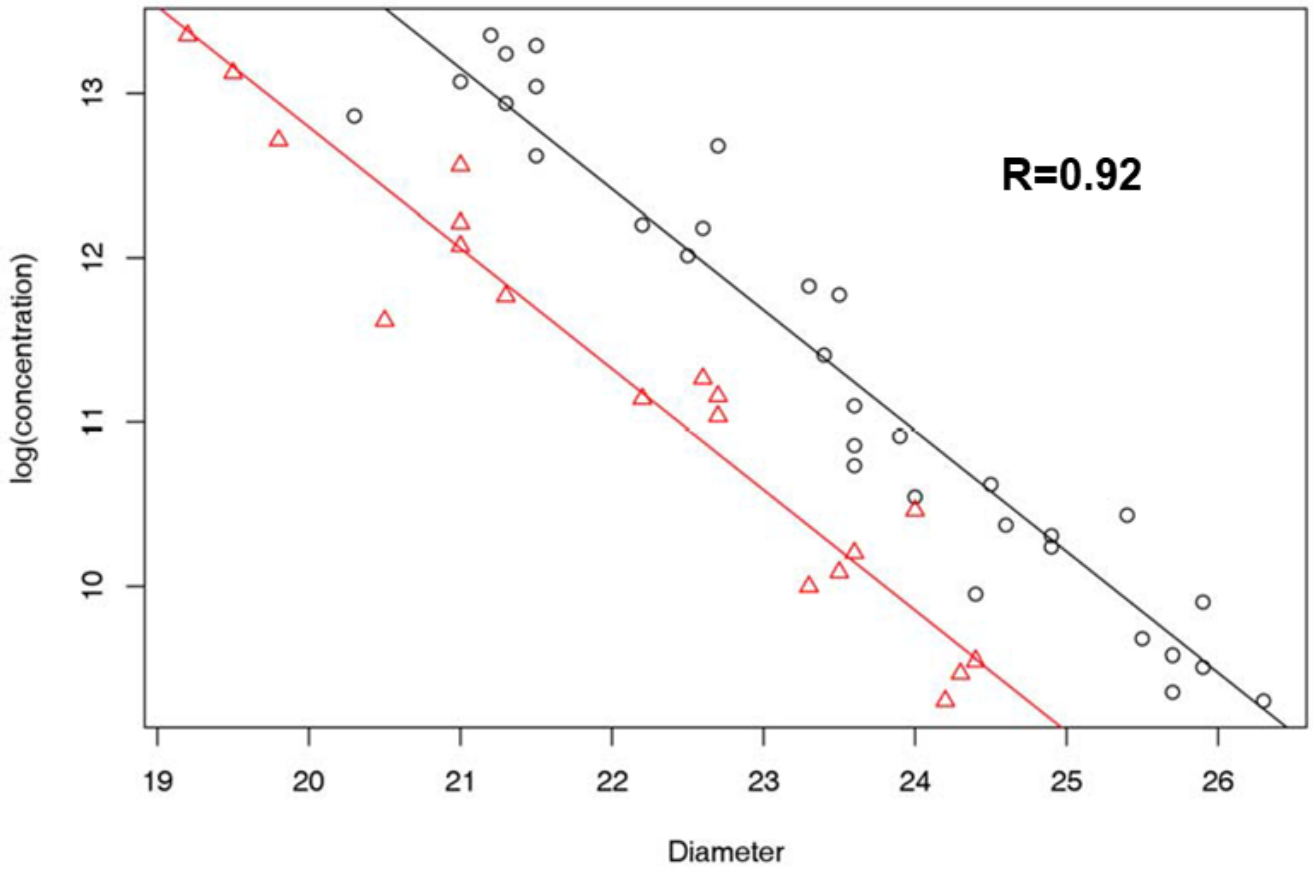
Concentration $= b_0 + b_1$ Diameter + b_2 « Glucose » + error

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

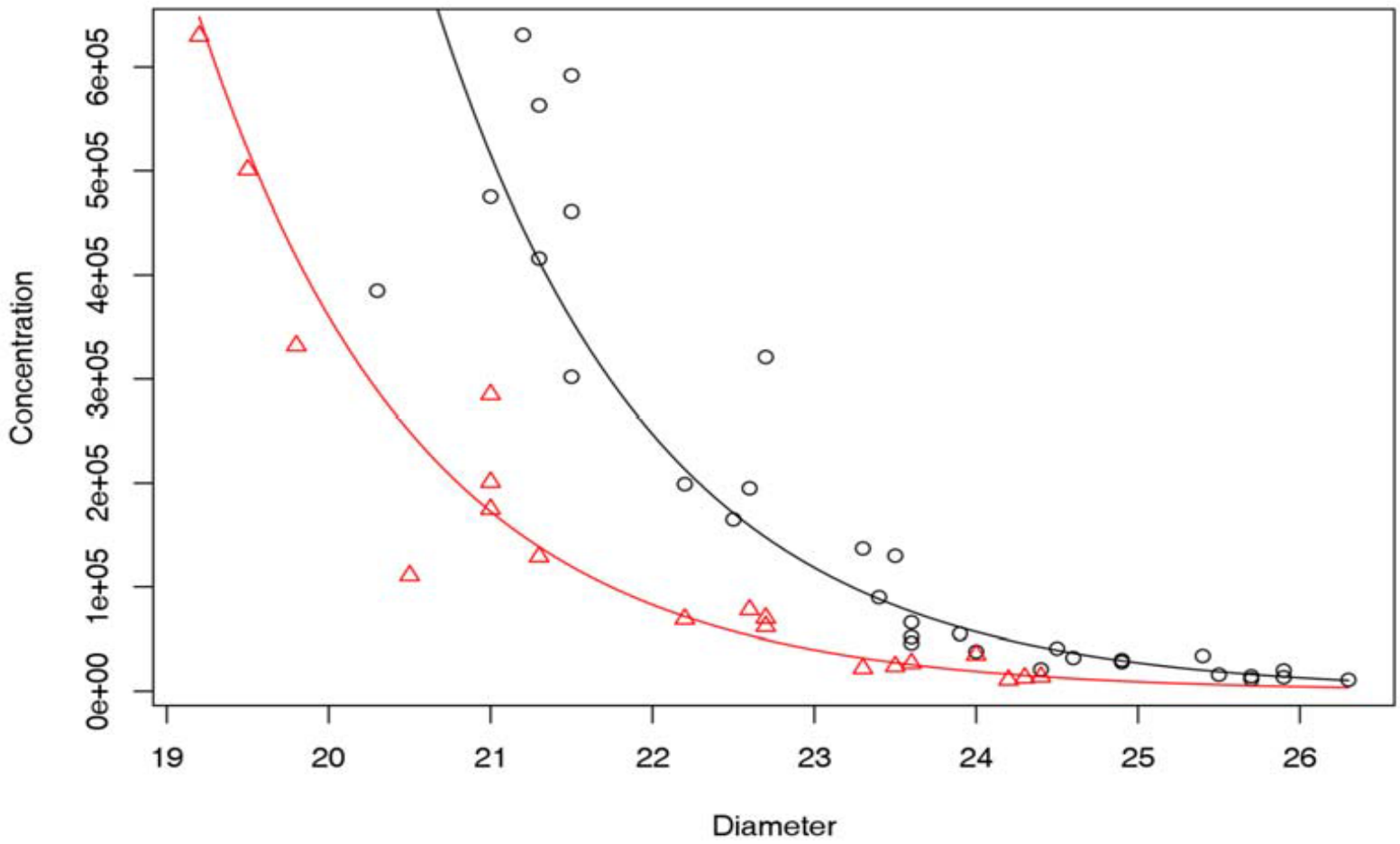
Log(concentration) according to diameter and glucose



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.