Regression Methods

Summary of tests

- Significance
- Confidence Intervals
- Hypotheses
- tests: t-test and ANOVA
- Equal/Not equal variance
- Post hoc test

Correlation analyisis

We have two 2 variables (bivariate)…

Volume of trees

(it is difficult to measure)

Heigh of trees

(it is easy to measure)

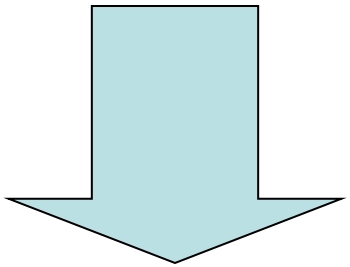
We want to know if there is any correlation between them, so in the future, we will only measure the height of the trees to measure the volume.

We take a sample and we check this correlation

Correlation

Study correlation between 2 variables

- We compare how a variable can explain the other' variance
- To describe this relationship, we use an analysis of correlation



pearson correlation coefficient

R

Correlation analysis

Study correlation between 2 variables

When we use pearson correlation coefficient, we assume:

- Variables are in interval scale
- Variables are **NORMALLY** distributed

Person coefficient can take values between -1 and +1

-1: Strongest correlation, indirect relationship

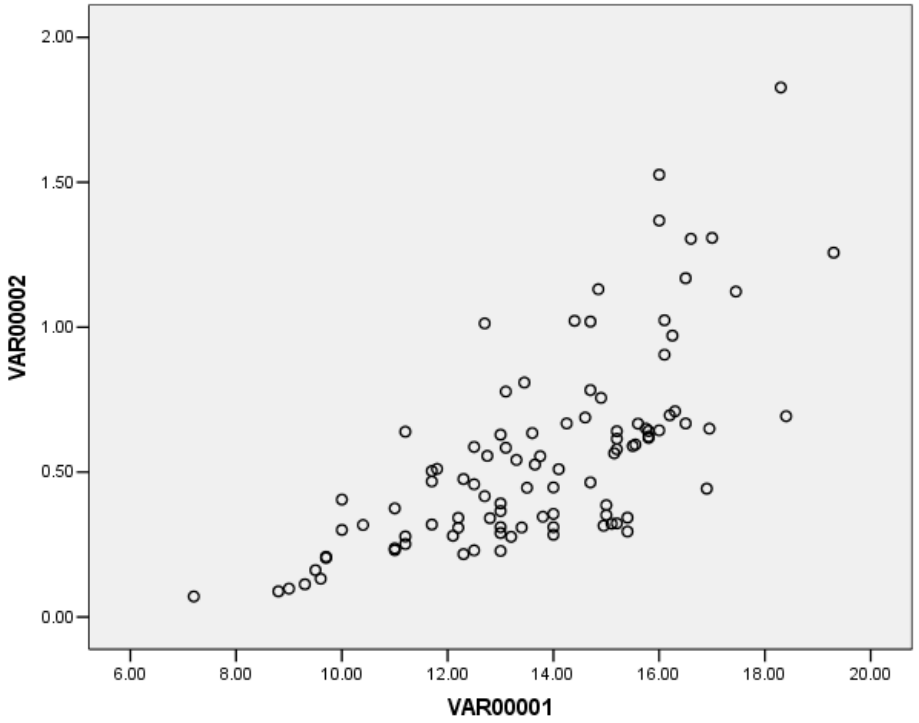
+1: Strongest correlation, direct relationship

Correlation analyisis

Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	.687**
	Sig. (2-tailed)		.000
	N	99	99
VAR00002	Pearson Correlation	.687**	1
	Sig. (2-tailed)	.000	
	N	99	99

**. Correlation is significant at the 0.01 level (2-tailed).



Correlation analyisis

Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	.687**
	Sig. (2-tailed)		.000
	N	99	99
VAR00002	Pearson Correlation	.687**	1
	Sig. (2-tailed)	.000	
	N	99	99

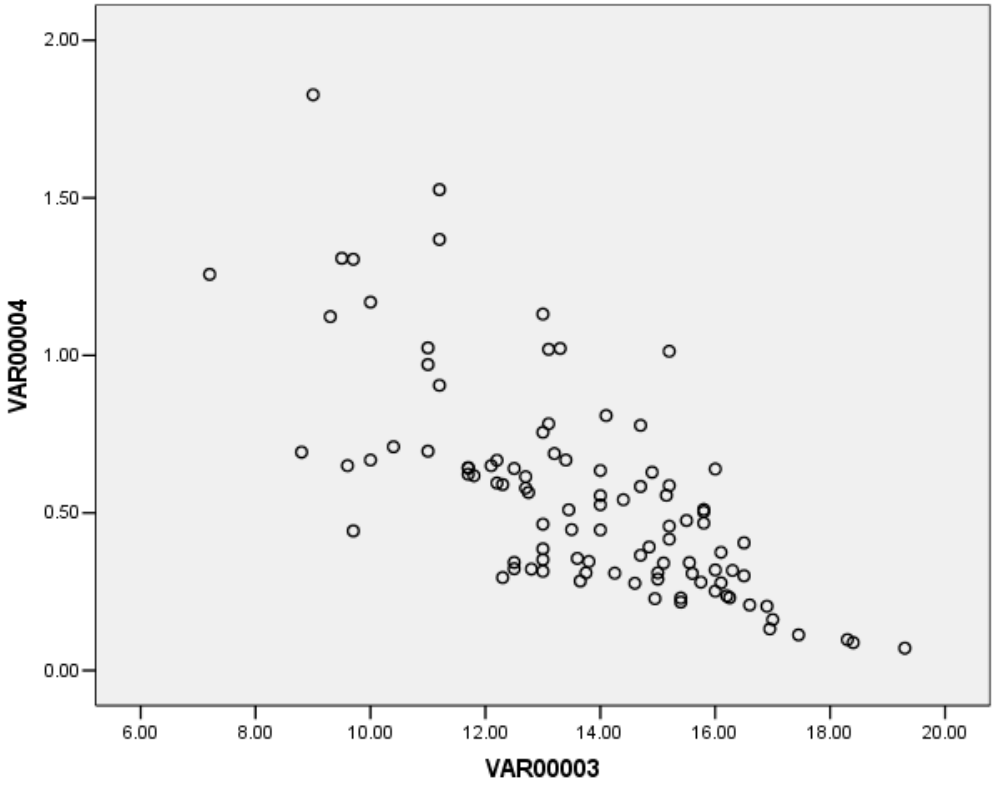
**. Correlation is significant at the 0.01 level (2-tailed).

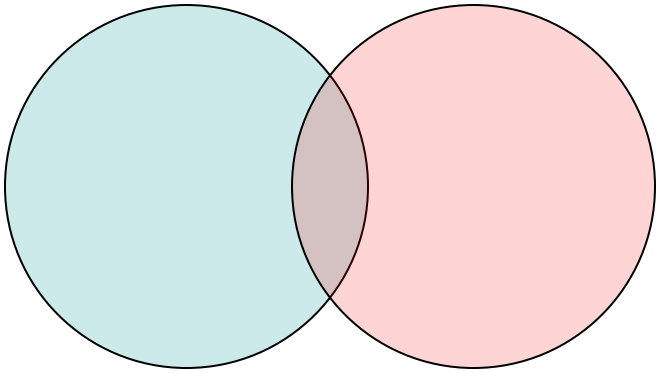
Correlation analysis

Correlations

		VAR00003	VAR00004
VAR00003	Pearson Correlation	1	-.716**
	Sig. (2-tailed)		.000
	N	99	99
VAR00004	Pearson Correlation	-.716**	1
	Sig. (2-tailed)	.000	
	N	99	99

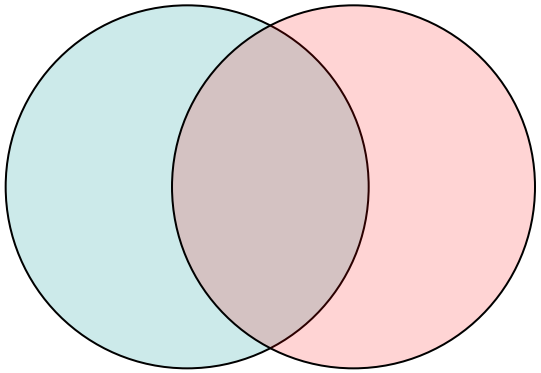
**. Correlation is significant at the 0.01 level (2-tailed).



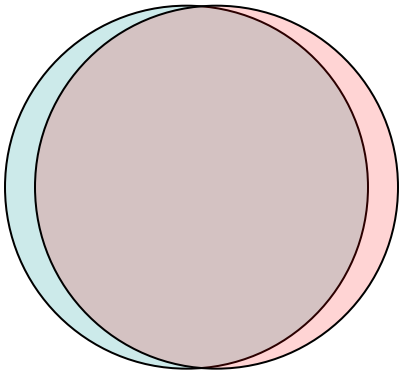


r=0,25

r=Pearson correlation coefficient



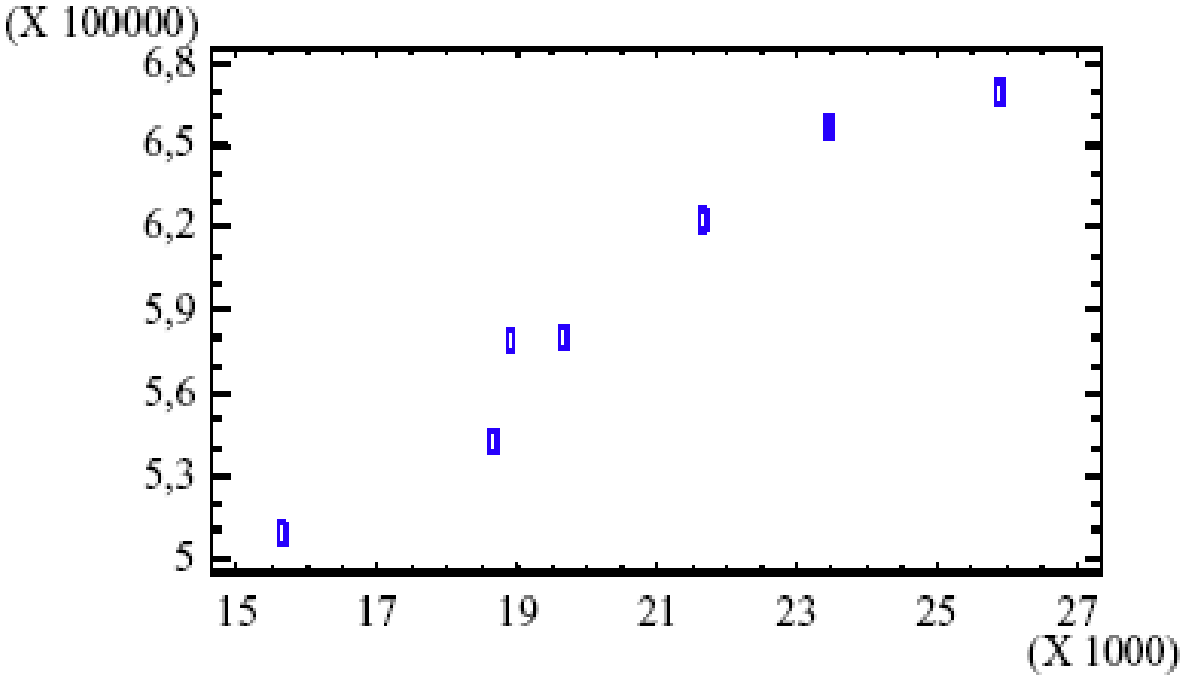
r=0,50



r=0,90

Correlation

Does it make sense?



Birth rate and stork correlate very well… but, does it make sense?

Models

Model: a simplified representation of some aspect of reality.

We frequently use models unconsciously: making mental maps to anticipate and explain the behaviour of systems.

Specifically in a **statistical context**, a model is a formalized expression of a theory.

A mathematical model is like a verbal model, but uses mathematical language (more concise, less ambiguous). Statistics are tools we use in the models, as well as computers.

[1] VANCLAY, J. 1994. “Modelling Forest Growth and Yield. Application to Mixed Tropical Forests” CAB International..

Who uses models?

Engineering

Economy

Biology

Sociology

Linguistics

Computer design

…

The right model

Every model will be wrong in some sense.

The right model is the one that is most useful for the application needed, and the choice has to be based in the application and the resources.

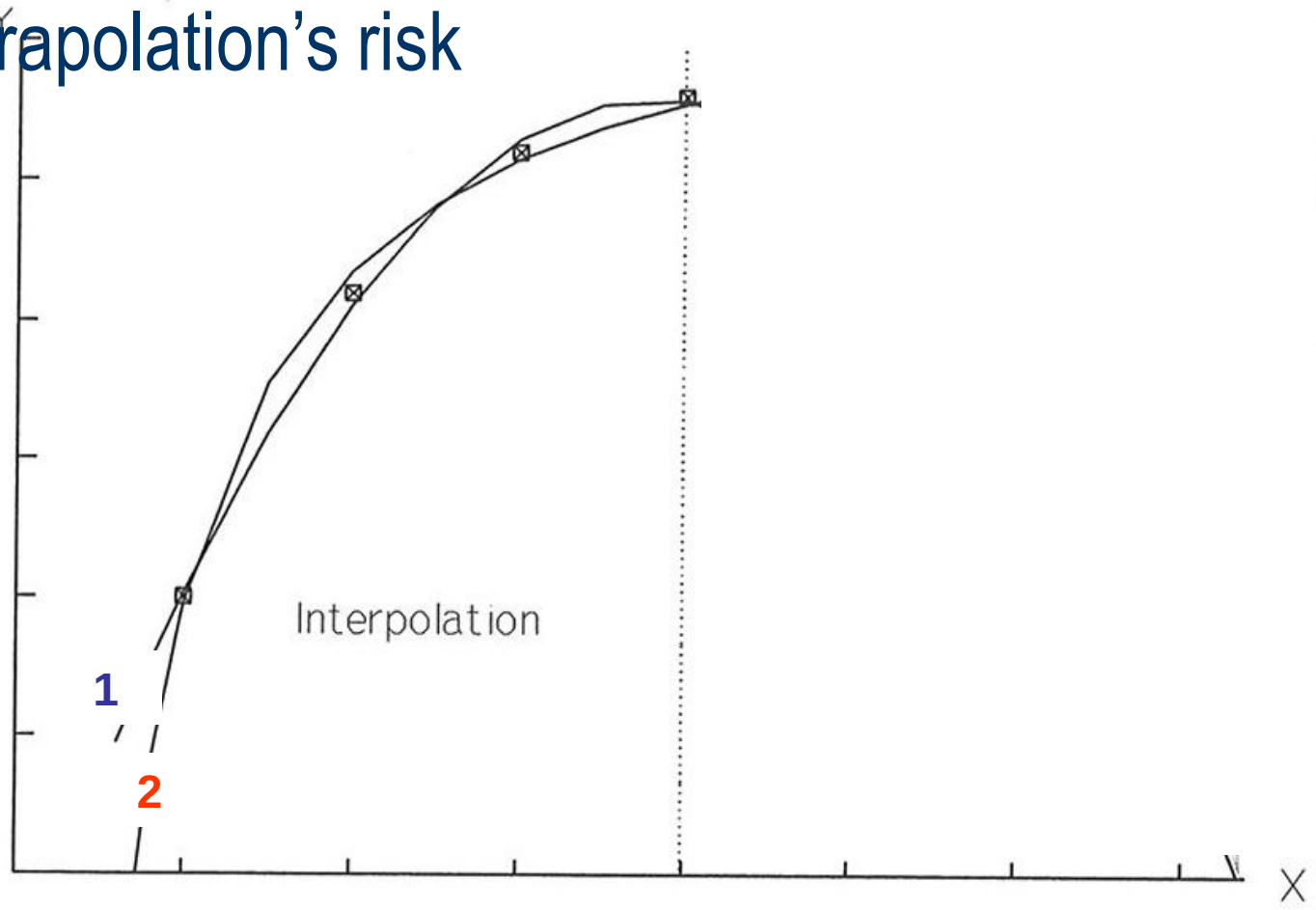
1. Does the approach make sense?
2. Will the model work for my application and input data?
3. What range of data was used to develop the model?
4. Do model assumptions and inferences apply to my situation?
5. What confidence can I place in model predictions?
6. Be sceptical and DEMAND PROOF!

[1] from: VANCLAY, J. 1994. “Modelling Forest Growth and Yield. Application to Mixed Tropical Forests” CAB International..

Models

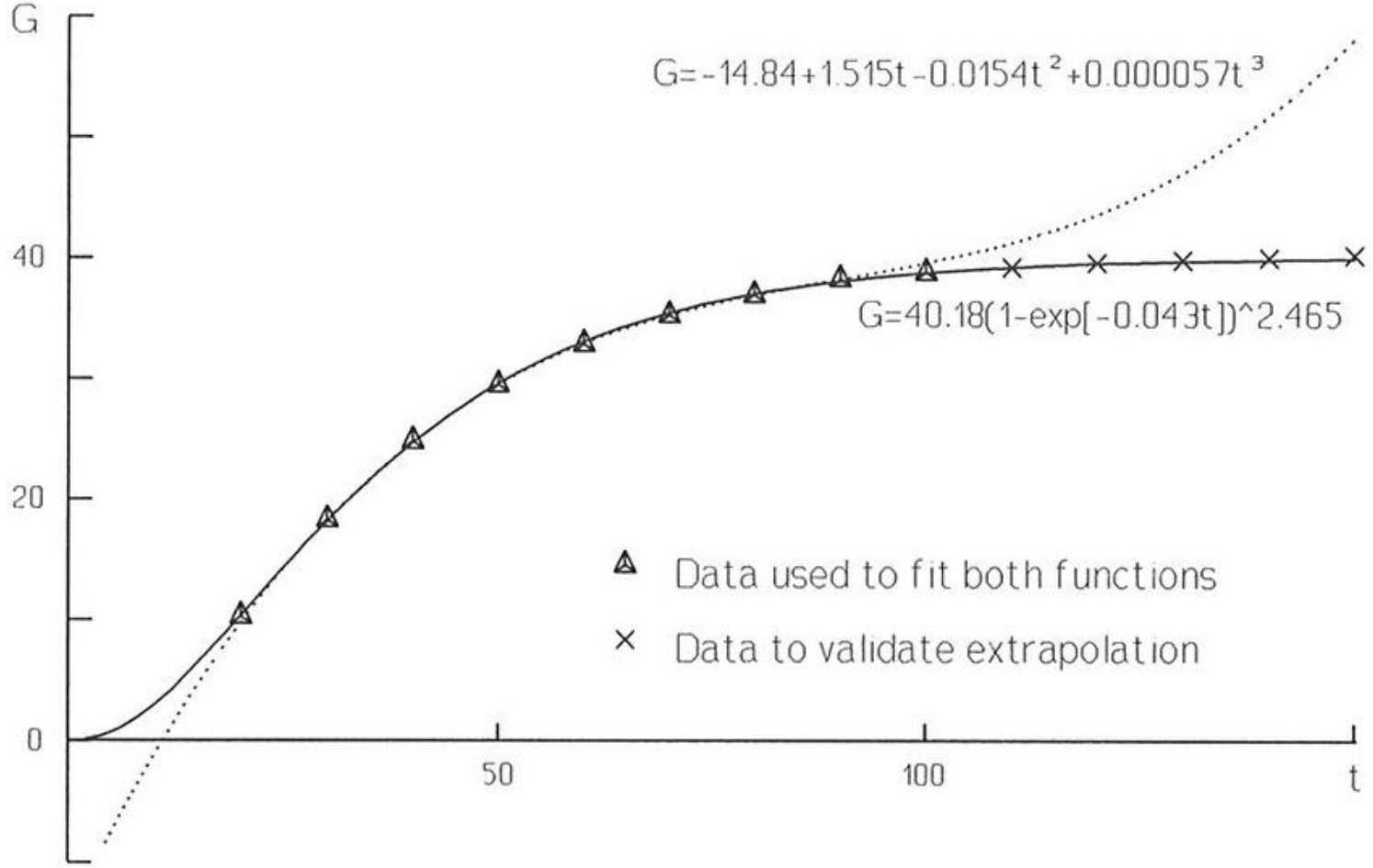
Extrapolation's risk

R^2>0.996

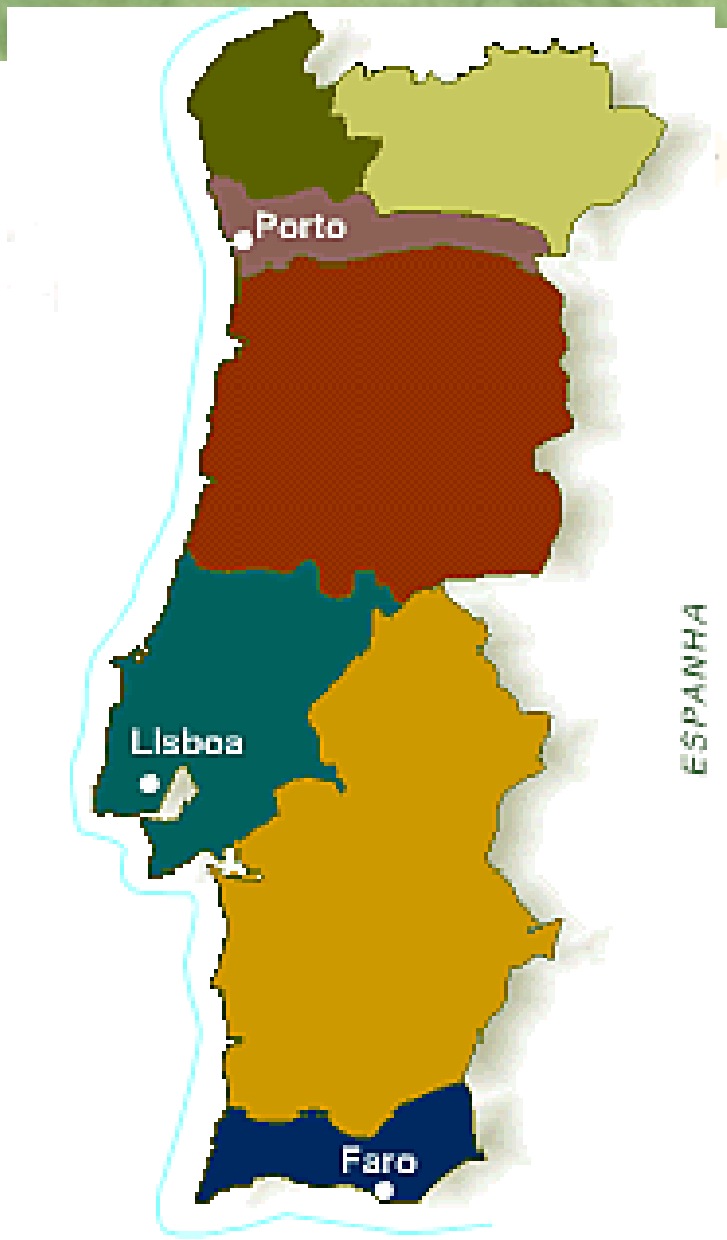


Models

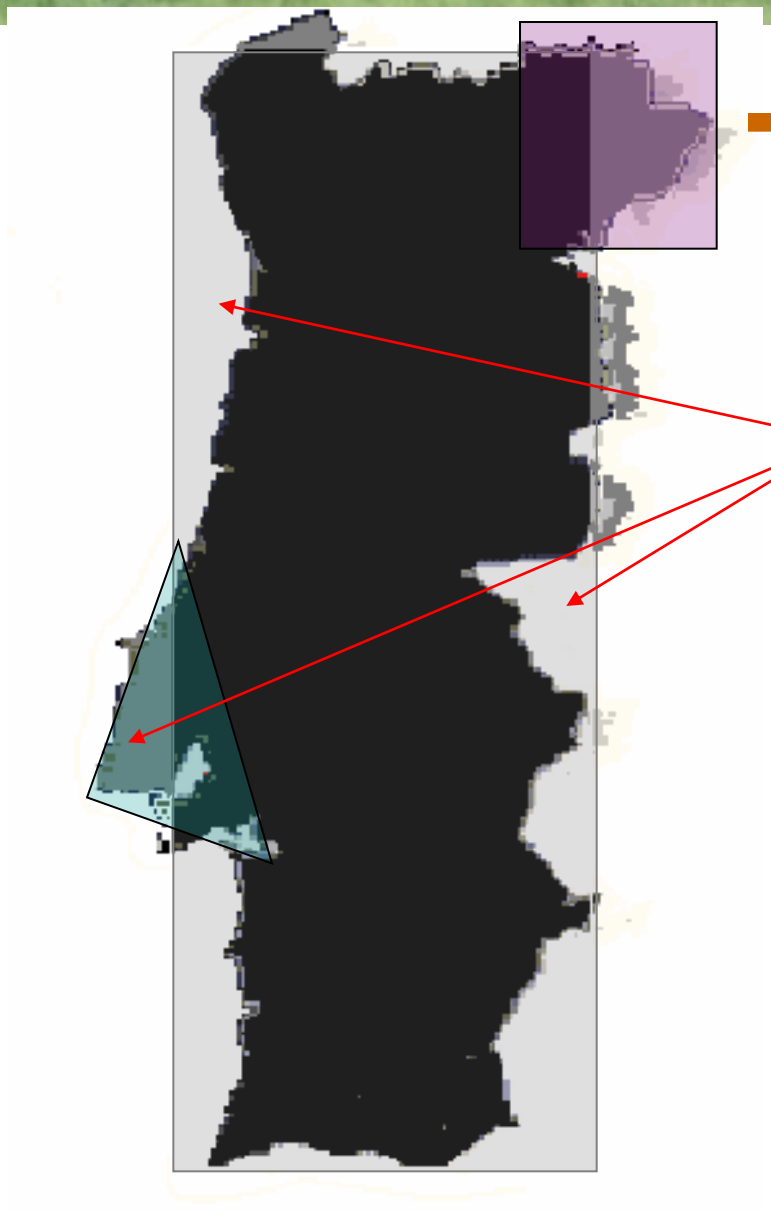
Validation



Modelling Portugal



Modelling Portugal



ε (*error*): We can not explain these areas with our model

New Variables: *can increase the accuracy, but the model might turn too complex.*

Models

Construction steps

General principles:

1) **Parsimony** (or *Occam's razor)*: Entities should not be multiplied beyond necessity. Do not include unnecessary variables and parameters in the model

2) **Simplicity**: Unnecessary complexity does not improve the model, and may create many problems. Keep it simple.

- What variables to include?
- What equation to use?
- How to fit the equation to the data?

[1] from: VANCLAY, J. 1994. “Modelling Forest Growth and Yield. Application to Mixed Tropical Forests” CAB International..

Linear Regression

Using some models

The linear regression model assumes that there is a linear, or "straight line," relationship between the dependent variable and each predictor. This relationship is described in the following formula.

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i$$

Simple linear regression (1 Variable)

Multiple linear regression (Several Variables)

Regression

Using some models

For the purpose of testing hypotheses about the values of model parameters, the linear regression model also assumes the following:

- The error term has a normal distribution with a mean of 0.

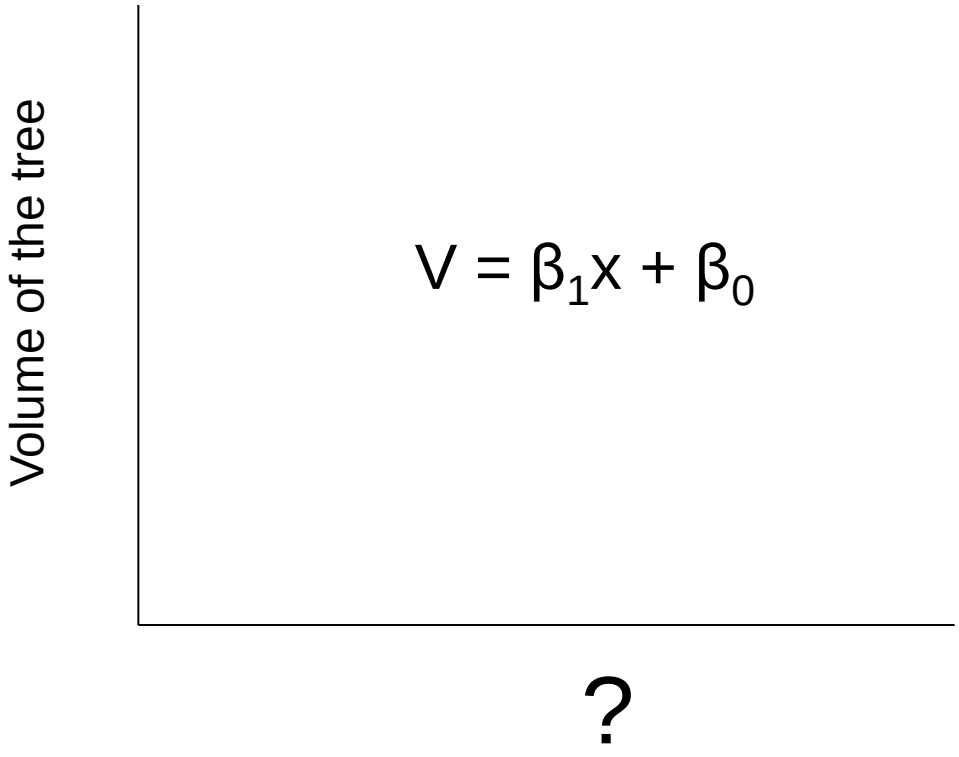
- The variance of the error term is constant across cases and independent of the variables in the model. An error term with non-constant variance is said to be **heteroscedastic**.

- The value of the error term for a given case is independent of the values of the variables in the model and of the values of the error term for other cases.

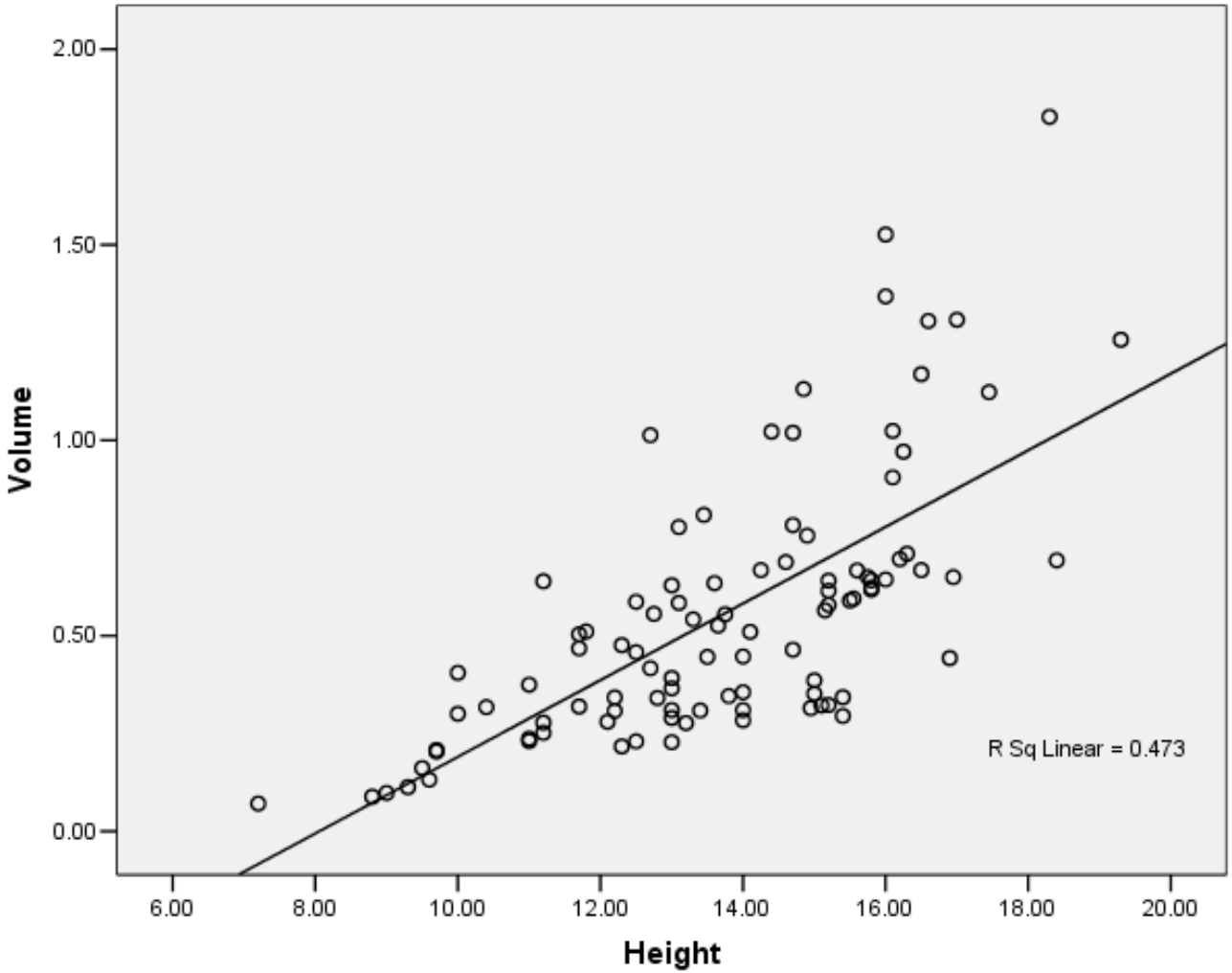
Regression

Example

Measuring trees



Regression


2.00
1.50
1.00
0.50
0.00
Volume
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
Height
R Sq Linear = 0.473

Regression

The strength of the relationship between the model and the dependent variable can be explained by R2.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.687[a]	.473	.467	.24480

a. Predictors: (Constant), Height

b. Dependent Variable: Volume

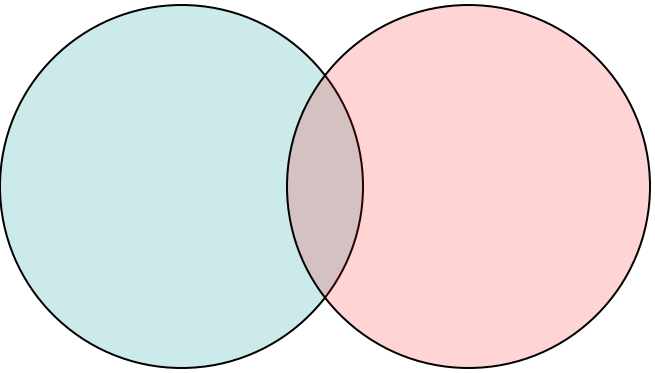
ANOVA table tests the model's ability to explain any variation in the dependent variable (it does not directly address the strength of that relationship).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.208	1	5.208	86.910	.000[a]
	Residual	5.813	97	.060		
	Total	11.021	98			

a. Predictors: (Constant), Height

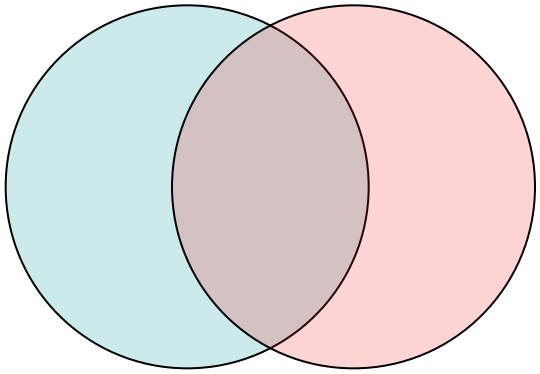
b. Dependent Variable: Volume



r=0,25

r^2=0,06

r=Pearson correlation coefficient



r=0,50

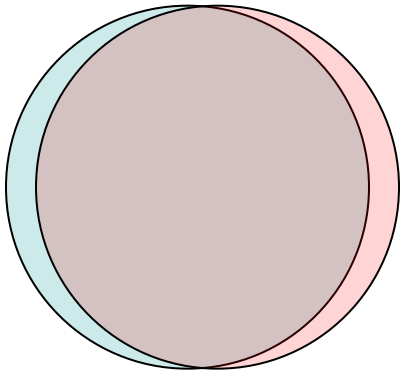
r^2=0,25

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.687[a]	.473	.467	.24480

a. Predictors: (Constant), Height

b. Dependent Variable: Volume



r=0,90

r^2=0,81

Descriptive Statistics

	N	Mean	Std. Deviation
Volume	99	.5556	.33535
Valid N (listwise)	99		

Regression

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.208	1	5.208	86.910	.000[a]
	Residual	5.813	97	.060		
	Total	11.021	98			

a. Predictors: (Constant), Height

b. Dependent Variable: Volume

It is a valid model

The regression and residual sums of squares are approximately equal, which indicates that about half of the variation of volume is explained by the model.

Variation not explained by our model

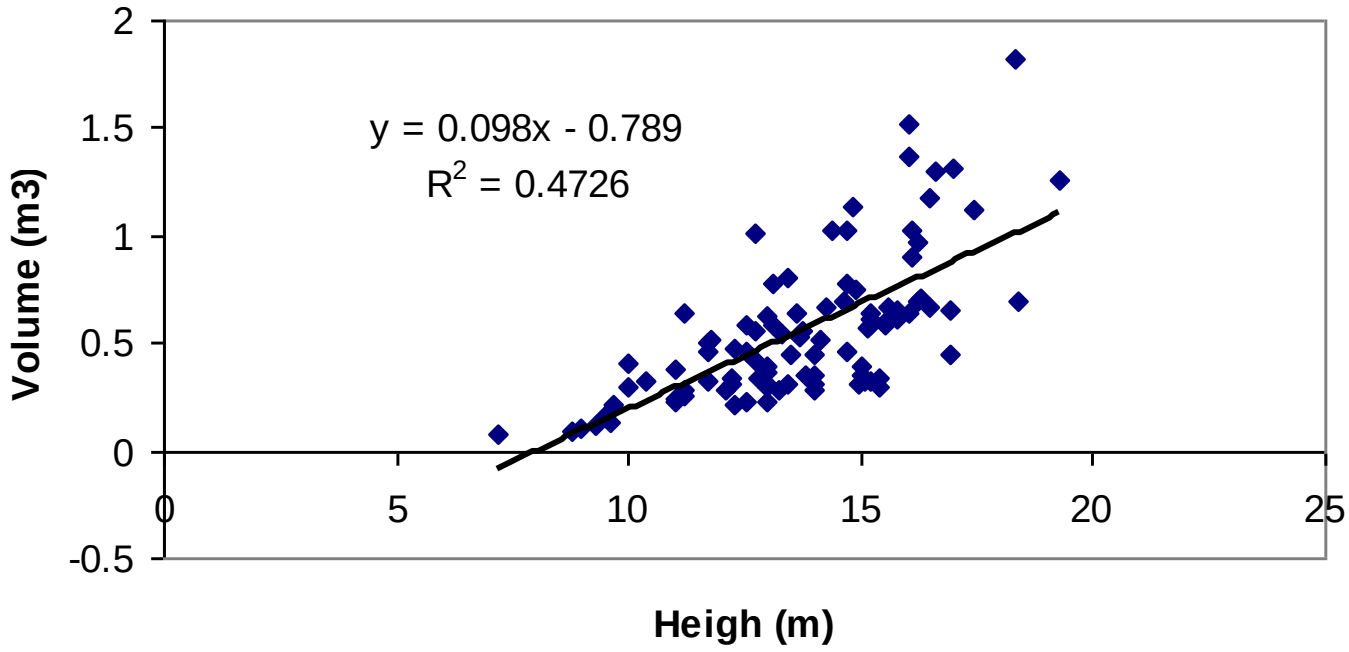
Variation accounted by our model

Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.789	.146		-5.392	.000
	Height	.098	.011	.687	9.323	.000

a. Dependent Variable: Volume



Exercise

MSc European Forestry

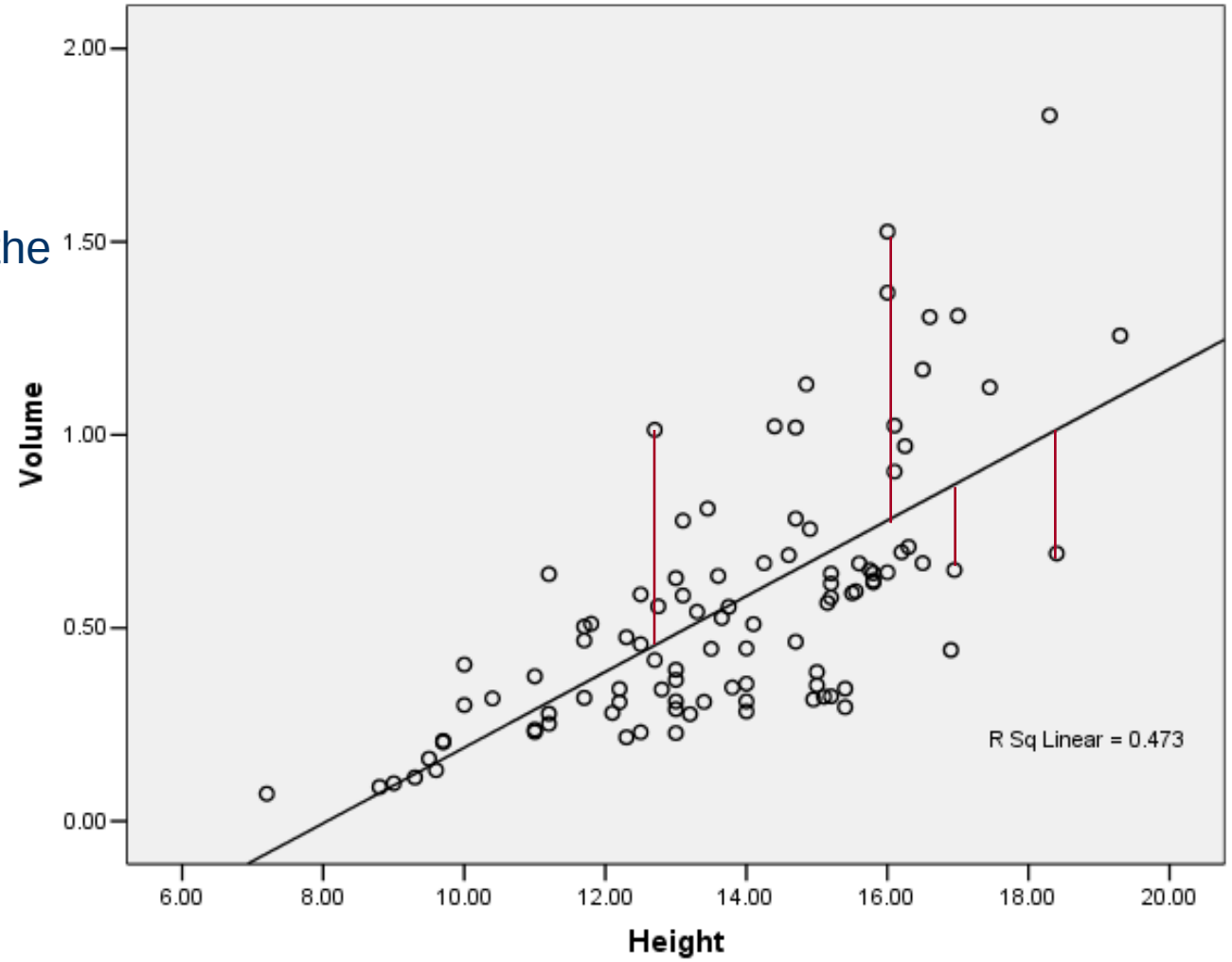
Regression methods

- Analysis of the residuals

Regression

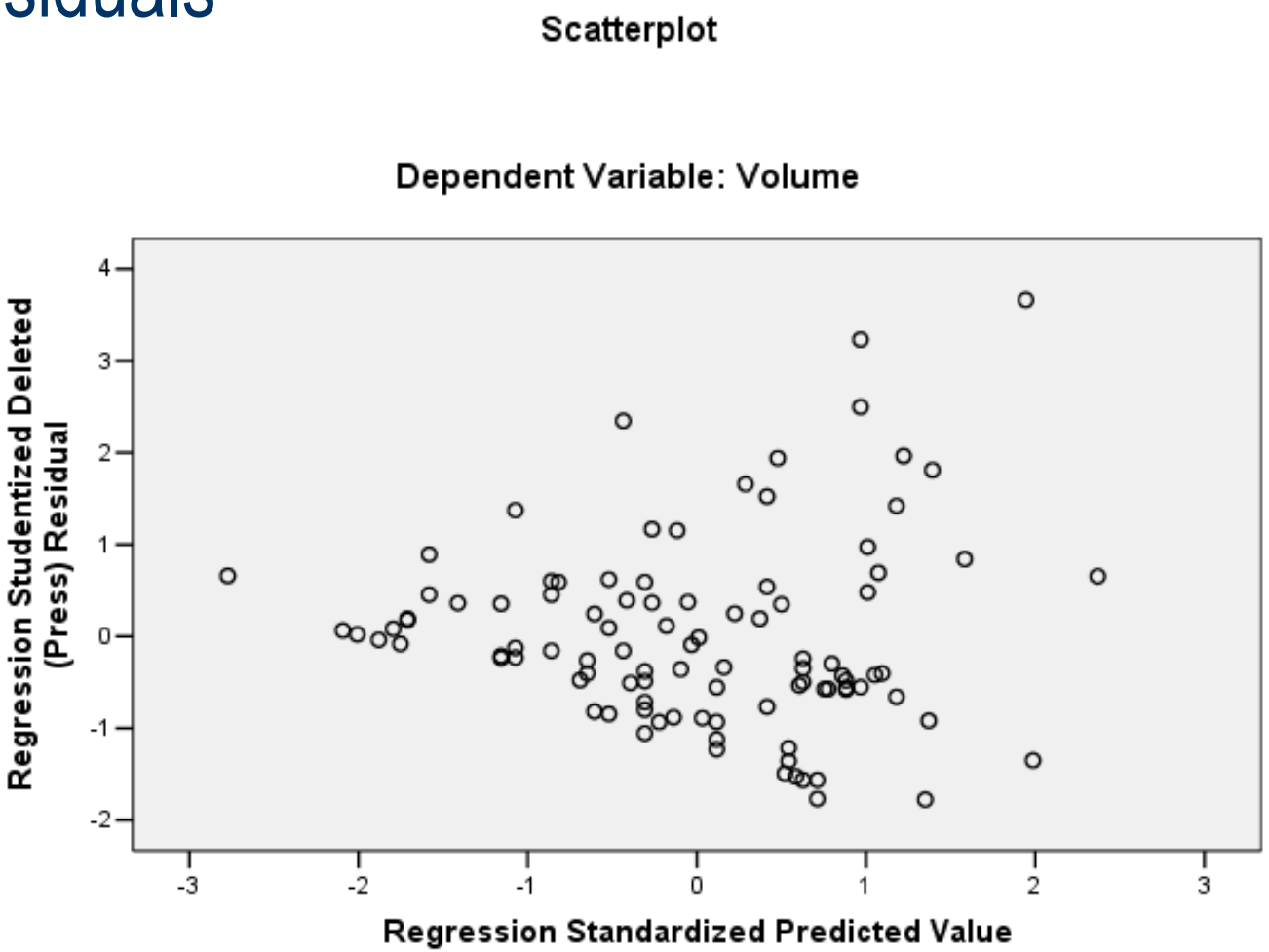
Residuals

A residual is the difference between the observed and model-predicted values of the dependent variable.



Regression

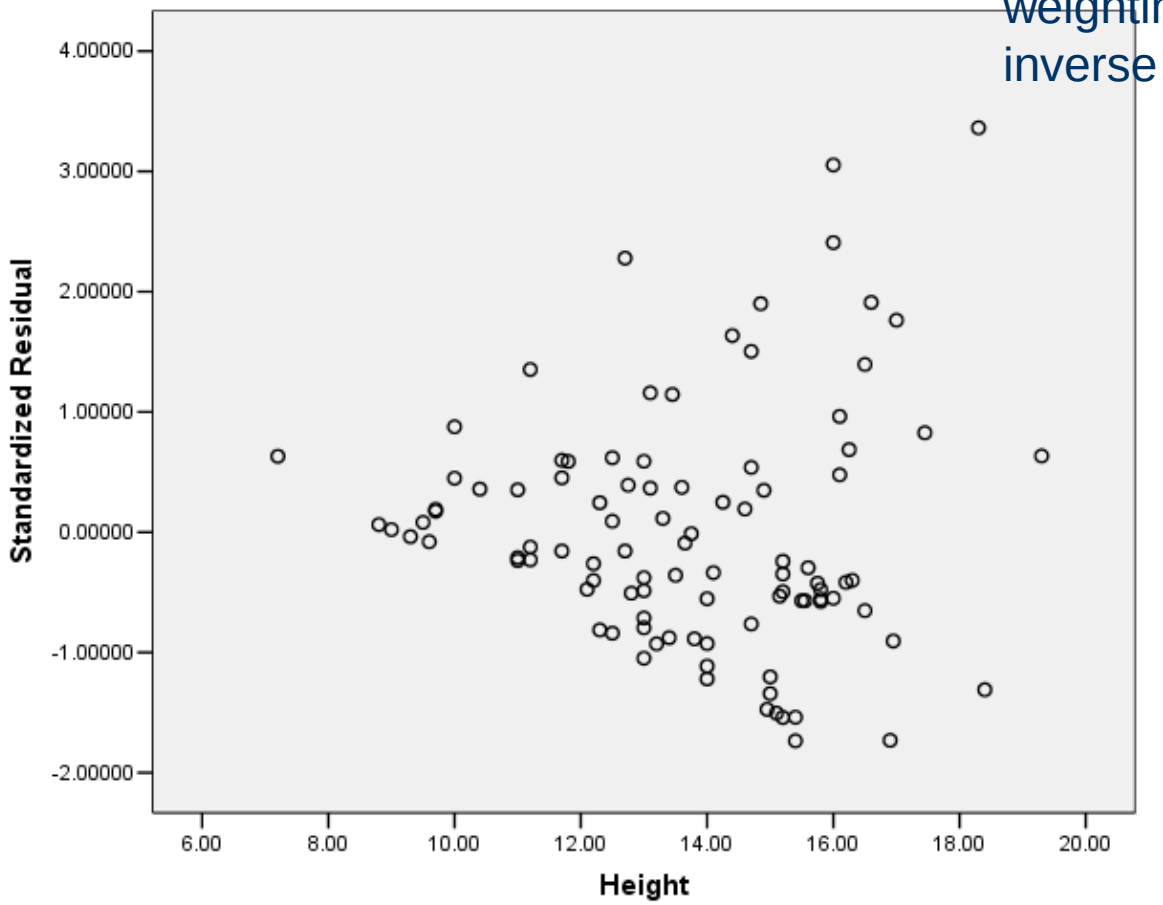
Residuals



Regression

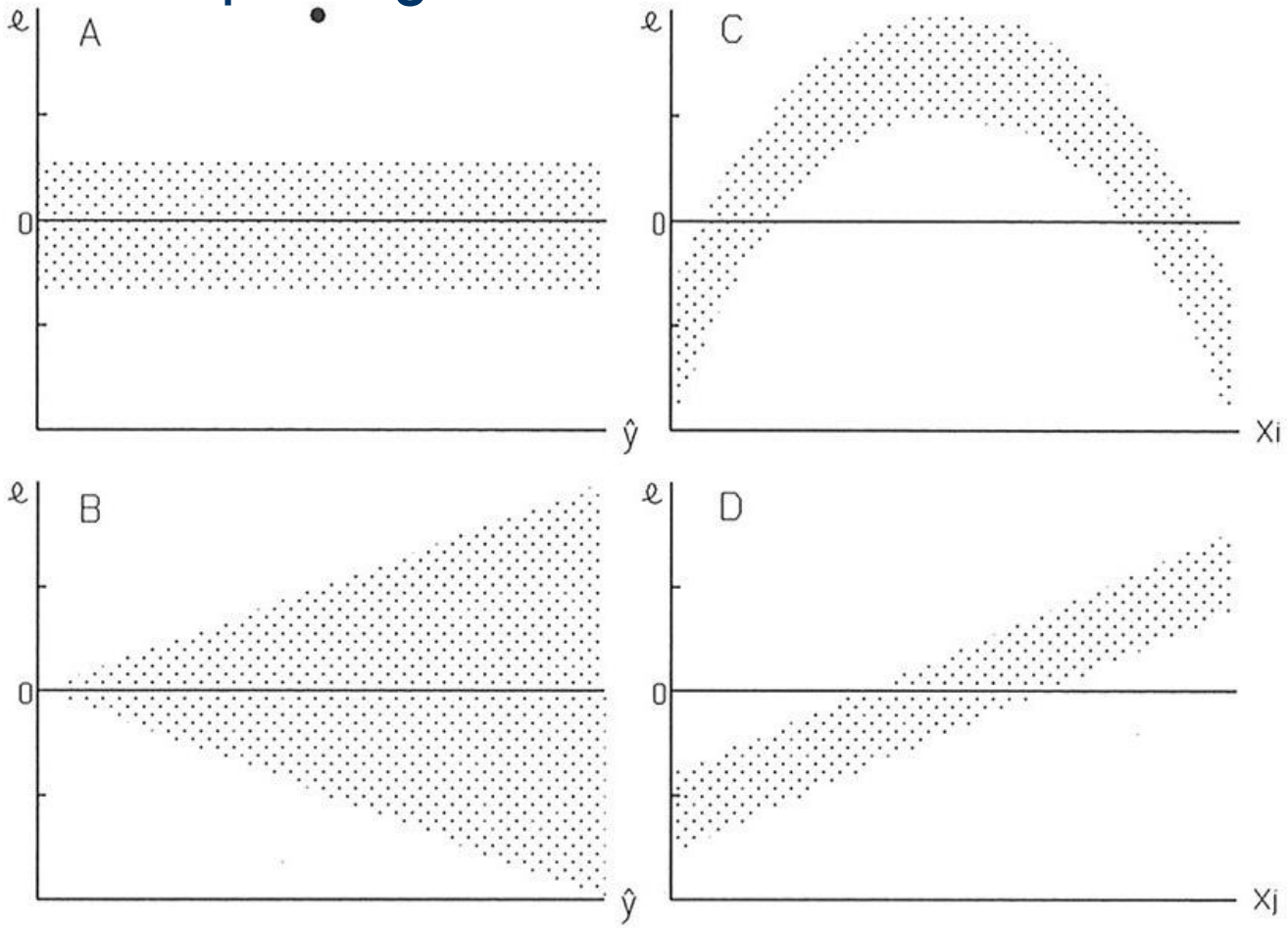
Residuals

To correct the heteroscedasticity in the residuals in further analyses, you should define a weighting variable based on the inverse of height (in this case)



Regression

Interpreting Residual Plots



a) Outlier

b) non-constant variance

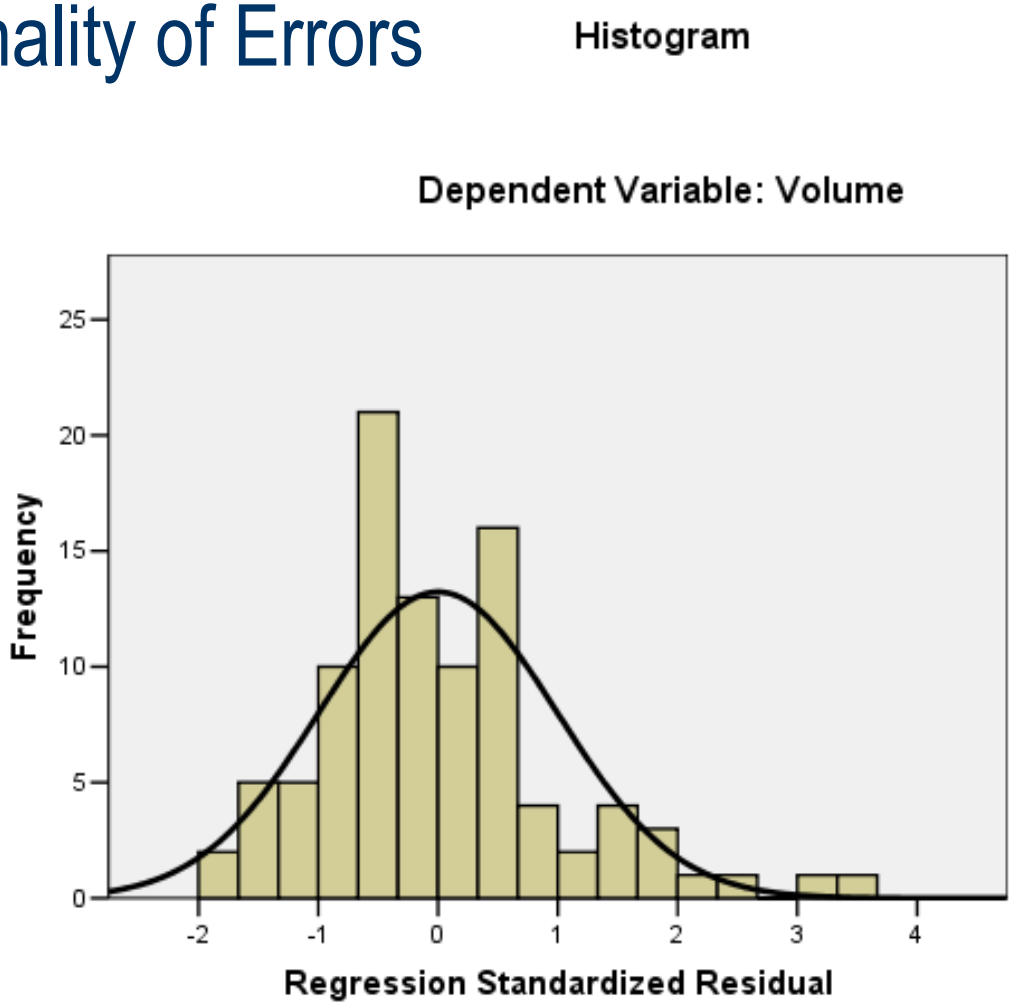
c) transformation required

d) variable Xj should be included in the model

[1] from: VANCLAY, J. 1994. "Modelling Forest Growth and Yield. Application to Mixed Tropical Forests" CAB International..

Regression

Normality of Errors



The residual for a given product is the observed value of the error term for that product. A histogram or P-P plot of the residuals will help you to check the assumption of normality of the error term.

Regression

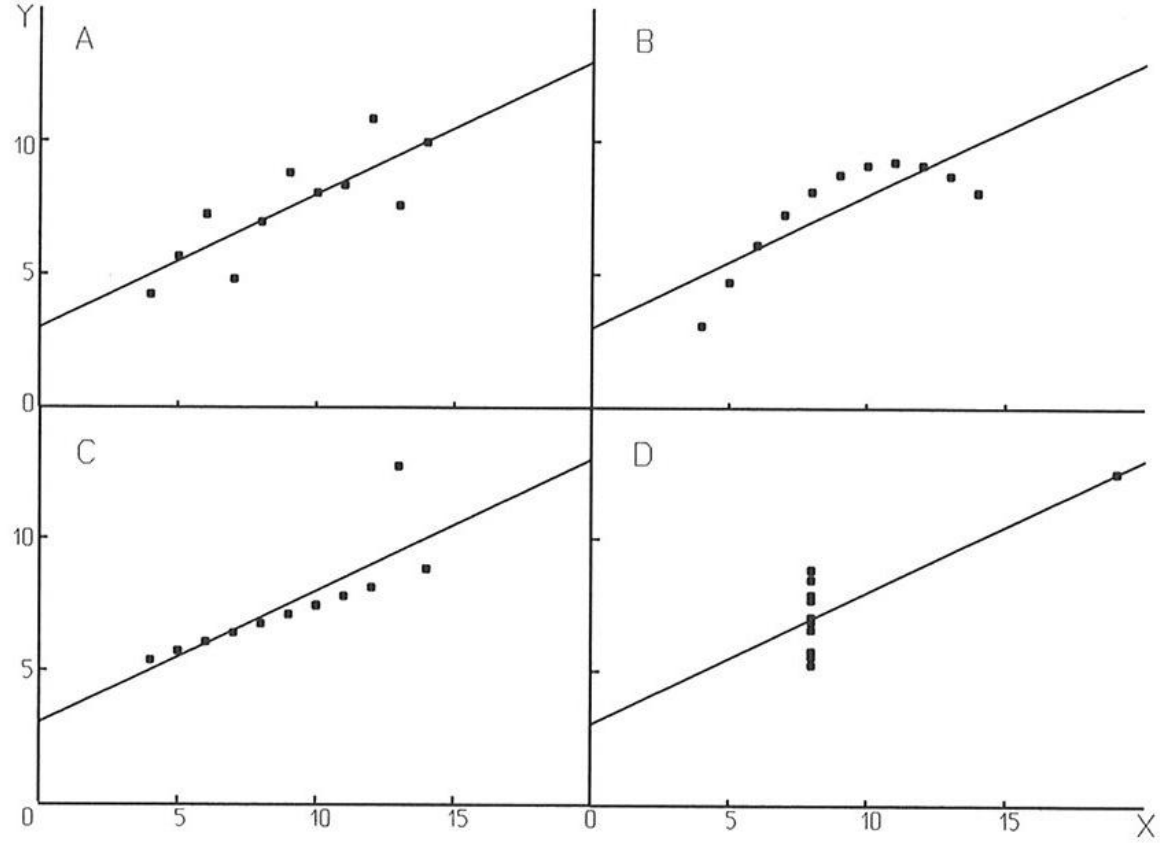
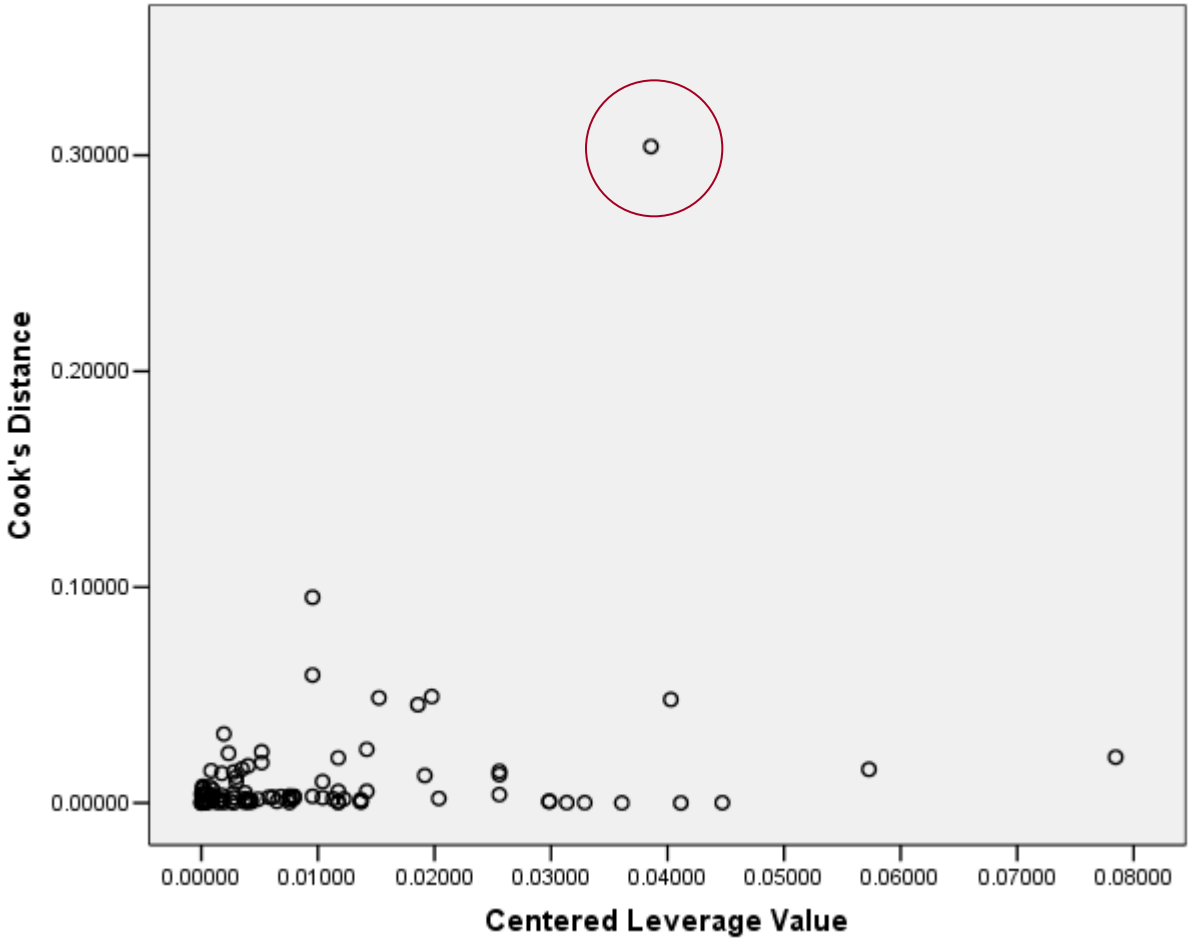


Fig. 6.9. The R^2 does not indicate how well a model fits the data. Here the plots reveal (A) pure error, (B) the wrong model, (C) an outlier, and (D) a point with high leverage, but all cases have $R^2 = 0.667$ (redrawn from Anscombe 1973).

[1] from: VANCLAY, J. 1994. “Modelling Forest Growth and Yield. Application to Mixed Tropical Forests” CAB International..

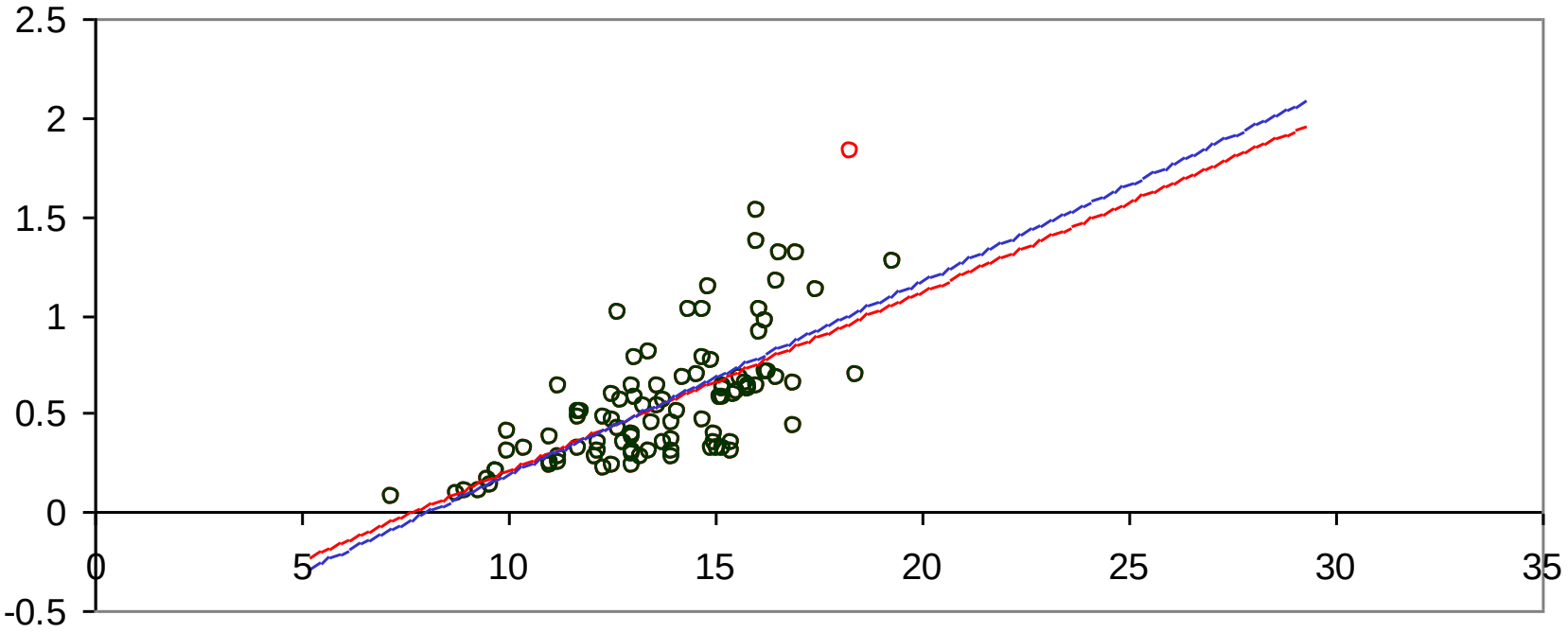
Regression

Influential values



Regression

Influential values



Some recomendations

1) Design of the model should be first. It has to be simple and it has to meet the needs.

2) Understanding of the limitations of the data and the implications they have for the model and analyses

3) Always plot the data and the fitted model to visually examine the quality of the fit

4) To know one's own limitations.

Summary

Basic concepts

- Correlation
- Pearson's correlation coefficient (r)
- Models
- Simple Regression
- Errors in Simple regression
- R^2