Tools for complex data analysis
University of Koblenz-Landau 2018/19

Introduction to the linear model: simple linear regression
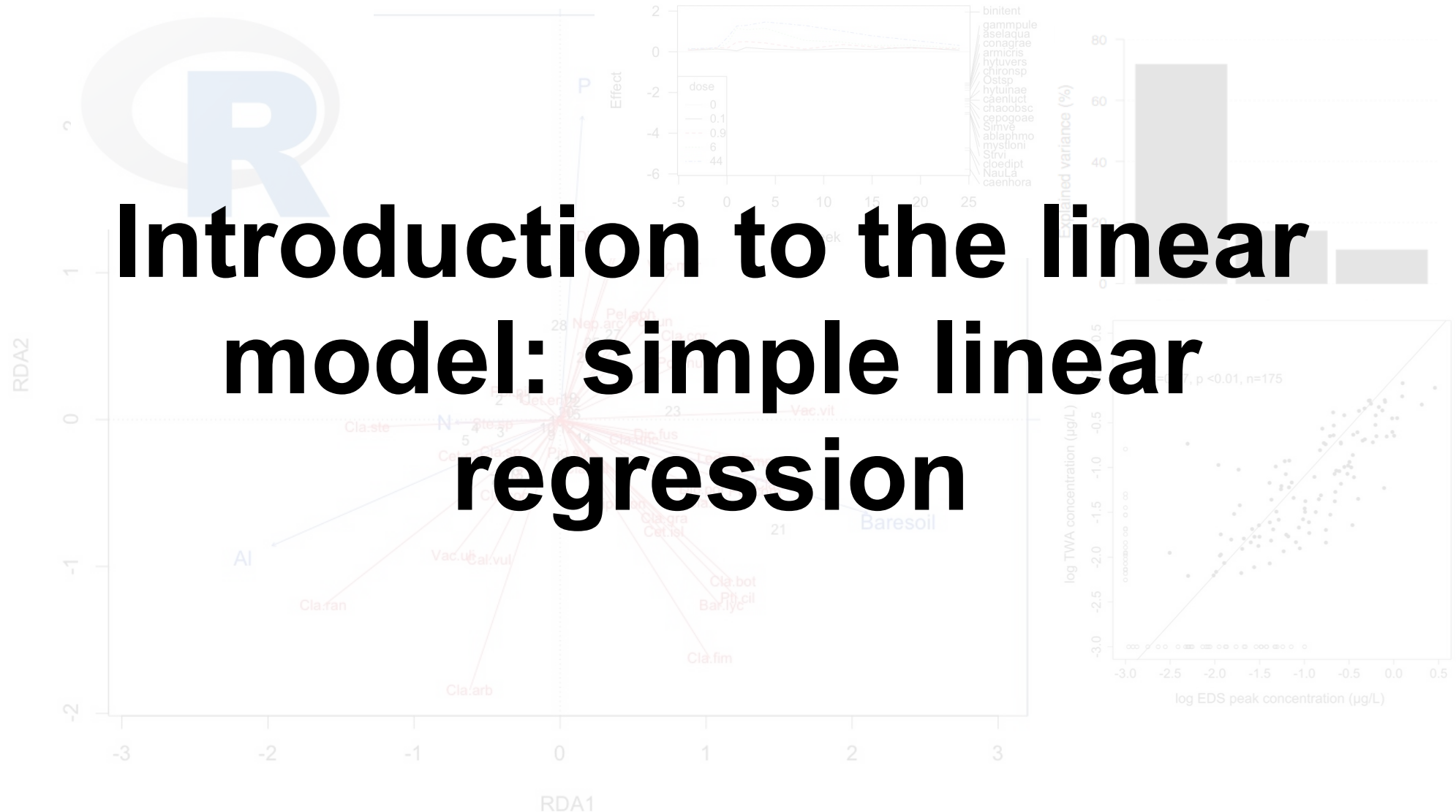
Ralf B. Schäfer

Learning targets

- Understanding the research goals and concept of the linear model

- Knowledge on the calculation of the model and of accuracy measures

- Understanding confidence intervals in general and in the regression context

- Ability to assess overall model accuracy and interpret model output

- Understanding and diagnosing model assumptions

Learning targets and study questions

- Understanding the research goals and concept of the linear model

  - Give an example of a study question for each research goal that can be tackled with a linear regression model.

  - What does the linear regression model predict? Explain the assumption of linearity in this context.

  - Define the residual and explain its role in finding optimal regression coefficients.

- Knowledge on the calculation of the model and of accuracy measures

  - Explain how b_1 relates to the joint variance of x and y.

  - Explain the concept of the standard error regarding accuracy.

  - What is the MSE?

  - How do the formulas inform a) that the SE of regression coefficients decreases with a wider range of x values and b) that the SE of fit decreases towards the centre of data?

Learning targets and study questions

- Understanding confidence intervals (CIs) in general and in the regression context

  - Explain the correct interpretation of CIs and two misleading interpretations.

  - Describe the change of the CI with the confidence level.

- Ability to assess overall model accuracy and interpret model output

  - Define R^2 and outline its calculation.

  - Which elements should be considered when interpreting linear models?

- Understanding and diagnosing linear model assumptions

  - Outline the model assumptions along with tools for their diagnosis.

  - Mention a few ways to deal with the violation of assumptions.

  - Which categories of unusual observations exist?

  - What is the hat matrix and discuss the relation of hat values to outliers.

  - Outline the concept of Cook's distance.

Introduction to the linear model: simple linear regression

Contents

1. **Research goals and general concept**

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. Model assumptions and diagnostics II + wrap up

Linear model: Research goals

1. Prediction

Example: Establish linear relationship between mean plant growth and nutrient concentrations from observations that allows for prediction of mean plant growth for non-observed nutrient concentrations
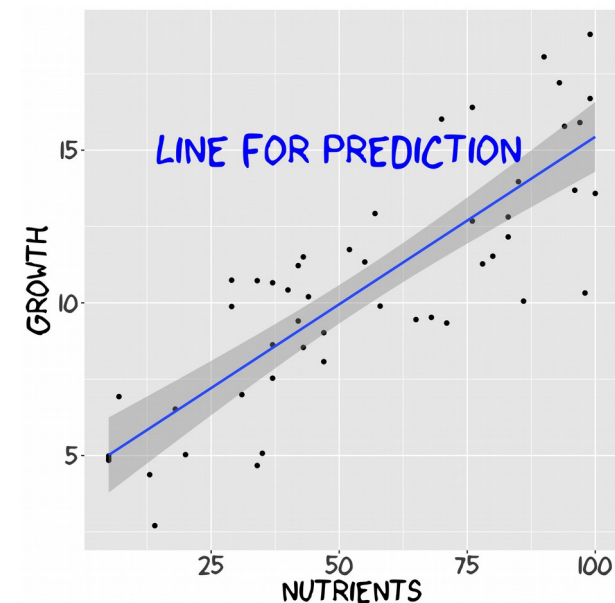
2. (Parameter) estimation

Example: Estimate slope for linear relationship between mean plant growth and nutrient concentrations → How much does growth increase per additional unit of nutrients?

3. Assessing hypotheses

Example: Assess hypotheses related to relationship between plant growth and nutrient concentrations.

4. Explanation

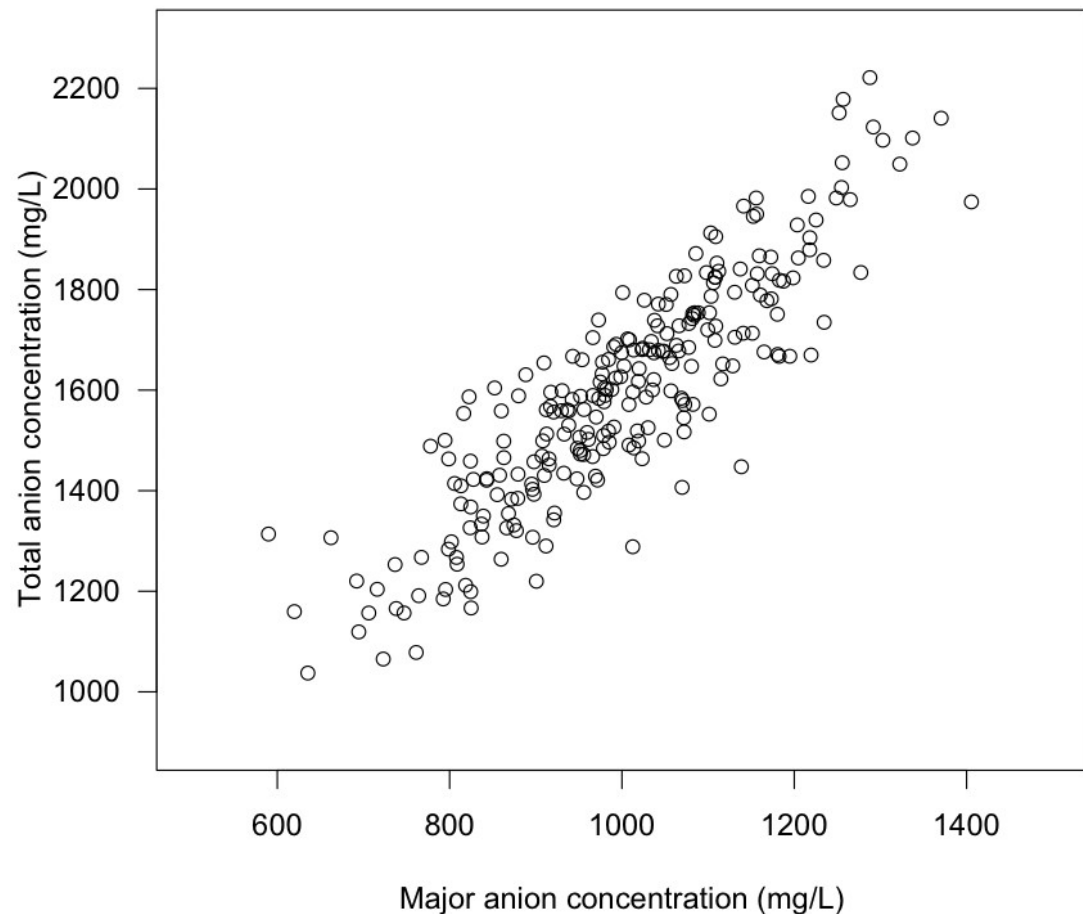Example: Use nutrient concentrations to explain mean plant growth.

Case study: Water concentrations

Research question: Can we predict the total anion concentration in water from the concentration of major anions (Cl^-, SO_4^{2-}, PO_4^{3-})?

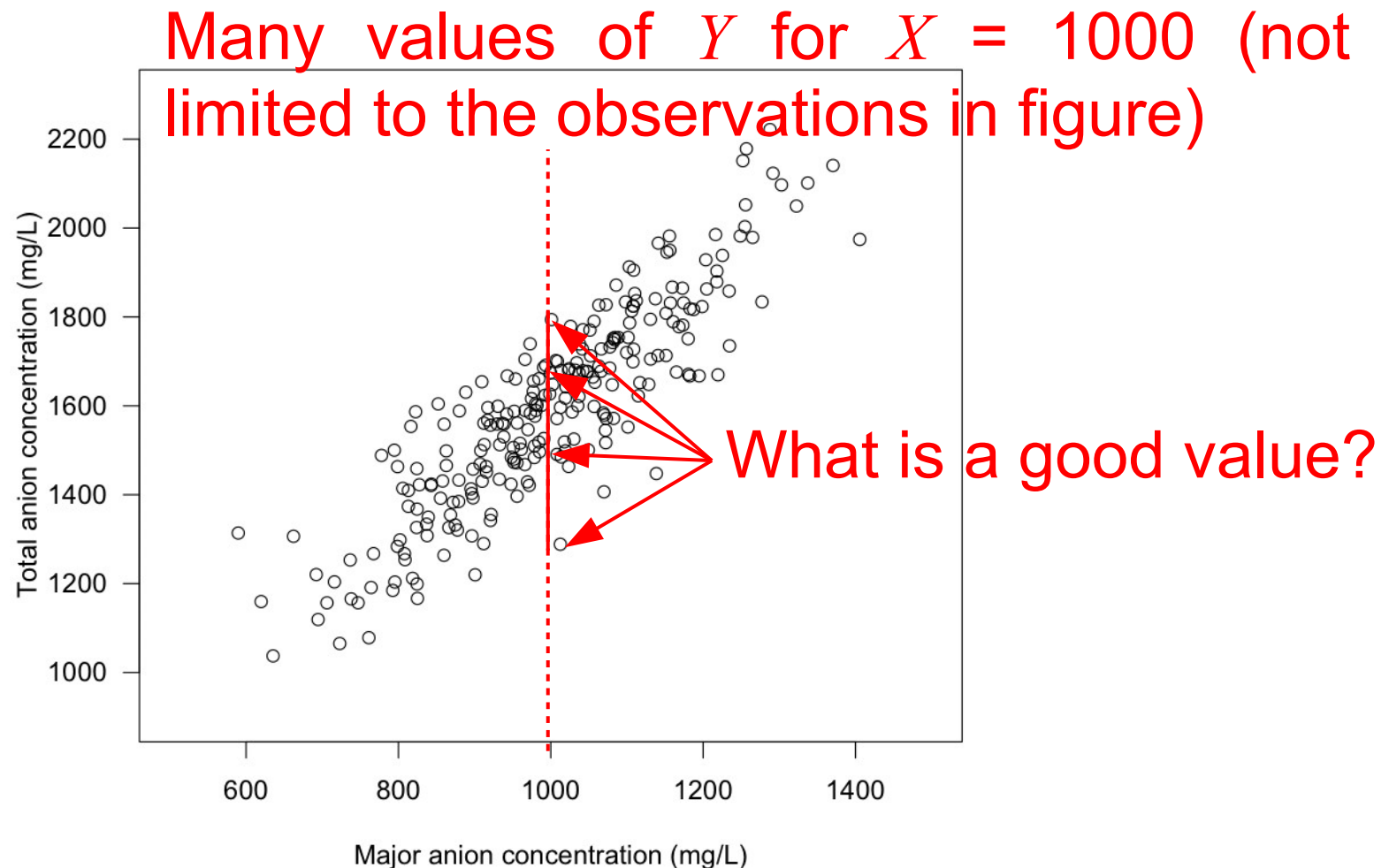Study: Samples of total anion and major anion concentrations from 250 streams.

https://upload.wikimedia.org/wikipedia/commons/1/11/Water_resources%2C_taking_a_water_sample.jpg
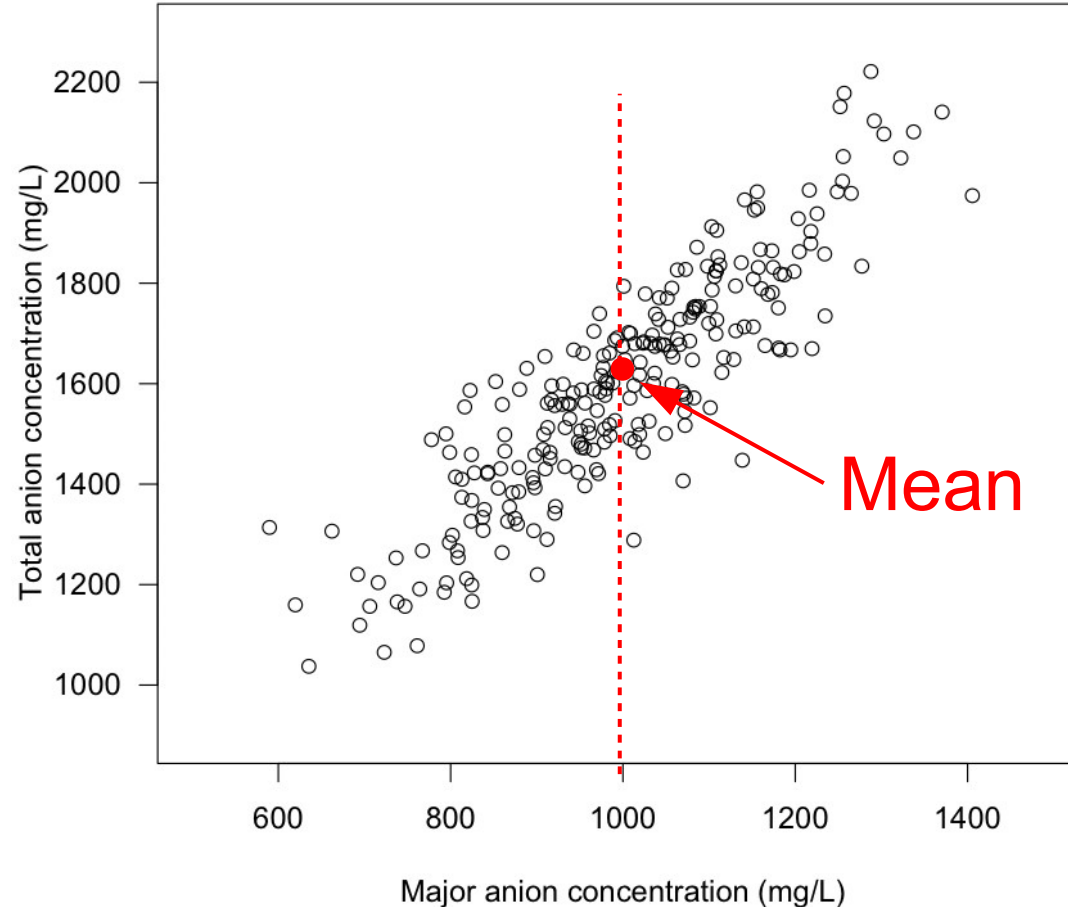
Research goal: Predicting Y from X

Y = Population of total ion conc., X = Population of major ion conc.

We define for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$, where ε represents all variables influencing Y omitted from the model → What is the optimal $f(X)$?

Many values of Y for X = 1000 (not limited to the observations in figure)
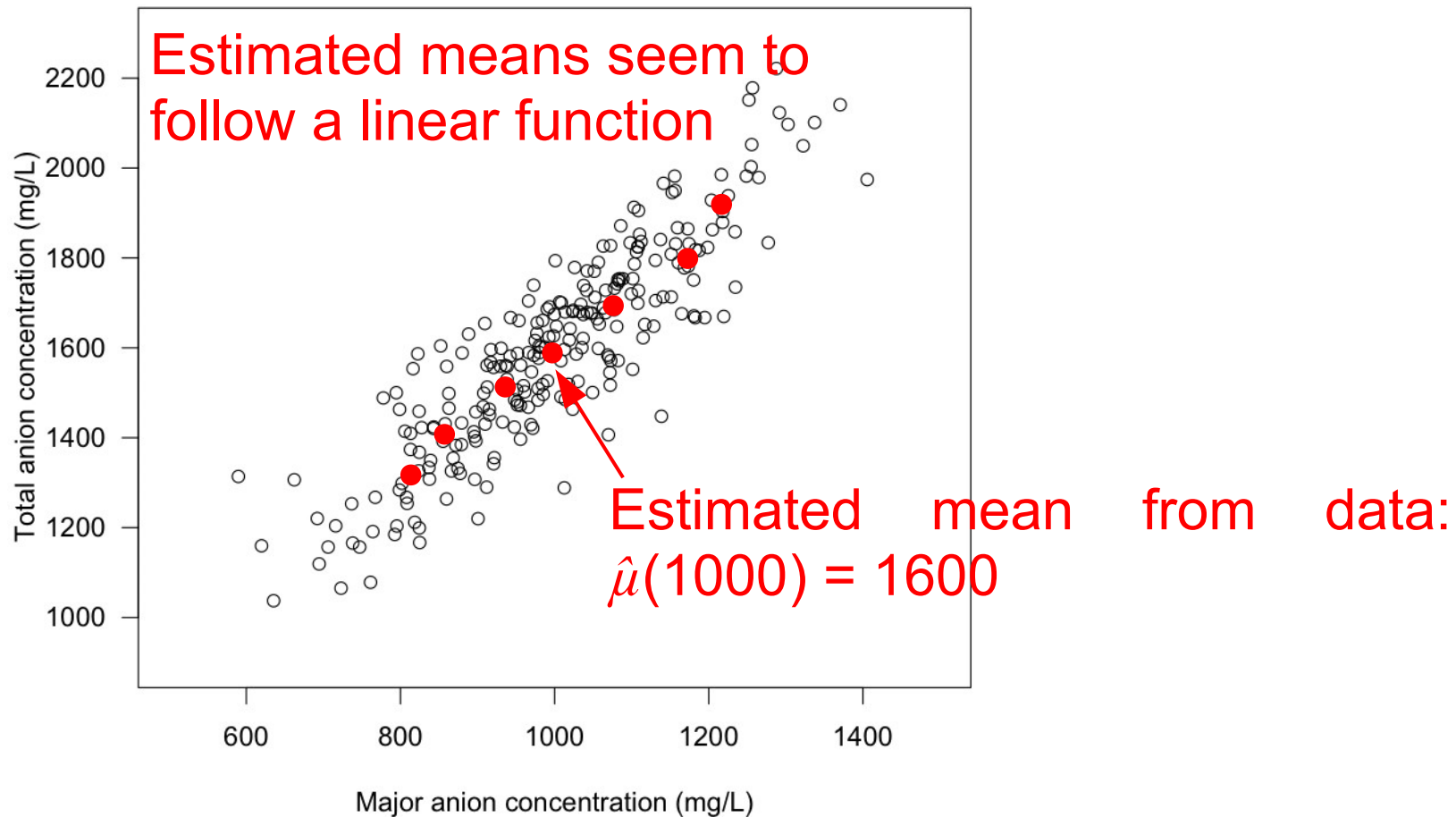
What is a good value?

Predicting Y from X: What is a good value?

A good value is the mean: $f(1000) = E(Y|X = 1000)$.

$\rightarrow$ Ideal $f(X) = E(Y|X = x) = \boxed{\mu(X)}$ *regression function*

Predicting Y from X with a linear function

Estimated means seem to follow a linear function

Estimated mean from data: $\hat{\mu}(1000) = 1600$

We assume a linear function of the true population $\mu(X)$:

$\mu(X) = \beta_0 + \beta_1 X$ from which follows that: $Y = \beta_0 + \beta_1 X + \varepsilon$

β_0 and β_1 : regression coefficients

Introduction to the linear model: simple linear regression

Contents

(Simple) Linear regression model

Assuming that the true relationship is a linear function of the form $Y = \beta_0 + \beta_1 X + \varepsilon$, we can use sample data to obtain estimates of β_0 and β_1, denoted as $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$, and subsequently predict $\hat{Y}$:

$$\hat{Y} = b_0 + b_1 X \quad \text{for realisations of } X \text{ we can rewrite this to:}$$

$$\hat{y}_i = b_0 + b_1 x_i$$

What are β_0 and β_1?

$\beta_0 = E(Y|X = 0)$ "intercept"

$\beta_1 = \dfrac{dy}{dx}$ "slope"

$b_0 = 138$

$b_1 = 1.5$

What is the optimal regression line?

We defined for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$

$$\Rightarrow Y = \hat{Y} + \varepsilon \quad \Leftrightarrow \quad \varepsilon = Y - \hat{Y}$$

For sample data ($i = 1,2,3, \ldots, n$) and the regression model, we defined: $\hat{y}_i = b_0 + b_1 x_i$

We define the residual e_i as: $e_i = y_i - \hat{y}_i = y_i - b_0 + b_1 x_i$

Example for observation $i = 12$

What is the optimal regression line?

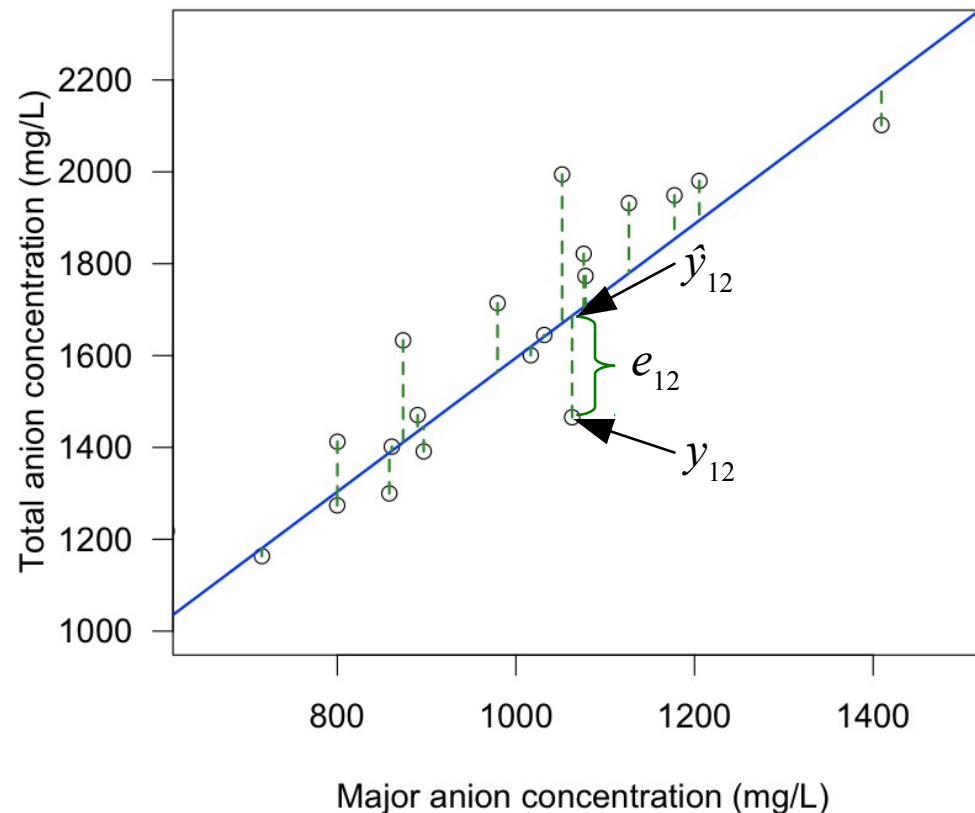Example for observation i = 12

Optimal line minimises the Residual Sum of Squares (RSS):

$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

$$= \left(y_1 - b_0 + b_1 x_1\right)^2 + \left(y_2 - b_0 + b_1 x_2\right)^2 + ... + \left(y_n - b_0 + b_1 x_n\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_i\right)^2 \Rightarrow \text{Find} \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_i\right)^2$$

Introduction to the linear model: simple linear regression

Contents

Determining the regression coefficients

Find $\arg\min\limits_{b_0, b_1} \sum\limits_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2$

It can be shown that the minimizing values are:

$$b_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Matrix notation of linear regression clarifies calculation

$\hat{y}_i = b_0 + b_1 x_i$ Notation for the observations $i = 1, 2, 3, \ldots, n$:

$\hat{y}_1 = b_0 + b_1 x_1$
$\hat{y}_2 = b_0 + b_1 x_2$
$\vdots$
$\hat{y}_n = b_0 + b_1 x_n$

matrix $\longrightarrow$

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

$$\hat{Y} = X b$$

Example calculation of coefficients

$$\hat{Y} = X\,b$$

It can be shown that:

Inverse

$$b = (X^T X)^{-1}(X^T Y)$$

Example: Calculation for trivial data

Let our set of data be $\{(10,4)(20,5)\} \Rightarrow x=\{10,\,20\}\,,\,y=\{4,\,5\}$

Matrix notation: $X=\begin{pmatrix} 1 & 10 \\ 1 & 20 \end{pmatrix},\, X^T=\begin{pmatrix} 1 & 1 \\ 10 & 20 \end{pmatrix},\, Y=\begin{pmatrix} 4 \\ 5 \end{pmatrix}$

Example calculation of coefficients

$$\hat{Y} = X b$$

It can be shown that:

$$b = (X^T X)^{-1}(X^T Y)$$

Example: Calculation for trivial data

Let our set of data be $\{(10,4)(20,5)\} \Rightarrow x=\{10, 20\}, y=\{4, 5\}$

Matrix notation: $X = \begin{pmatrix} 1 & 10 \\ 1 & 20 \end{pmatrix}, X^T = \begin{pmatrix} 1 & 1 \\ 10 & 20 \end{pmatrix}, Y = \begin{pmatrix} 4 \\ 5 \end{pmatrix}$

Calculation of b:

$$b = (X^T X)^{-1}(X^T Y) = \left(\begin{pmatrix} 1 & 1 \\ 10 & 20 \end{pmatrix}\begin{pmatrix} 1 & 10 \\ 1 & 20 \end{pmatrix}\right)^{-1}\left(\begin{pmatrix} 1 & 1 \\ 10 & 20 \end{pmatrix}\begin{pmatrix} 4 \\ 5 \end{pmatrix}\right)$$

$$= \begin{pmatrix} 1\cdot1+1\cdot1 & 1\cdot10+1\cdot20 \\ 10\cdot1+20\cdot1 & 10\cdot10+20\cdot20 \end{pmatrix}^{-1}\begin{pmatrix} 1\cdot4+1\cdot5 \\ 10\cdot4+20\cdot5 \end{pmatrix} = \begin{pmatrix} 2 & 30 \\ 30 & 500 \end{pmatrix}^{-1}\begin{pmatrix} 9 \\ 140 \end{pmatrix}$$

Example calculation of coefficients

Calculation of the inverse For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the inverse is $A^{-1} = \dfrac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$$b = \begin{pmatrix} 2 & 30 \\ 30 & 500 \end{pmatrix}^{-1} \begin{pmatrix} 9 \\ 140 \end{pmatrix} = \left(\frac{1}{2 \cdot 500 - 30 \cdot 30} \begin{pmatrix} 500 & -30 \\ -30 & 2 \end{pmatrix} \right) \begin{pmatrix} 9 \\ 140 \end{pmatrix}$$

$$= \left(\frac{1}{100} \begin{pmatrix} 500 & -30 \\ -30 & 2 \end{pmatrix} \right) \begin{pmatrix} 9 \\ 140 \end{pmatrix} = \begin{pmatrix} 5 & -0.3 \\ -0.3 & 0.02 \end{pmatrix} \begin{pmatrix} 9 \\ 140 \end{pmatrix}$$

$$= \begin{pmatrix} 5 \cdot 9 - 0.3 \cdot 140 \\ -0.3 \cdot 9 + 0.02 \cdot 140 \end{pmatrix} = \begin{pmatrix} 45 - 42 \\ -2.7 + 2.8 \end{pmatrix} = \begin{pmatrix} 3 \\ 0.1 \end{pmatrix}$$

$$\Rightarrow b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 3 \\ 0.1 \end{pmatrix}$$ The intercept $b_0 = 3$ and the slope $b_1 = 0.1$

20

Confirmation in R and accuracy of results

Calculation in R

Although the matrix calculation may seem complicated, you are now able to conduct a regression analysis during a power outage. Of course, there are easier ways. R can do the calculation within the split of a second and confirms our result:

```
x <- c(10,20)
y <- c(4,5)
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)                x
##         3.0              0.1
```

But how good is the model?

Depends on what we mean by "good", we can evaluate different aspects, e.g.:

- Accuracy of the estimated regression coefficients

- Accuracy of predictions

- Overall model (as compared to other models)

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. **Accuracy of regression coefficients and predictions**

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. Model assumptions and diagnostics II + wrap up

How accurate are the estimates for the regression coefficients?

Recall that the standard error informs on the error of a parameter estimate (e.g. mean of X):

$$\text{SE}_{\overline{X}} \approx \frac{s}{\sqrt{n}} \quad \text{where} \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}$$

Also recall the concept of RSS :

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

Average squared residual: Mean squared error

$$\text{MSE} = \frac{1}{\text{DoF}} \text{RSS} \quad \text{where} \, \text{DoF} = \text{Degrees of freedom}$$

DoF for regression model: $n - p - 1$ where $p = $ no of parameters in model excluding the intercept

How accurate are the estimates for the regression coefficients?

Now that we have introduced MSE, we provide the standard errors for the regression coefficients:

$$SE_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad SE_{b_0} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Standard errors of regression coefficients increase with residuals and decrease with a wider range of x values (and sample size).

How accurate are the predictions?

We have two standard errors for predictions: The standard error for the predicted mean $\hat{y}$ and for a new observation y_{new}.

$$SE_{\hat{y}_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad \text{SE of fit}$$

$$SE_{y_{new}} = \sqrt{MSE + MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad \begin{array}{l}\text{SE of} \\ \text{prediction}\end{array}$$

x_h is the value of the predictor. Both SEs are lowest at the mean of $x \to$ accuracy decreases from centre

Agrees with intuition that the predicted (fitted) mean of y, i.e. $\hat{y}$, or a new y is most accurate in the centre given the many values on both sides that support the prediction.

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. **Confidence intervals – intro and application**

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. Model assumptions and diagnostics II + wrap up

Confidence intervals

For graphical illustration, we need to understand the concept of the confidence interval.

Confidence interval (CI):

- core concept of frequentist statistics

- interval estimate for parameter (e.g. mean) from statistical population

- confidence level (typically 95%) defines frequency the interval contains the true parameter in (hypothetical) repeated studies. I.e. 95% of CIs constructed from repeated samples, contain true parameter.

$\rightarrow$ Probabilistic statement about performance of procedure deriving interval, and not directly about a specific CI

Visualisation of CIs: CI for mean

95% CIs for mean of quail sperms

https://upload.wikimedia.org/wikipedia/commons/d/db/Japanese_Quail.jpg

These intervals don't contain the 'true' value of the mean

Field, Miles & Field 2013: 46
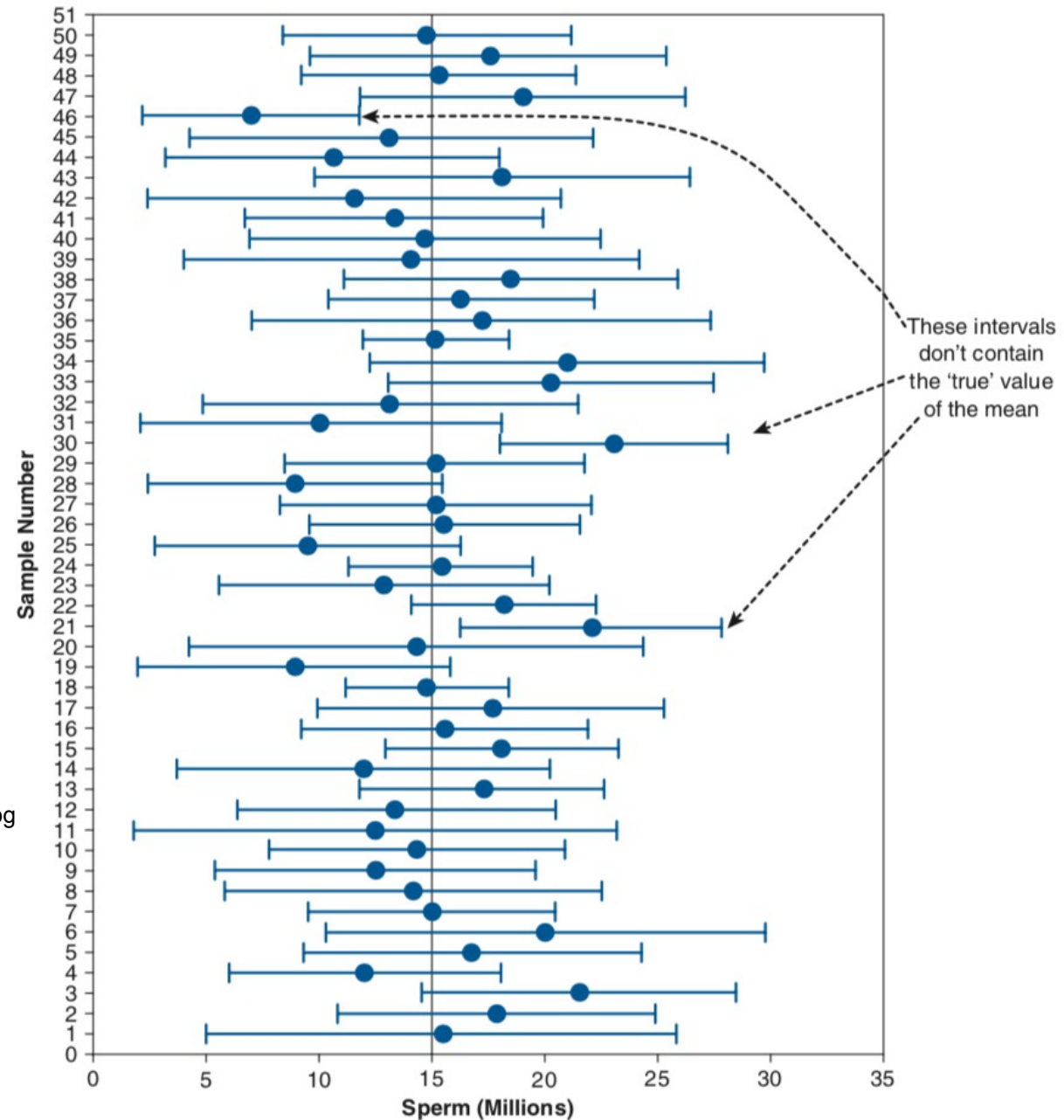
Visualisation of CIs: Regression bands

95% CI for mean S, i.e. $E(S|K = k)$, (blue) and
95% prediction interval for S, i.e. $S(K = k)$, (red)

Example: Study on soil ion
concentrations
S = Sulphur
K = Potassium

Calculation and interpretation of CIs

General calculation of confidence interval for parameter θ:

$\theta \pm$ value related to confidence level and probability distribution $\times$ SE

(e.g. 95%, 99%)

Known or assumed probability distribution of parameter

Example: Values (= z-scores) for normal probability distribution

The Normal Distribution

Probability

Values

68 % CI

95 % CI
95% of values

99 % CI
99% of values

-1.96σ   1.96σ

-2.58σ   2.58σ

Probability of Cases in portions of the curve	≈0.0013	≈0.0214	≈0.1359	≈0.3413	≈0.3413	≈0.1359	≈0.0214	≈0.0013

Standard Deviations From The Mean	-4σ	-3σ	-2σ	-1σ	0	+1σ	+2σ	+3σ	+4σ
Cumulative %		0.1%	2.3%	15.9%	50%	84.1%	97.7%	99.9%	
Z Scores	-4.0	-3.0	-2.0	-1.0	0	+1.0	+2.0	+3.0	+4.0

31

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. **Accuracy of overall model and sum of squares**

7. Model assumptions and diagnostics I

8. Model assumptions and diagnostics II + wrap up

How accurate is the overall model?

Concept of Root mean square error (RMSE), allows for comparison across models:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{\text{DoF}}\text{RSS}} = \sqrt{\frac{1}{\text{DoF}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Concept of explained variance, allows for comparison across linear models:

$$R^2 = r^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad \text{where TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Pearson correlation coefficient

Total sum of squares

What are these different sum of squares?

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

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
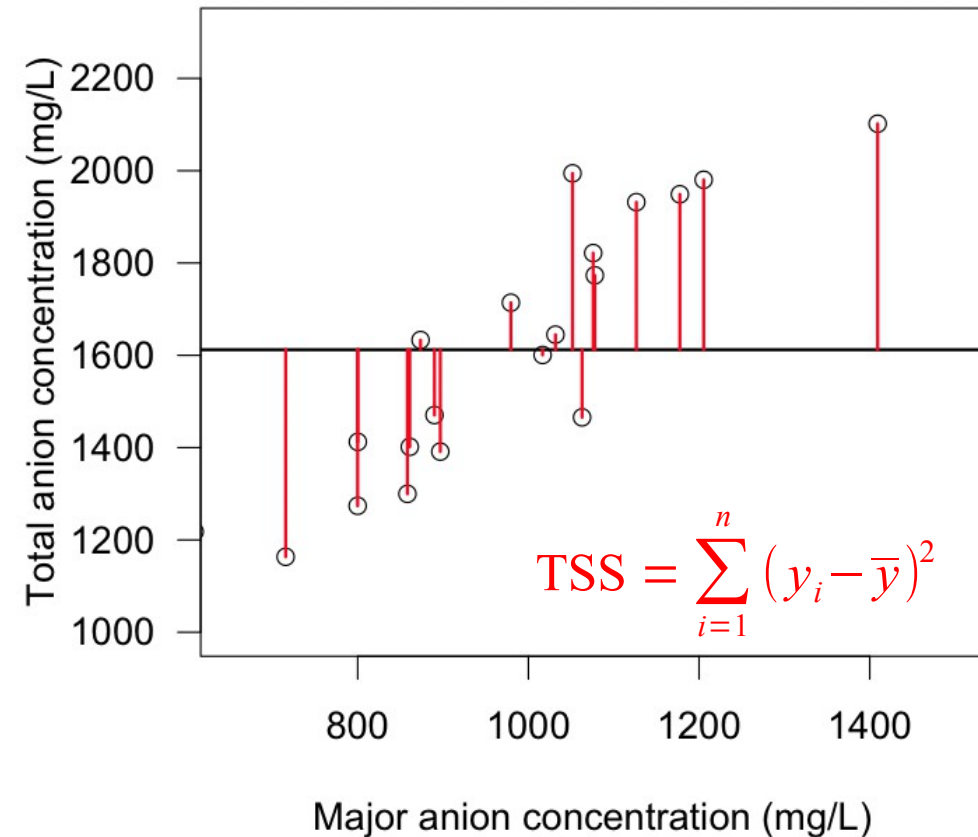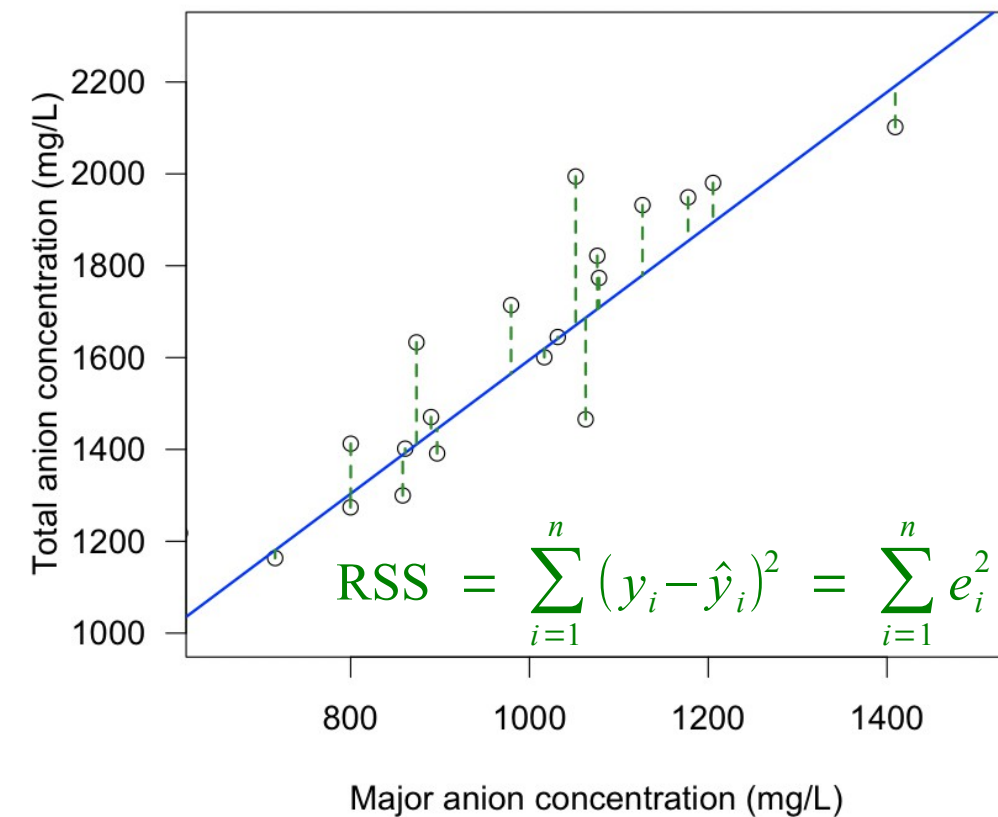
TSS (total Var) = MSS (explained Var) + RSS (unexplained Var)

What are these different sum of squares?
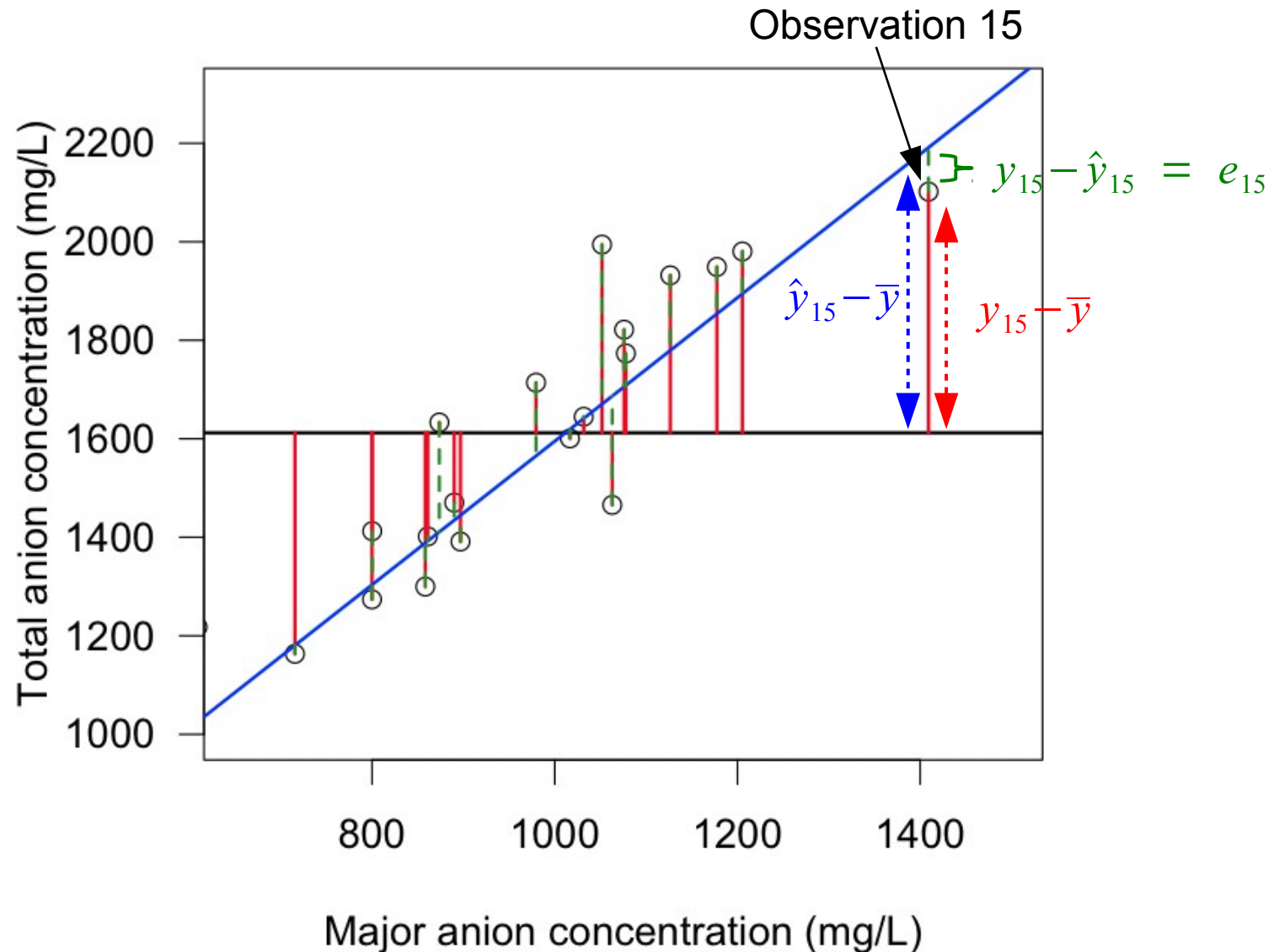
Observation 15

$y_{15} - \hat{y}_{15} = e_{15}$

$\hat{y}_{15} - \overline{y}$

$y_{15} - \overline{y}$

TSS (total Var) = MSS (explained Var) + RSS (unexplained Var)

R output for linear regression model

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -352.96  -71.49   -2.75   69.60  323.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.79468   31.92061   4.317 1.95e-05 ***
## X             1.45722    0.03152  46.231  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103 on 448 degrees of freedom
## Multiple R-squared:  0.8267,  Adjusted R-squared:  0.8263
## F-statistic:  2137 on 1 and 448 DF,  p-value: < 2.2e-16
```

Residuals: $y_i - \hat{y}_i = e_i$

b_0 SE_{b_0}

b_1 SE_{b_1}

?

RMSE

R^2

Never interpret a linear model without a visual representation

We have 4 regression models with the same regression coefficients and standard errors:

b_0 SE_{b_0}

b_1 SE_{b_1}

```
## $lm1
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0000909   1.1247468 2.667348 0.025734051
## x1          0.5000909   0.1179055 4.241455 0.002169629
##
## $lm2
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.000909   1.1253024 2.666758 0.025758941
## x2          0.500000   0.1179637 4.238590 0.002178816
##
## $lm3
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0024545   1.1244812 2.670080 0.025619109
## x3          0.4997273   0.1178777 4.239372 0.002176305
##
## $lm4
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0017273   1.1239211 2.670763 0.025590425
## x4          0.4999091   0.1178189 4.243028 0.002164602
```
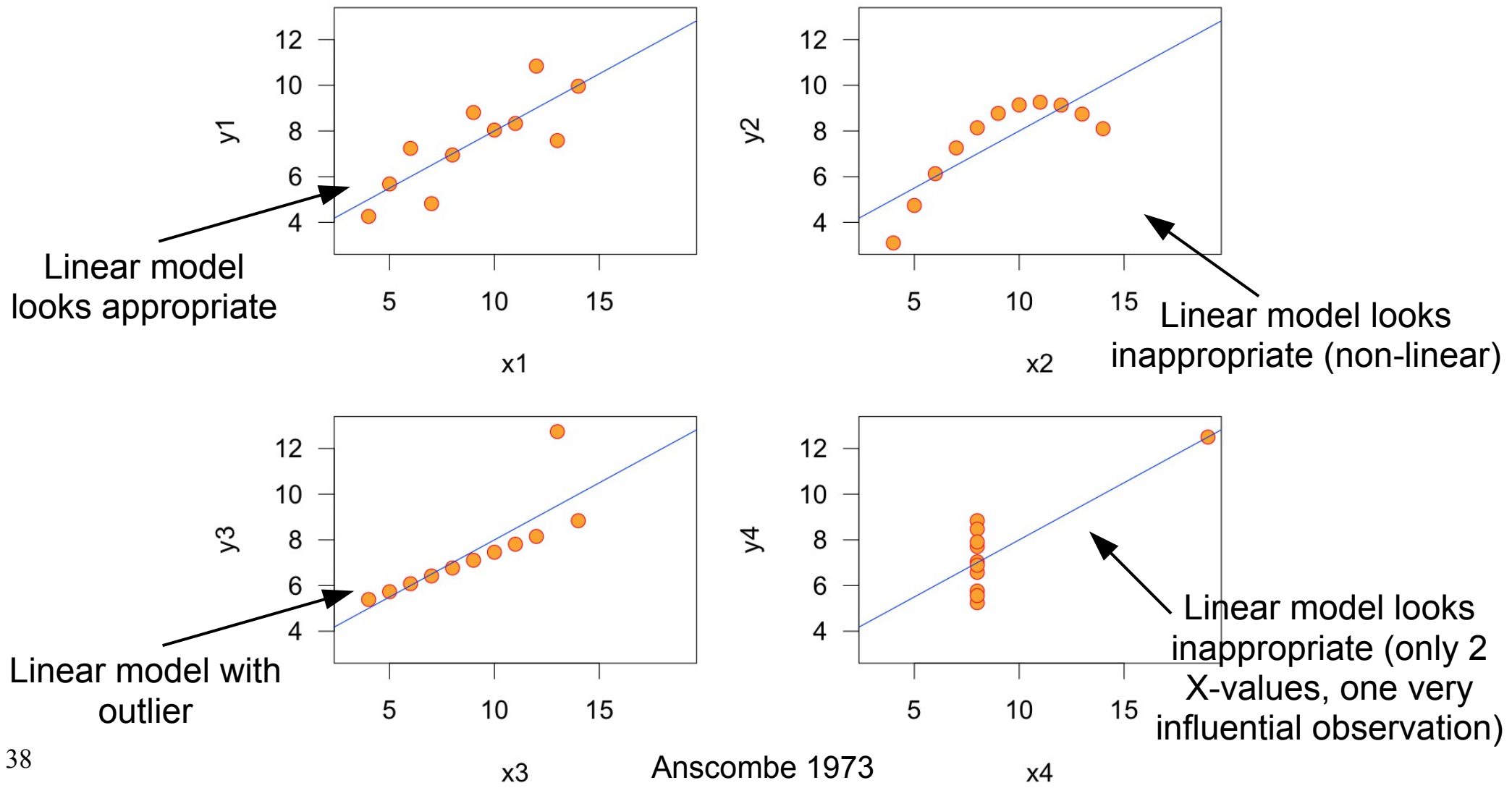
Model 1

Model 2

Model 3

Model 4

Never interpret a linear model without a visual representation

We have 4 regression models with the same regression coefficients and standard errors, but the data differ strongly:

Linear model looks appropriate

Linear model looks inappropriate (non-linear)

Linear model with outlier

Linear model looks inappropriate (only 2 X-values, one very influential observation)

38

Anscombe 1973

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. **Model assumptions and diagnostics I**

8. Model assumptions and diagnostics II + wrap up

Model assumptions

Recap of model and derivation of model assumptions:

Classical model definition

Probability distribution-centric model definition

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$
$$\varepsilon \sim \text{Normal}(0, \sigma)$$

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 x_i$$
$$\varepsilon = y_i - \mu_i$$

$$\beta_0 + \beta_1 x_i$$
$$=$$
$$\mu_i \quad \sigma$$
normal
$$\sim$$
$$y_i$$

- Assumptions:
  - Linear relationship
  - Normal distribution of error
  - Variance homogeneity (Homoscedasticity)
  - Independence of errors

Visualisation of some model assumptions
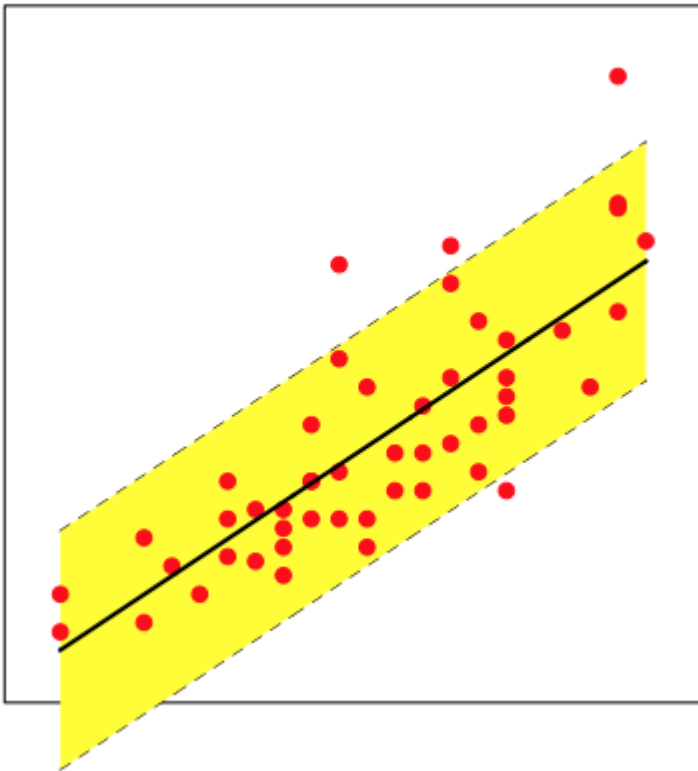
Linear relationship

Normal distribution and constant variance

http://freakonometrics.hypotheses.org/tag/poisson

Diagnostics for model assumptions of the simple linear regression model

- Assumptions:

  - Linear relationship: $\mu_i = \beta_0 + \beta_1 x_i$ $\rightarrow$ checking via model visualisation and residuals *vs.* fitted values plots

  - Normal distribution of error $\rightarrow$ checking via QQ-plots

  - Variance homogeneity: $\mathrm{Var}(Y|\,X = x)$ is constant in $x \rightarrow$ checking via residuals *vs.* fitted values plots

  - Independence of errors $\rightarrow$ checking via serial correlation plots

Model diagnostics: Variance homogeneity
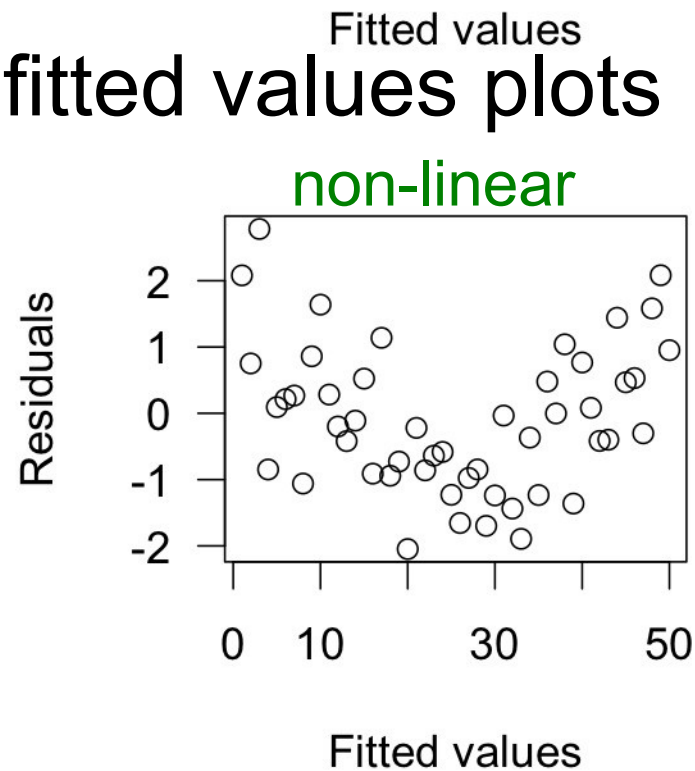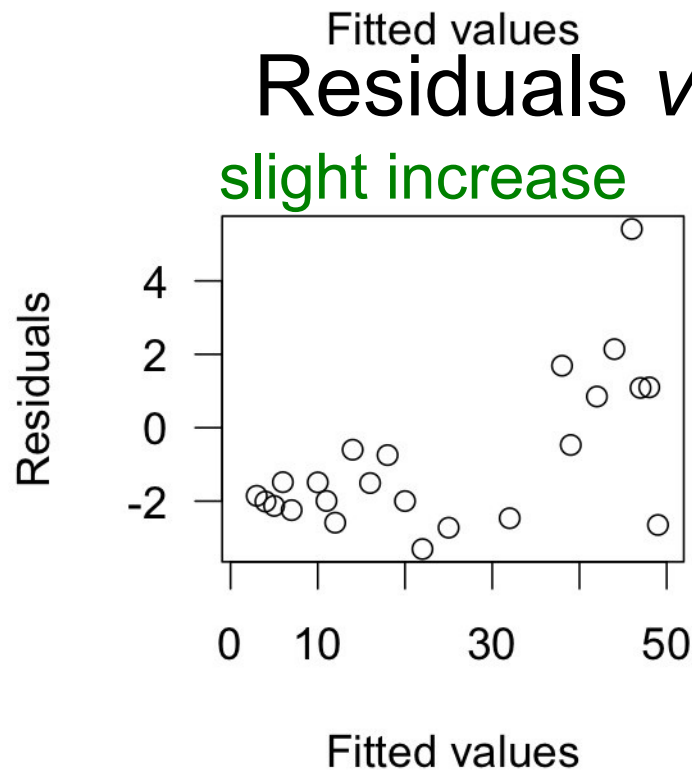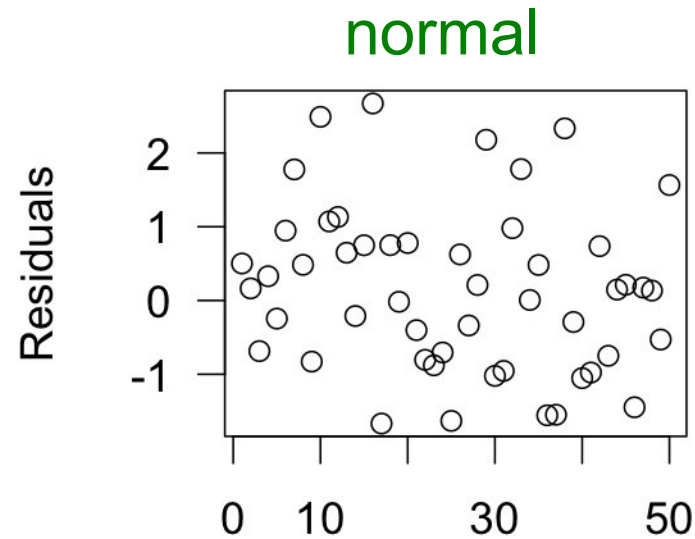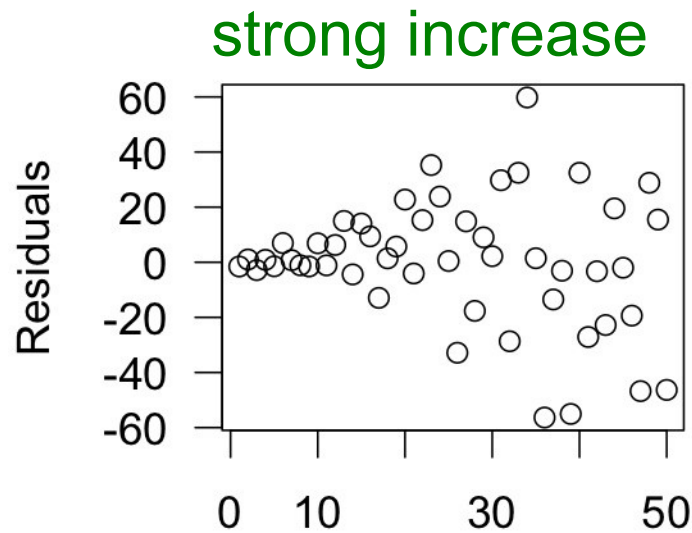
Residuals *vs.* fitted values plots

How to deal with violation of model assumptions

- Violation of assumptions can invalidate specific or all regression estimates → results may be unreliable

- Often multiple assumptions violated simultaneously (e.g. non-linear relationship can cause non-constant error variance)

- Check whether important variables are missing from model → can lead to violation of assumptions (e.g. serial correlation, non-linear relationship)

- Data is dependent, shows serial correlation:

  - Aggregate data to achieve independence

  - Model error structure (e.g. Generalised least squares)

  - Use different model (e.g. Linear mixed effect model)

How to deal with violation of model assumptions

- Relationship is non-linear, discernible in model plot or residuals *vs.* fitted values plot:

    - Use different model (e.g. Generalised linear model, Generalised additive model, Random forest)

    - Transform variable(s)

- Variance is not constant (i.e. is heteroscedastic), discernible in residuals *vs.* fitted values plot:

    - Correct standard errors (e.g. Heteroscedasticity-consistent standard errors)

    - Use different model (e.g. Generalised least squares, Generalised linear model)

    - Transform variable(s)

How to deal with violation of model assumptions

- Error not normally distributed, discernible in QQ-plot with residuals:

  - Least important assumption, model relatively robust towards violation

  - Violation often associated with violation of other assumptions → curing other violations often achieves normality

  - If any other assumption is violated, a non-normal error distribution is particularly problematic for heavy-tailed or strongly skewed error distributions. In this case:

    - Use robust linear regression method (e.g. median regression)
    - Use different model (e.g. Generalised linear model)
    - Transform variable(s)

Be careful with transformations!

Change model (e.g. use a GLM) if other model fits better in terms of distribution or shape of relationship

Ieno & Zuur 2015: 36

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. **Model assumptions and diagnostics II + wrap up**

Further model diagnostics: Outliers

Some terminology:

- Outlier: Unusual observation in $Y|X$ deviating from model

- Leverage point: Unusual observation regarding distribution of X

- Influential point: Observation influencing model fit (e.g. coefficients)

(a) Outlier, no leverage or influential point

(b) Leverage point and influential point, no outlier

(c) Leverage point, no outlier or influential point

Model without black observation

Model with black observation

Fox 2015: 268

Diagnosing leverage: Hat values

Recap

Linear model in matrix form: $\qquad \hat{Y} = X b$

Estimate regression coefficients via: $\quad b = (X^T X)^{-1}(X^T Y)$

Substitution of b by $X^{-1}\hat{Y}$: $\qquad \hat{Y} = \underbrace{X(X^T X)^{-1}(X^T}Y)$

Hat matrix H (named for "putting a hat on Y")

So-called *hat values* h_i (calculated from H) summarise influence of Y_i on <u>all</u> fitted $\hat{Y} \rightarrow$ higher h, higher influence on fitted Y (but not necessarily influential point)

h measures distance to mean of X

Diagnosing leverage: Hat values
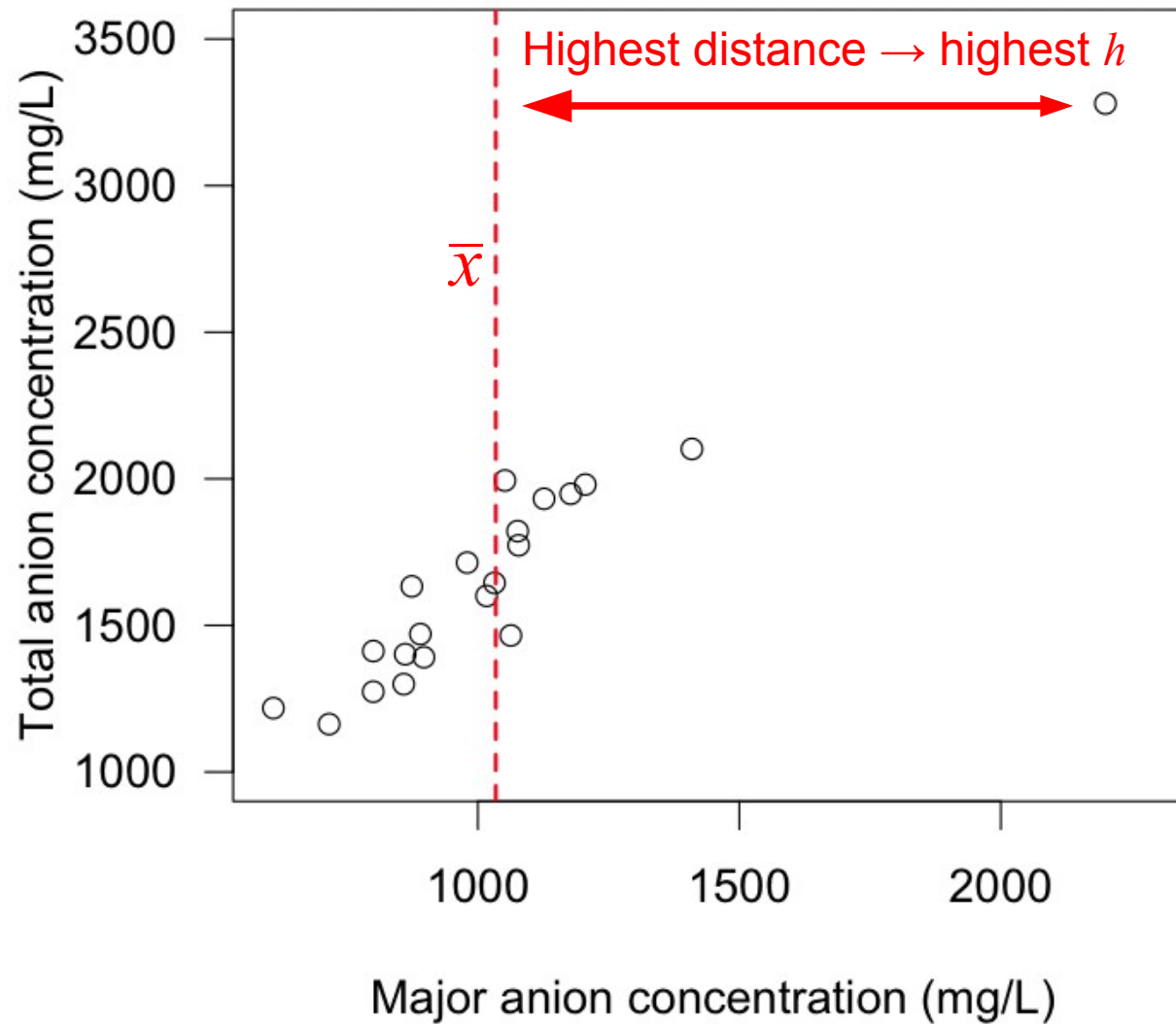
Diagnosing leverage: Hat values

High h relative to other observations $\rightarrow$ Leverage point
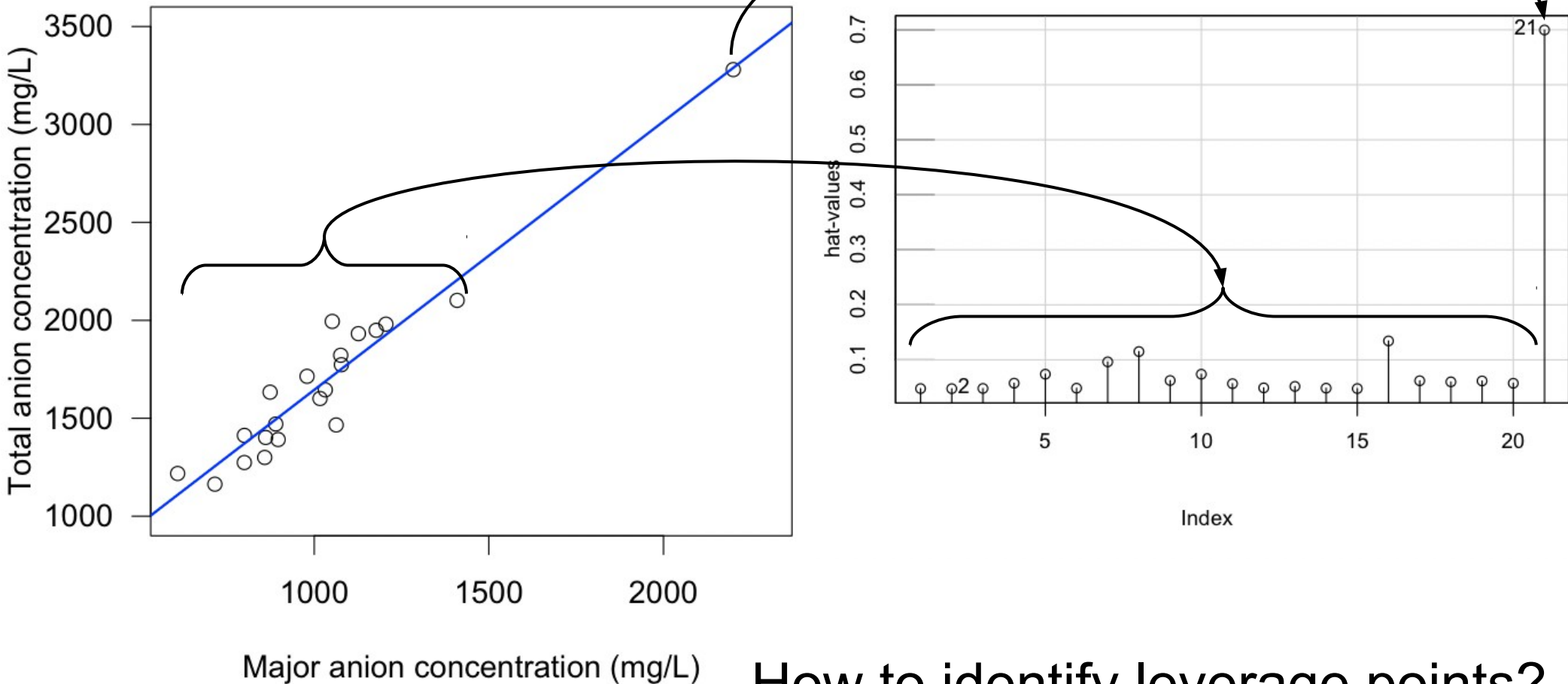
How to identify leverage points?

$$\bar{h} = \frac{p}{n} \qquad \text{Check observations with } h > 2\frac{p}{n}$$

Diagnosing outliers and influential points

Outliers:

- can be identified via residuals $e \rightarrow$ higher e, higher deviation from model

- Standardisation of residuals by their standard error and by h, because they are influenced by h:

$$\text{Standardised residual } r_i \;=\; \frac{e_i}{\sqrt{\text{MSE}(1-h_i)}}$$

Influential points:

- Intuition: Outliers with high leverage $\rightarrow$ Cook's distance D combines standardised residuals with hat values:

$$D_i \;=\; \frac{r_i}{p}\,\frac{h_i}{(1-h_i)}$$

How to deal with unusual observations?

- Unusual observations are model-dependent: may disappear if model specification (e.g. variables included), type of model (e.g. GLM instead of linear model) or variables (e.g. transformation) change

- Check whether values are plausible

- Unusual observations that are not influential points → not much to worry

- Check robustness of model results when removing influential observation → report (and plot) both results

- If other model assumptions are met, but model results respond strongly to removal of influential observation:

  - Use robust or quantile regression model
  - Transform data

Flowchart for simple linear regression

Fit a model based on subject matter expertise and/or observation of the scatter plot

Assess the adequacy of the model in particular:
Is the functional form of the model correct?
Do the errors have constant variance?

YES — Do outliers and/or leverage points exist?

NO — Add new terms to the model and/or transform x and/or Y

NO — Is the sample size large?

YES — Are the outliers and leverage points valid?

YES — Based on Analysis of Variance or a t-test decide if there is a significant association between Y and x?

NO — Are the errors normally distributed?

NO — Remove them and refit the model

YES — Are the errors normally distributed?

YES — Based on Analysis of Variance or a t-test...

NO — Use the bootstrap for inference

YES — Remove them and refit the model

Consider modifications to the model

YES — Use the final model to form confidence intervals

NO — Stop!

56

Introduction to the linear model: simple linear regression

Ralf B. Schäfer

These slides and notes complement the lecture with exercises "Tools for complex data analysis" for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

- Understanding the research goals and concept of the linear model

- Knowledge on the calculation of the model and of accuracy measures

- Understanding confidence intervals in general and in the regression context

- Ability to assess overall model accuracy and interpret model output

- Understanding and diagnosing model assumptions

2

Learning targets and study questions

- Understanding the research goals and concept of the linear model

    - Give an example of a study question for each research goal that can be tackled with a linear regression model.

    - What does the linear regression model predict? Explain the assumption of linearity in this context.

    - Define the residual and explain its role in finding optimal regression coefficients.

- Knowledge on the calculation of the model and of accuracy measures

    - Explain how b_1 relates to the joint variance of x and y.

    - Explain the concept of the standard error regarding accuracy.

    - What is the MSE?

    - How do the formulas inform a) that the SE of regression coefficients decreases with a wider range of x values and b) that the SE of fit decreases towards the centre of data?

Learning targets and study questions

- Understanding confidence intervals (CIs) in general and in the regression context

    - Explain the correct interpretation of CIs and two misleading interpretations.

    - Describe the change of the CI with the confidence level.

- Ability to assess overall model accuracy and interpret model output

    - Define R^2 and outline its calculation.

    - Which elements should be considered when interpreting linear models?

- Understanding and diagnosing linear model assumptions

    - Outline the model assumptions along with tools for their diagnosis.

    - Mention a few ways to deal with the violation of assumptions.

    - Which categories of unusual observations exist?

    - What is the hat matrix and discuss the relation of hat values to outliers.

    - Outline the concept of Cook's distance.

Introduction to the linear model: simple linear regression

Contents

1. **Research goals and general concept**
2. Simple linear regression model
3. Calculation of regression coefficients
4. Accuracy of regression coefficients and predictions
5. Confidence intervals – intro and application
6. Accuracy of overall model and sum of squares
7. Model assumptions and diagnostics I
8. Model assumptions and diagnostics II + wrap up

5

Linear model: Research goals

1. Prediction

Example: Establish linear relationship between mean plant growth and nutrient concentrations from observations that allows for prediction of mean plant growth for non-observed nutrient concentrations
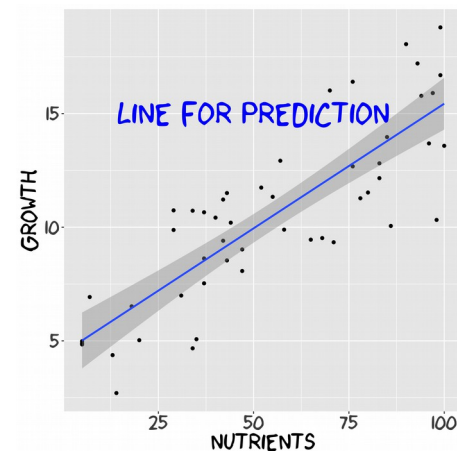
2. (Parameter) estimation

Example: Estimate slope for linear relationship between mean plant growth and nutrient concentrations → How much does growth increase per additional unit of nutrients?

3. Assessing hypotheses

Example: Assess hypotheses related to relationship between plant growth and nutrient concentrations.

4. Explanation

Example: Use nutrient concentrations to explain mean plant growth.

Hypotheses related to the relationship between plant growth and nutrient concentrations could for example be: "There is no relationship between between plant growth and nutrient concentrations." or "An increase of 1 unit of nutrients results in an increase in growth of 1 unit."

Research goals related to explanation are typically more relevant in situations with multiple explanatory variables, i.e. multiple linear regression, which we will discuss later.

For exploration, using correlations is typically more relevant than linear regression analysis.
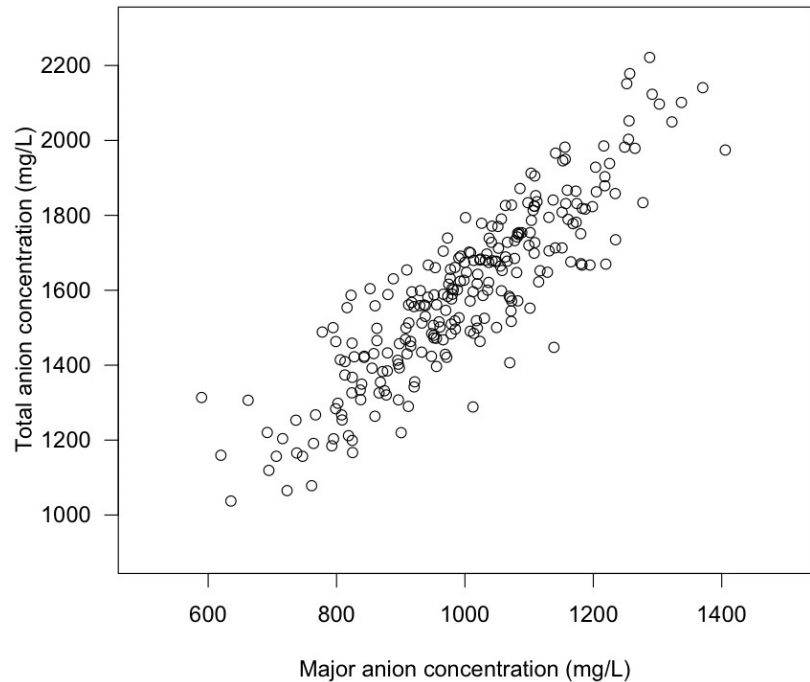
Case study: Water concentrations

<u>Research question:</u> Can we predict the total anion concentration in water from the concentration of major anions (Cl^-, SO_4^{2-}, PO_4^{3-})?

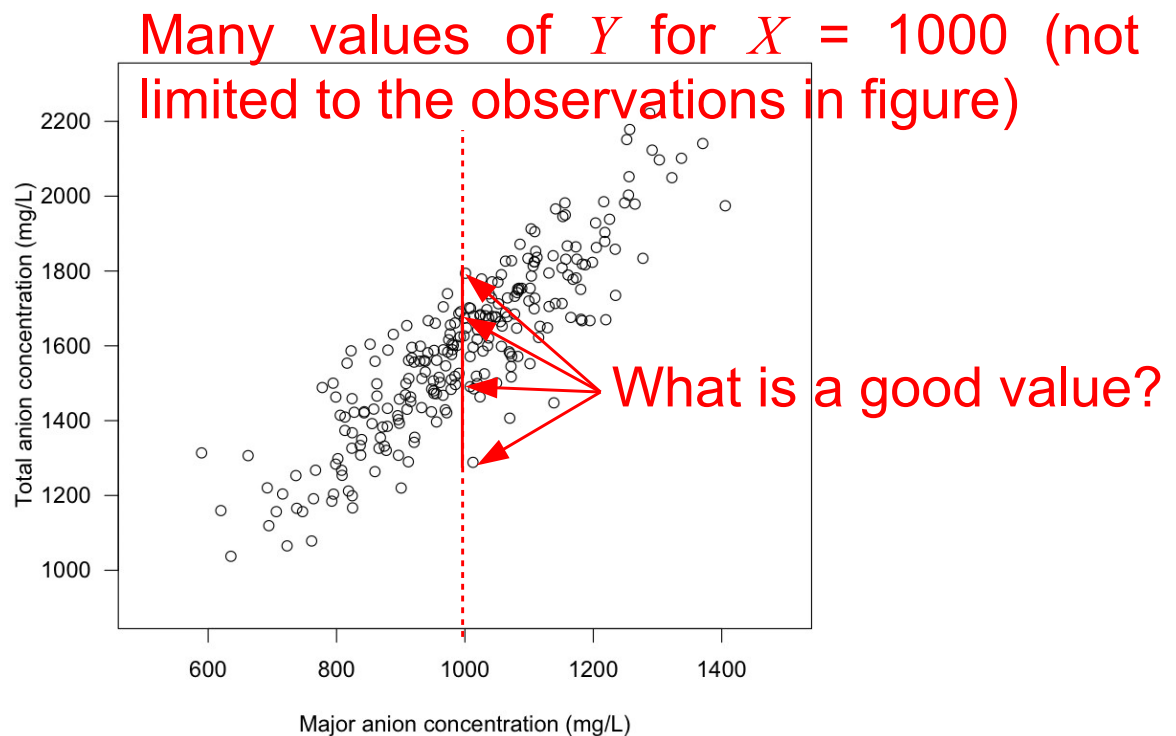<u>Study:</u> Samples of total anion and major anion concentrations from 250 streams.

https://upload.wikimedia.org/wikipedia/commons/1/11/Water_re
sources%2C_taking_a_water_sample.jpg

Research goal: Predicting Y from X

Y = Population of total ion conc., X = Population of major ion conc.

We define for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$, where ε represents all variables influencing Y omitted from the model → What is the optimal $f(X)$?

Many values of Y for X = 1000 (not limited to the observations in figure)
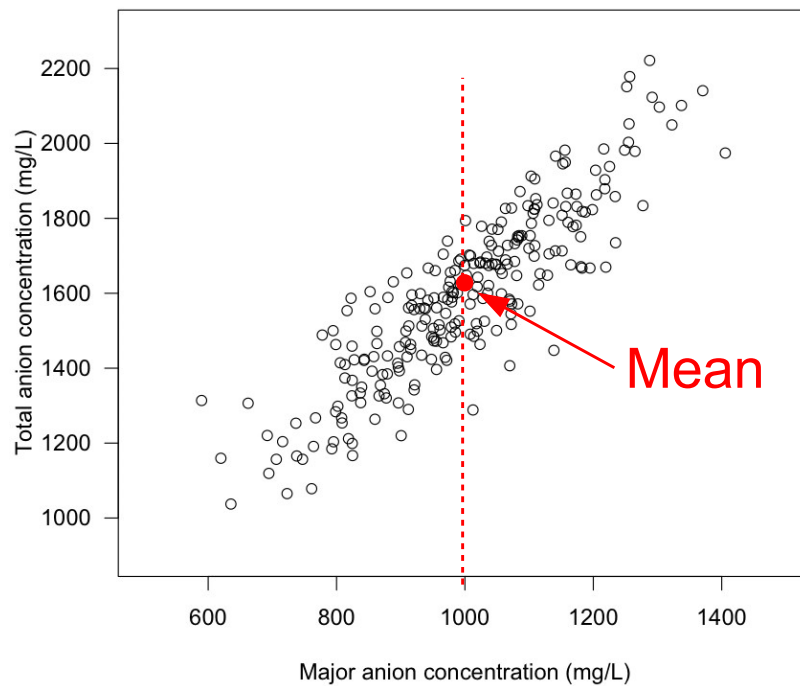
What is a good value?

Note that we refer here to the statistical populations of Y and X. $\hat{Y}$ is the prediction of the random variable Y using all population values. It is clear that with the function $f(X)$ we can only approximate Y (i.e. we have to accept that we make an error), given that we omit variables influencing Y from the model. This is always the case unless Y is a direct function of X. From a statistical perspective, for each $X = x$, we have a distribution of Y (for X = 1000 (dashed line) the range of the distribution is given by the solid red line). The observations in the figure represent only a few realisations of Y. Thus, we make an error when we predict a value for the distribution. The variable ε captures the difference between the prediction $\hat{Y}$ and Y and is called the statistical error.

In linear regression analysis, we usually do not take the measurement error in x into account. This is discussed in detail in Warton et al. (2006). They also provide information on alternatives to linear regression that should be used if the measurement error is relevant (and known).

Warton D.I., Wright I.J., Falster D.S. & Westoby M. (2006) Bivariate line-fitting methods for allometry. *Biological Reviews* 81, 259–291.

Predicting Y from X: What is a good value?

A good value is the mean: $f(1000) = E(Y|X = 1000)$.

$\rightarrow$ Ideal $f(X) = E(Y|X = x) = \boxed{\mu(X)}$ *regression function*

The regression function minimizes the Mean Squared Prediction error (MSPE): $E[(Y - f(X))^2]$, a measure that informs on the accuracy of a predictor (i.e. of all possible functions $f()$). For the mathematical proof see Matloff (2017: 49-50). Again note that our notation refers to the population, i.e. the MSPE relates to all pairs of X and Y.

Predicting Y from X with a linear function
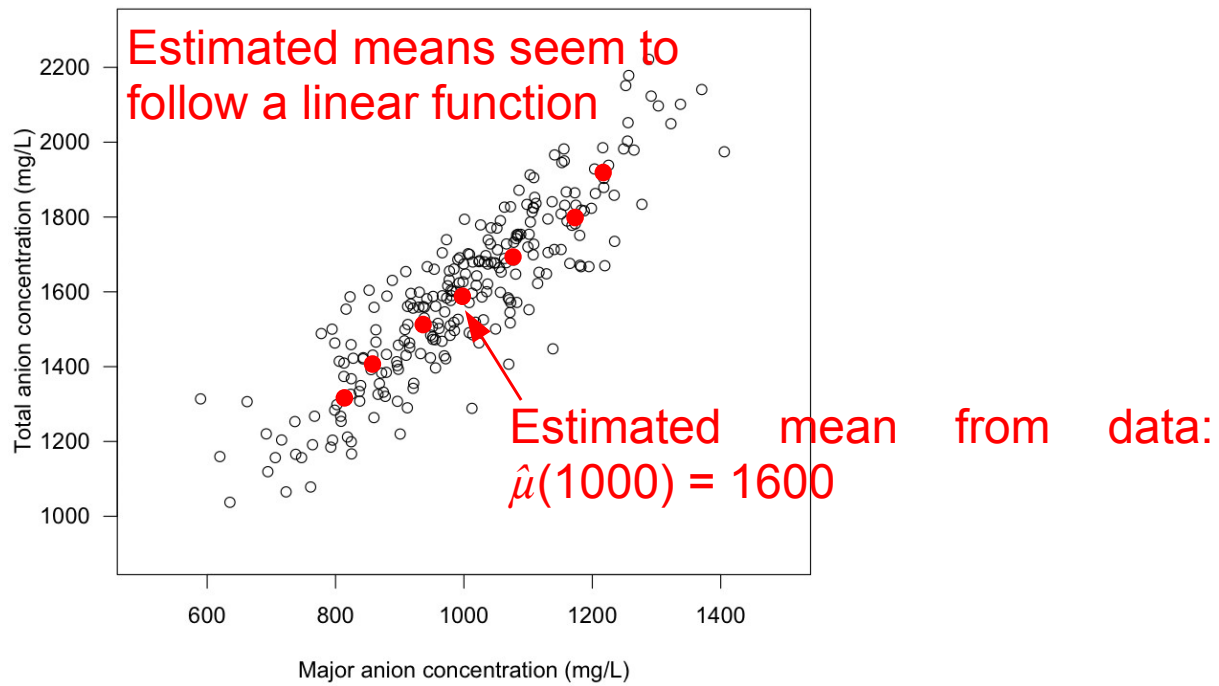
We assume a linear function of the true population $\mu(X)$:

$\mu(X) = \beta_0 + \beta_1 X$ from which follows that: $Y = \beta_0 + \beta_1 X + \varepsilon$

β_0 and β_1 : regression coefficients

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept
2. **Simple linear regression model**
3. Calculation of regression coefficients
4. Accuracy of regression coefficients and predictions
5. Confidence intervals – intro and application
6. Accuracy of overall model and sum of squares
7. Model assumptions and diagnostics I
8. Model assumptions and diagnostics II + wrap up

(Simple) Linear regression model

Assuming that the true relationship is a linear function of the form $Y = \beta_0 + \beta_1 X + \varepsilon$, we can use sample data to obtain estimates of β_0 and β_1, denoted as $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$, and subsequently predict $\hat{Y}$:

$$\hat{Y} = b_0 + b_1 X \quad \text{for realisations of } X \text{ we can rewrite this to:}$$
$$\hat{y}_i = b_0 + b_1 x_i$$

What are β_0 and β_1?

$\beta_0 = E(Y|X = 0)$ "intercept"

$\beta_1 = \dfrac{dy}{dx}$ "slope"

$b_0 = 138$

$b_1 = 1.5$

Fitted line. Total anion concentration (mg/L) vs Major anion concentration (mg/L). Δy, Δx, $x = 0$.
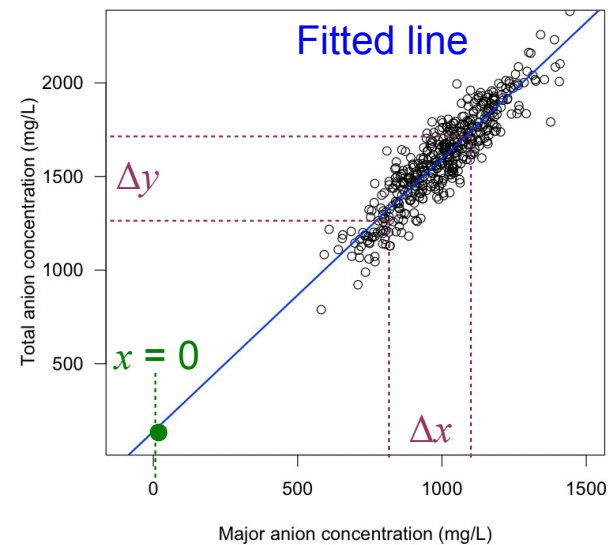
The model that contains only one explanatory variable/predictor is called simple linear regression model.

The intercept of a linear regression model relates to the expected value for $X = 0$. In a figure, it is the intersection of the regression line with the Y axis at $X = 0$.

The slope is represented in the figure as the change in y per change in x. In the figure, we use Δx for the difference between two given numbers for x. By contrast, dx refers to an infinitesimal small change in x.

What is the optimal regression line?

We defined for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$

$\Rightarrow Y = \hat{Y} + \varepsilon \quad \Leftrightarrow \quad \varepsilon = Y - \hat{Y}$

For sample data ($i = 1,2,3, \ldots, n$) and the regression model, we defined: $\hat{y}_i = b_0 + b_1 x_i$
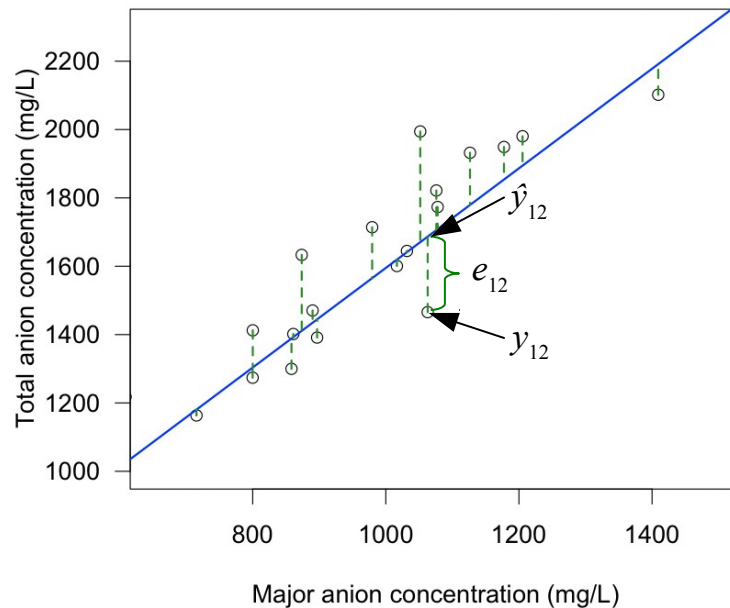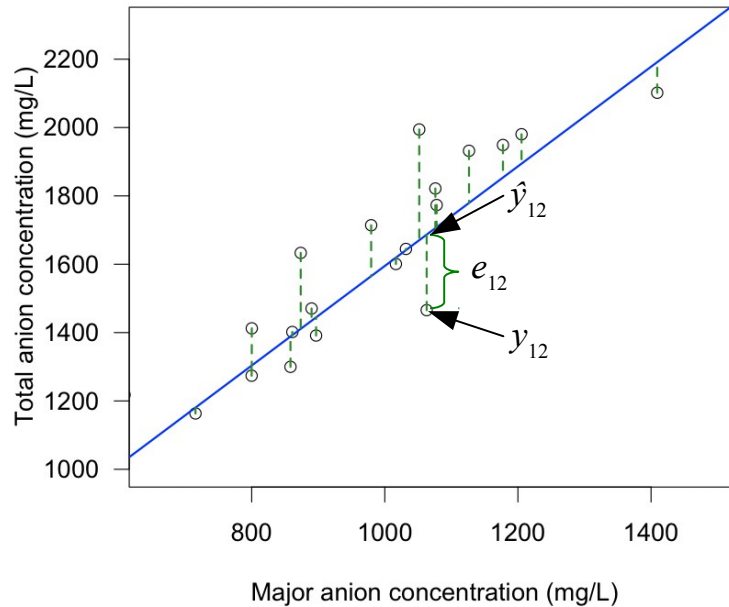
We define the residual e_i as: $e_i = y_i - \hat{y}_i = y_i - b_0 + b_1 x_i$

Example for
observation $i = 12$

What is the optimal regression line?

Example for observation $i = 12$

Optimal line minimises the Residual Sum of Squares (RSS):

$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

$$= \left(y_1 - b_0 + b_1 x_1\right)^2 + \left(y_2 - b_0 + b_1 x_2\right)^2 + ... + \left(y_n - b_0 + b_1 x_n\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_i\right)^2 \Rightarrow \text{Find} \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_i\right)^2$$

14

We have defined ε as the statistical error. However, we can only determine the statistical error, if we know the true linear relationship, i.e. the true values for β_0 and β_1. Given that we obtain our predictions from the estimates b_0 and b_1, we cannot determine the true error but only an estimate of the error, called residual e. Consequently, we call the sum of the squared estimates of the errors the residual sum of squares (RSS). Other terms used are sum of squared residuals (SSR) and sum of squared residual errors (SSE). Sometimes the misleading term sum of squared errors (SSE) is used, but this should be avoided, because we only deal with estimates of errors. We will return to these terms later.

Note also that in some text books the residuals are simply called error and it is not distinguished between the population error and the sample estimate of the error (residual).

Due to the focus on the minimisation of the sum of squares, the model is also called *ordinary least squares* (OLS). The term "ordinary" was added to distinguish the model from the many others, later developed, relying on least squares approaches (e.g. weighted least squares, generalised linear models).

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept
2. Simple linear regression model
3. **Calculation of regression coefficients**
4. Accuracy of regression coefficients and predictions
5. Confidence intervals – intro and application
6. Accuracy of overall model and sum of squares
7. Model assumptions and diagnostics I
8. Model assumptions and diagnostics II + wrap up

15

15

Determining the regression coefficients

$$\text{Find} \arg\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2$$

It can be shown that the minimizing values are:

$$b_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

<u>Matrix notation of linear regression clarifies calculation</u>

$$\hat{y}_i = b_0 + b_1 x_i \quad \text{Notation for the observations } i = 1,2,3, \ldots, n:$$

$$\hat{y}_1 = b_0 + b_1 x_1$$
$$\hat{y}_2 = b_0 + b_1 x_2$$
$$\vdots$$
$$\hat{y}_n = b_0 + b_1 x_n$$

matrix →

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

$$\hat{Y} = X b$$

16

Note that a column of 1s is included in the matrix X that are multiplied with the intercept. Hence, this matrix is not identical with the values for the random variable X.

The derivation of the minimizing values is presented in text books, e.g. Acevedo (2013: 180-182). Acevedo (2013:183) also describes the relationship between correlation and regression through a different derivation of b_1:

$$b_1 = r_{XY} \frac{s_Y}{s_X}$$

In words, b_1 is the sample Pearson correlation coefficient multiplied with the ratio of the sample standard deviations.

Acevedo M.F. (2013) Data analysis and statistics for geography, environmental science, and engineering. CRC Press, Boca Raton.

16

Example calculation of coefficients

$$\hat{Y} = Xb$$

It can be shown that:

$$b = (X^T X)^{-1}(X^T Y)$$

Example: Calculation for trivial data

Let our set of data be $\{(10,4)(20,5)\} \Rightarrow x = \{10, 20\}, y = \{4, 5\}$

Matrix notation: $X = \begin{pmatrix} 1 & 10 \\ 1 & 20 \end{pmatrix}, X^T = \begin{pmatrix} 1 & 1 \\ 10 & 20 \end{pmatrix}, Y = \begin{pmatrix} 4 \\ 5 \end{pmatrix}$

17

For a mathematical derivation of the second statement see Matloff (2017): 70-72 (or shorter: Hastie et al. (2017: 45)).

The data represent observed tupels of (x,y).

Example calculation of coefficients

$$\hat{Y} = X\,b$$

It can be shown that:

$$b = (X^T X)^{-1}(X^T Y)$$

Inverse

Example: Calculation for trivial data

Let our set of data be $\{(10,4)(20,5)\} \Rightarrow x=\{10,\,20\},\, y=\{4,\,5\}$

Matrix notation: $X=\begin{pmatrix}1 & 10\\ 1 & 20\end{pmatrix},\, X^T=\begin{pmatrix}1 & 1\\ 10 & 20\end{pmatrix},\, Y=\begin{pmatrix}4\\ 5\end{pmatrix}$

Calculation of b:

$$b = (X^T X)^{-1}(X^T Y) = \left(\begin{pmatrix}1 & 1\\ 10 & 20\end{pmatrix}\begin{pmatrix}1 & 10\\ 1 & 20\end{pmatrix}\right)^{-1}\left(\begin{pmatrix}1 & 1\\ 10 & 20\end{pmatrix}\begin{pmatrix}4\\ 5\end{pmatrix}\right)$$

$$= \begin{pmatrix}1\cdot 1+1\cdot 1 & 1\cdot 10+1\cdot 20\\ 10\cdot 1+20\cdot 1 & 10\cdot 10+20\cdot 20\end{pmatrix}^{-1}\begin{pmatrix}1\cdot 4+1\cdot 5\\ 10\cdot 4+20\cdot 5\end{pmatrix} = \begin{pmatrix}2 & 30\\ 30 & 500\end{pmatrix}^{-1}\begin{pmatrix}9\\ 140\end{pmatrix}$$

18

For an introduction into matrix algebra including an explanation of the inverse, see the course materials (Key terms and concepts …).

Example calculation of coefficients

Calculation of the inverse   For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the inverse is $A^{-1} = \dfrac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$$b = \begin{pmatrix} 2 & 30 \\ 30 & 500 \end{pmatrix}^{-1} \begin{pmatrix} 9 \\ 140 \end{pmatrix} = \left(\frac{1}{2 \cdot 500 - 30 \cdot 30} \begin{pmatrix} 500 & -30 \\ -30 & 2 \end{pmatrix} \right) \begin{pmatrix} 9 \\ 140 \end{pmatrix}$$

$$= \left(\frac{1}{100} \begin{pmatrix} 500 & -30 \\ -30 & 2 \end{pmatrix} \right) \begin{pmatrix} 9 \\ 140 \end{pmatrix} = \begin{pmatrix} 5 & -0.3 \\ -0.3 & 0.02 \end{pmatrix} \begin{pmatrix} 9 \\ 140 \end{pmatrix}$$

$$= \begin{pmatrix} 5 \cdot 9 - 0.3 \cdot 140 \\ -0.3 \cdot 9 + 0.02 \cdot 140 \end{pmatrix} = \begin{pmatrix} 45 - 42 \\ -2.7 + 2.8 \end{pmatrix} = \begin{pmatrix} 3 \\ 0.1 \end{pmatrix}$$

$$\Rightarrow b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 3 \\ 0.1 \end{pmatrix} \qquad \text{The intercept } b_0 = 3 \text{ and the slope } b_1 = 0.1$$

19

20

Confirmation in R and accuracy of results

Calculation in R

Although the matrix calculation may seem complicated, you are now able to conduct a regression analysis during a power outage. Of course, there are easier ways. R can do the calculation within the split of a second and confirms our result:

```
x <- c(10,20)
y <- c(4,5)
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##         3.0          0.1
```

But how good is the model?

Depends on what we mean by "good", we can evaluate different aspects, e.g.:

- Accuracy of the estimated regression coefficients

- Accuracy of predictions

- Overall model (as compared to other models)

The meaning of accuracy is defined in the document "Key and concepts ..." in 4.1.

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. **Accuracy of regression coefficients and predictions**

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. Model assumptions and diagnostics II + wrap up

How accurate are the estimates for the regression coefficients?

Recall that the standard error informs on the error of a parameter estimate (e.g. mean of X):

$$\text{SE}_{\overline{X}} \approx \frac{s}{\sqrt{n}} \quad \text{where} \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Also recall the concept of RSS :

$$\text{RSS} = \sum_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Average squared residual: Mean squared error

$$\text{MSE} = \frac{1}{\text{DoF}} \text{RSS} \quad \text{where DoF} = \text{Degrees of freedom}$$

DoF for regression model: $n - p - 1$ where $p = \begin{array}{l} \text{no of parameters in model} \\ \text{excluding the intercept} \end{array}$

23

See the course materials (Key terms and concepts …) for an explanation of the concept of the standard error and of accuracy.

The MSE informs on the accuracy of the overall model. You may wonder, why it is called MSE and not mean squared residual – so do I and others: https://stats.stackexchange.com/questions/183311/shouldnt-the-root-mean-square-error-rmse-be-called-root-mean-square-residual
As discussed before, the terminology unfortunately varies within the literature and is not fully consistent.
The square root of MSE is called root mean square error (RMSE), other terms used are: square root of the mean squared residual, residual standard error (e.g. in R and in Hastie et al. (2017)) or residual standard deviation.

In a simple linear regression model, i.e. a model with the regression coefficients b_0 and b_1, we have $n - 2$ degrees of freedom (DoF). This is to account for the number of parameters estimated. In a regression model without intercept the denominator, i.e. DoF, turns into n-p.

How accurate are the estimates for the regression coefficients?

Now that we have introduced MSE, we provide the standard errors for the regression coefficients:

$$\text{SE}_{b_1} = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{SE}_{b_0} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Standard errors of regression coefficients increase with residuals and decrease with a wider range of x values (and sample size).

For details on how the design influences the standard errors of regression coefficients (and further standard errors that will be discussed next), see Maindonald & Braun (2010: p. 151-152). An important point is that the accuracy can also be increased by increasing the range of x values. Thus, increasing the sample size is not the only measure to increase accuracy.

How accurate are the predictions?

We have two standard errors for predictions: The standard error for the predicted mean $\hat{y}$ and for a new observation y_{new}.

$$SE_{\hat{y}_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad \text{SE of fit}$$

$$SE_{y_{new}} = \sqrt{MSE + MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad \text{SE of prediction}$$

x_h is the value of the predictor. Both SEs are lowest at the mean of $x \rightarrow$ accuracy decreases from centre

Agrees with intuition that the predicted (fitted) mean of y, i.e. $\hat{y}$, or a new y is most accurate in the centre given the many values on both sides that support the prediction.

Both equations differ by one MSE term. Therefore, the SE for $\hat{y}$ will always be narrower than for a new observation y_{new}. This is not surprising as predicting a new observation combines the uncertainty from predicting the mean with the uncertainty from predicting an observation varying around the mean.

The standard error of the fit can get close to zero, if x_h is close to the mean and n is large. This is not the case for the standard error of the prediction.

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept
2. Simple linear regression model
3. Calculation of regression coefficients
4. Accuracy of regression coefficients and predictions
5. **Confidence intervals – intro and application**
6. Accuracy of overall model and sum of squares
7. Model assumptions and diagnostics I
8. Model assumptions and diagnostics II + wrap up

27

Confidence intervals

For graphical illustration, we need to understand the concept of the confidence interval.

Confidence interval (CI):

- core concept of frequentist statistics

- interval estimate for parameter (e.g. mean) from statistical population

- confidence level (typically 95%) defines frequency the interval contains the true parameter in (hypothetical) repeated studies. I.e. 95% of CIs constructed from repeated samples, contain true parameter.

$\rightarrow$ Probabilistic statement about performance of procedure deriving interval, and not directly about a specific CI

A classical misinterpretation of CIs is to make probabilistic statements such as: "The true parameter has a 95% probability to be inside the interval." and "The CI has a 95% probability to contain the true parameter.". Note that "the interval" and "The CI" refers to a specific CI calculated from a sample that has been drawn from a population. Such statements are incorrect because, the true parameter has a fixed value and once a sample has been drawn from a population, the parameter either is or is not inside the related CI. This may seem counter-intuitive at first, but bear in mind that the frequentist perspective is on (hypothetical) repeated studies. For one specific sample, resulting in a CI, we cannot say whether the true value is inside the interval. Probabilistic statements are only appropriate before a sample from a population is drawn. For example: "The confidence interval for the future study has a 95% probability of containing the true parameter."

For post-data inference, Bayesian approaches provide so-called credibility intervals that will be discussed later.

The background on confidence intervals can be found in virtually any statistical text book. A brief overview is given in Dorey (2010) and in more detail in Schober et al. (2018). Further misinterpretations are outlined in Hoekstra et al. (2014), Morey et al. (2015) and Greenland (2016). If you are interested in the debates that surrounded the introduction of confidence intervals including advice on their interpretation, a very readable overview is given in Lehmann (2011: 80-89). Confidence intervals and *p*-values, which we will discuss later, are subject to heated debates, see for example Morey et al. (2015), Miller & Ulrich (2016) and Morey et al. (2016).

Dorey, F. J. (2010). In Brief: Statistics in Brief: Confidence Intervals: What is the Real Result in the Target Population? *Clinical Orthopaedics and Related Research*, *468*(11), 3137–3138. Free to download: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2947664/

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N., et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31, 337–350. Free to download at: https://link.springer.com/article/10.1007/s10654-016-0149-3

Hoekstra R., Morey R.D., Rouder J.N. & Wagenmakers E.-J. (2014) Robust misinterpretation of confidence intervals. Psychonomic Bulletin & Review 21, 1157–1164.

Lehmann E.L. (2011) Fisher, Neyman, and the creation of classical statistics. Springer, New York.

Miller J. & Ulrich R. (2016) Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). Psychonomic Bulletin & Review 23, 124–130.

Morey R., Hoekstra R., Rouder J., Lee M. & Wagenmakers E.-J. (2015) The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review, 1–21.

Morey R.D., Hoekstra R., Rouder J.N. & Wagenmakers E.-J. (2016) Continued misinterpretation of confidence intervals: response to Miller and Ulrich. Psychonomic Bulletin & Review 23, 131–140.

Schober, P., Bossers, S. M., & Schwarte, L. A. (2018). Statistical Significance Versus Clinical Importance of Observed Effect Sizes: What Do *P* Values and Confidence Intervals Really Represent? *Anesthesia and Analgesia*, *126*(3), 1068–1072. Free to download: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5811238/

Visualisation of CIs: CI for mean

95% CIs for mean of quail sperms

https://upload.wikimedia.org/wikipedia/commons/d/db/Japanese_Quail.jpg
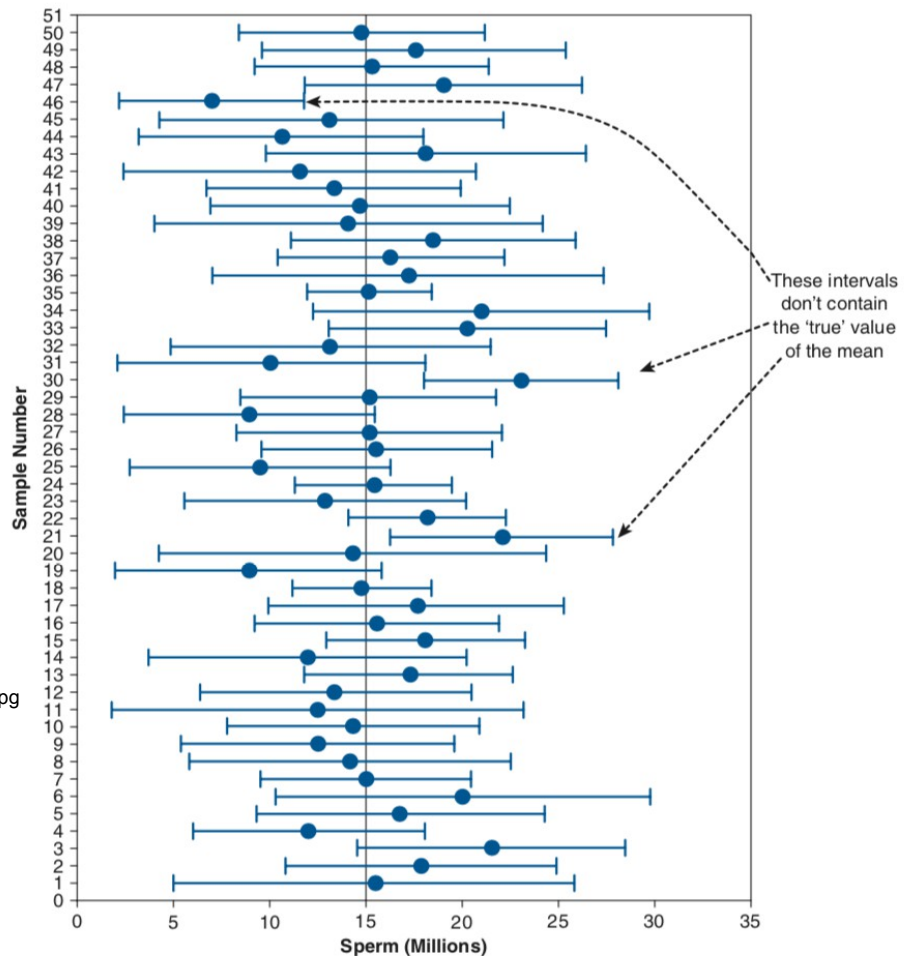
Field, Miles & Field 2013: 46

Such CIs could also be displayed for estimated regression coefficients, that is the regression parameter takes the place of the mean.

The figure displays 50 confidence intervals for sperm samples that have been taken from the quail population (in this case refers to the statistical and biological population). The true population mean (μ) is 15 million sperms, each blue dot represents an estimated mean $(\hat{\mu} = \bar{x})$ from the sample with 95% confidence intervals.

Confidence intervals are typically used with high confidence levels (e.g. 95% like in this example), because when interpreting them, one can assume that the given interval is one of those containing the true parameter. In other words: If the assumption holds that in 95% of cases the CI contains the true parameter and assuming that the specific CI is one of these 95%, then the parameter would be within the given CI limits.

In the example, 3 of 50 CIs do not contain the true parameter. Hence, if we had only drawn one sample from the population and this was sample 46, 30 or 21, we were mislead by the CI. To decrease the chance to be mislead by the CI, we could set a higher confidence level, e.g. 99%. This would mean that of 100 CIs related to samples from the population, we can expect that 99 contain the true parameter.

Visualisation of CIs: Regression bands

95% CI for mean S, i.e. $E(S|K = k)$, (blue) and
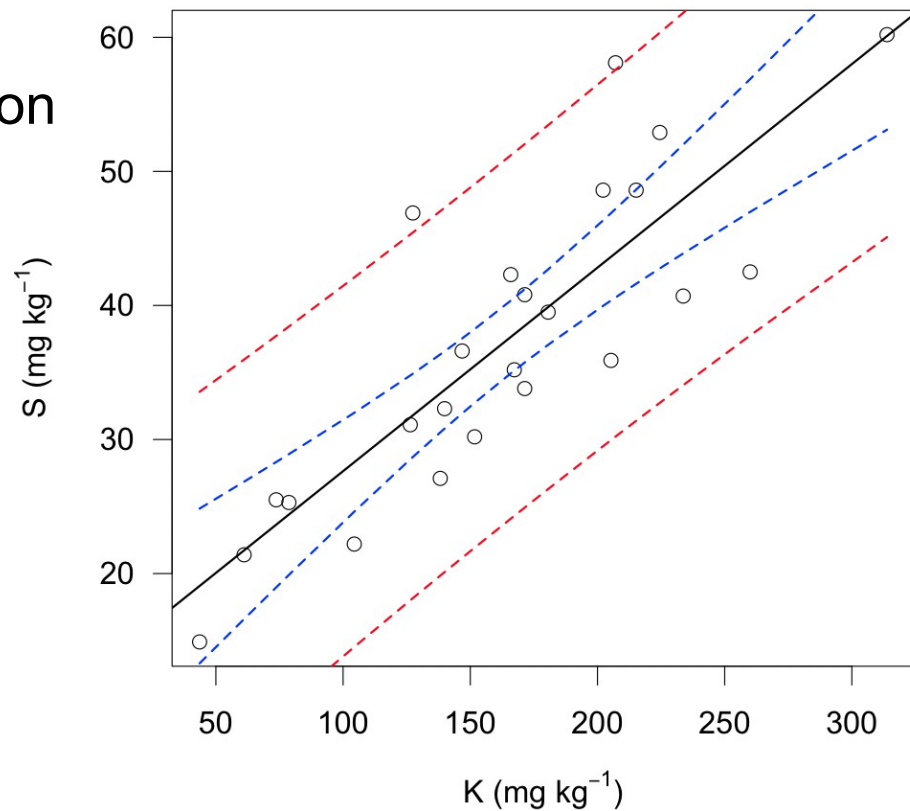95% prediction interval for S, i.e. $S(K = k)$, (red)

Example: Study on soil ion
concentrations
S = Sulphur
K = Potassium

Compared to the CI for a parameter such as regression coefficients, in a linear regression analysis we also obtain estimates of the conditional mean (denoted as fitted $\hat{y}$ before) for every value on the X-axis, in this example for each $K = k$. Thus, we have a point-wise interval estimate for each k, that is also for those not observed, between the minimum and maximum sampled k. These point-wise interval estimates can be displayed as continuous lines or bands, the so-called *regression bands*.

The CI is based on the standard error of the fit, whereas the prediction interval (PI) is based on the standard error of the prediction. The CI is related to the mean S predicted by the linear regression model. By contrast, the PI relates to a new observation for S (and not the mean). It is therefore much wider as it combines the uncertainty from predicting the mean with the uncertainty from predicting an observation varying around the mean. The interpretation of the PI follows that of the CI: For (hypothetical) repeated studies that produce sample data used in regression analysis, approximately a proportion equalling the confidence level of the PIs will contain the true value of y (S in our case).
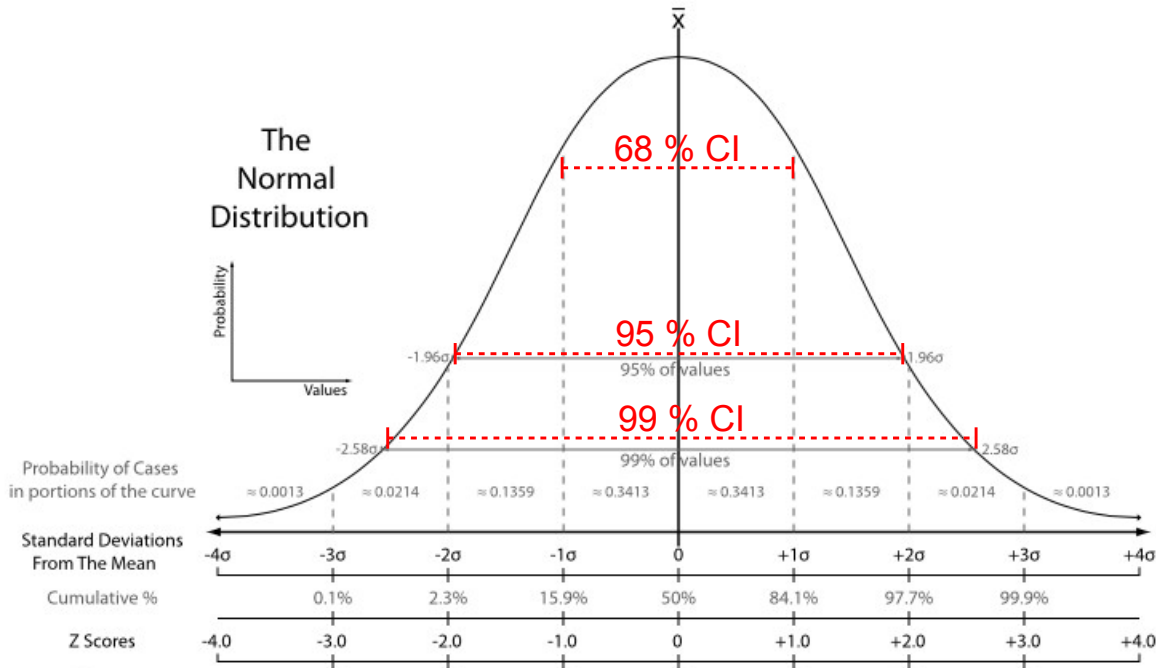
Calculation and interpretation of CIs

General calculation of confidence interval for parameter θ:
$\theta \pm$ value related to confidence level and probability distribution × SE

(e.g. 95%, 99%)   Known or assumed probability distribution of parameter

Example: Values (= z-scores) for normal probability distribution

The Normal Distribution

68 % CI

95 % CI

99 % CI

-1.96σ   95% of values   1.96σ

-2.58σ   99% of values   2.58σ

Probability of Cases in portions of the curve	≈ 0.0013	≈ 0.0214	≈ 0.1359	≈ 0.3413	≈ 0.3413	≈ 0.1359	≈ 0.0214	≈ 0.0013
Standard Deviations From The Mean	-4σ	-3σ	-2σ	-1σ	0	+1σ	+2σ	+3σ	+4σ
Cumulative %		0.1%	2.3%	15.9%	50%	84.1%	97.7%	99.9%	
Z Scores	-4.0	-3.0	-2.0	-1.0	0	+1.0	+2.0	+3.0	+4.0

31   https://upload.wikimedia.org/wikipedia/commons/2/25/The_Normal_Distribution.svg

Although CIs or the related bands only indicate the boundary of an interval, the probability density is not uniform within the boundaries. For example, the 95% CI has the double width of the 68% CI for an underlying normal probability distribution (and not an approximately 1.5 fold width as for a uniform distribution).

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept
2. Simple linear regression model
3. Calculation of regression coefficients
4. Accuracy of regression coefficients and predictions
5. Confidence intervals – intro and application
6. **Accuracy of overall model and sum of squares**
7. Model assumptions and diagnostics I
8. Model assumptions and diagnostics II + wrap up

How accurate is the overall model?

Concept of Root mean square error (RMSE), allows for comparison across models:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{\mathrm{DoF}} \mathrm{RSS}} = \sqrt{\frac{1}{\mathrm{DoF}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Concept of explained variance, allows for comparison across linear models:

$$R^2 = r^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \quad \text{where TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
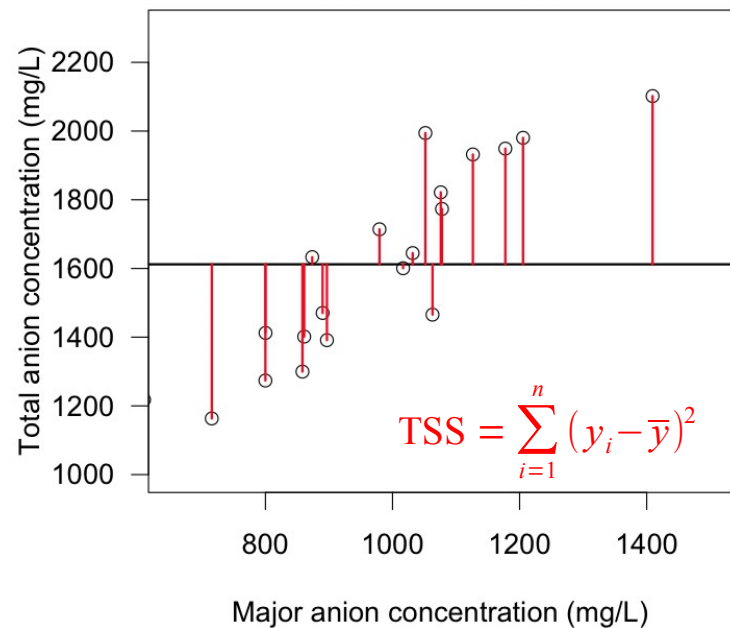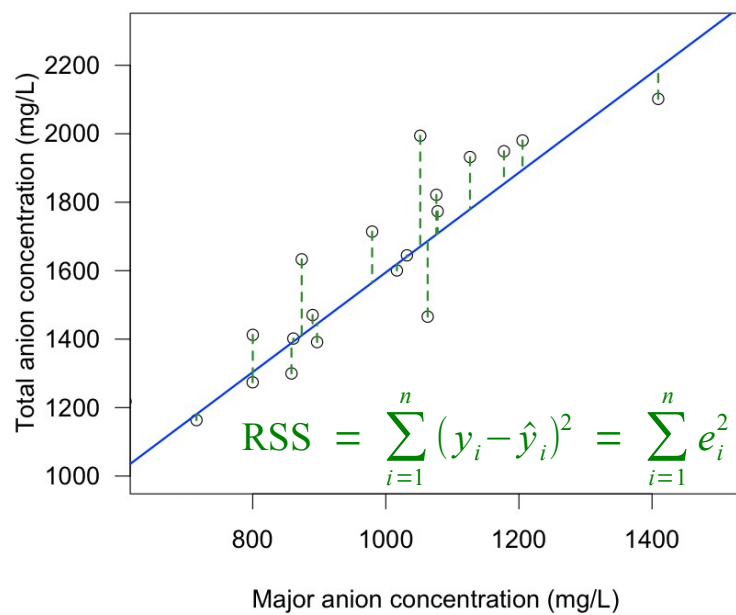
Pearson
correlation
coefficient

Total sum of squares

33

We will discuss the concept of the R^2 a bit more later in the course.

What are these different sum of squares?

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

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

TSS (total Var) = MSS (explained Var) + RSS (unexplained Var)

MSS = Model sum of squares – captures the sum of squares explained by the model, here regression model

Again, be very careful when using these terms (e.g. RSS, MSS), the terminology varies between sources. A first difference that is rather mild is that some authors use abbreviations starting with SS, e.g. SSR, SSM, SST. More irritating is the use of different letters:

TSS is widely used for the total sum of squares. Nevertheless, Crawley (2015) uses SSY for the total sum of squares, based on the fact that SSY in regression captures the total variation in Y.

Several synonyms for MSS are found in the literature: RSS - regression sum of squares (e.g. in Crawley 2015), (regression) SS (e.g. in Legendre & Legendre 2012), RegSS (for regression sum of squares, e.g. in Fox 2015) and ESS (explained sum of squares (e.g. in Wikipedia: https://en.wikipedia.org/wiki/Explained_sum_of_squares)).

An alternative term frequently used for RSS is SSE – error sum of squares (e.g. in Crawley 2015, Acevedo 2013).

Hence, you need to check the definition of terms, as for example both:

$$R^2 = \frac{RSS}{TSS} \text{ and } R^2 = 1 - \frac{RSS}{TSS}$$

can be correct, depending on whether RSS is defined as the residual sum of squares or regression sum of squares.

We stick to the terminology that is also used by Hastie et al. (2017) and Field et al. (2013).

Acevedo M.F. (2013) Data analysis and statistics for geography, environmental science, and engineering. CRC Press, Boca Raton.

What are these different sum of squares?

TSS (total Var) = MSS (explained Var) + RSS (unexplained Var)

R output for linear regression model

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:        $y_i - \hat{y}_i = e_i$
##      Min        1Q   Median         3Q        Max
## -352.96    -71.49     -2.75      69.60     323.95
##
## Coefficients:          $b_0$          $SE_{b_0}$
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 137.79468    31.92061     4.317 1.95e-05 ***
## X             1.45722     0.03152    46.231  < 2e-16 ***
## ---                  $b_1$        $SE_{b_1}$
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ?  ' 1
##                      RMSE
## Residual standard error: 103 on 448 degrees of freedom
## Multiple R-squared:  0.8267,  Adjusted R-squared:  0.8263
## F-statistic:  2137 on 1 and 448 DF,  p-value: < 2.2e-16
```

R^2

You should be able to interpret most of the output. The parts in the red squares will be discussed later in the course.

Never interpret a linear model without a visual representation

We have 4 regression models with the same regression coefficients and standard errors:

Model 1

```
## $lm1
##                  Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0000909   1.1247468 2.667348 0.025734051
## x1              0.5000909   0.1179055 4.241455 0.002169629
##
```

Model 2

```
## $lm2
##                  Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.000909   1.1253024 2.666758 0.025758941
## x2              0.500000   0.1179637 4.238590 0.002178816
##
```

Model 3

```
## $lm3
##                  Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0024545   1.1244812 2.670080 0.025619109
## x3              0.4997273   0.1178777 4.239372 0.002176305
##
```

Model 4

```
## $lm4
##                  Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0017273   1.1239211 2.670763 0.025590425
## x4              0.4999091   0.1178189 4.243028 0.002164602
```
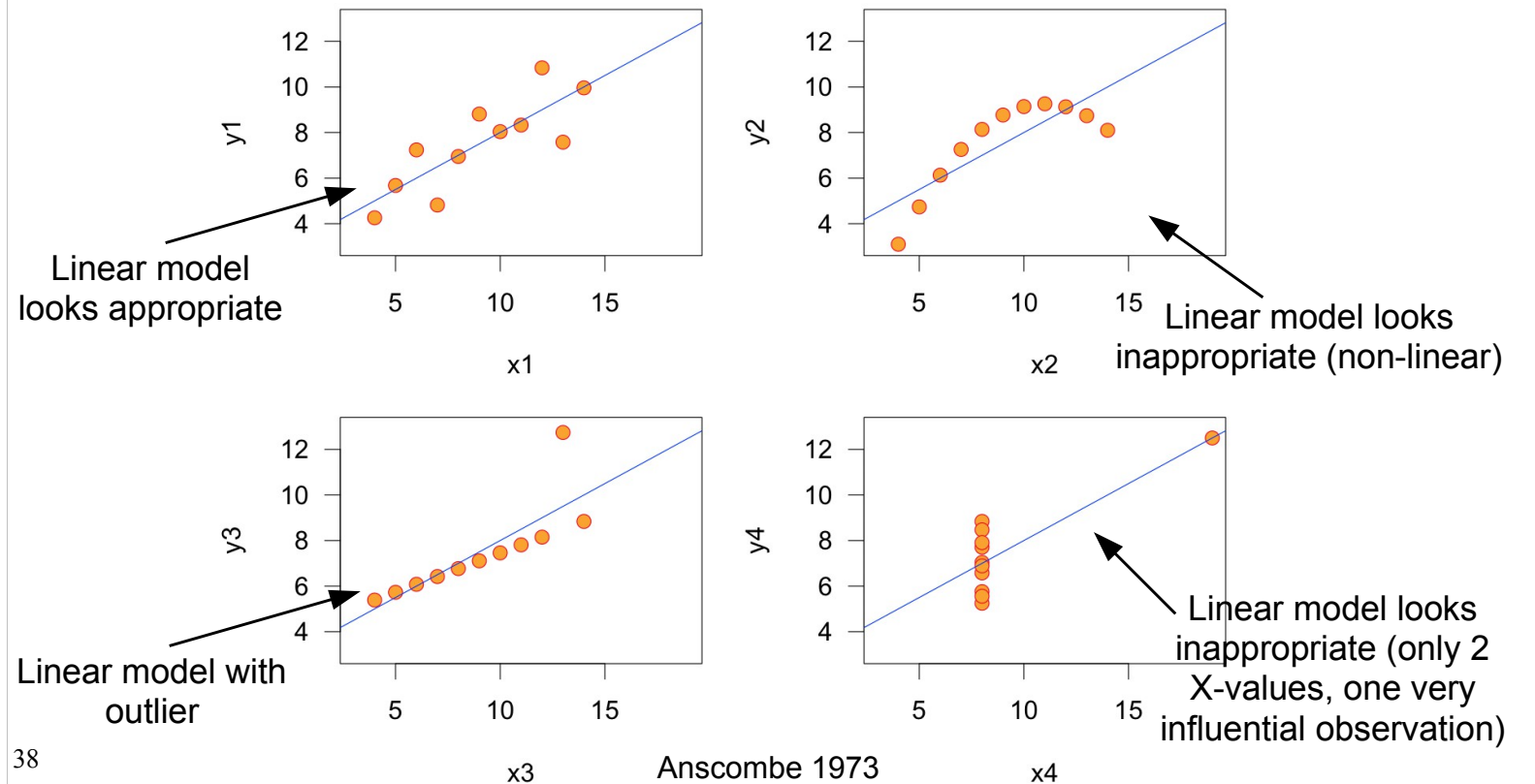
Annotations: b_0, SE_{b_0}, b_1, SE_{b_1}

Never interpret a linear model without a visual representation

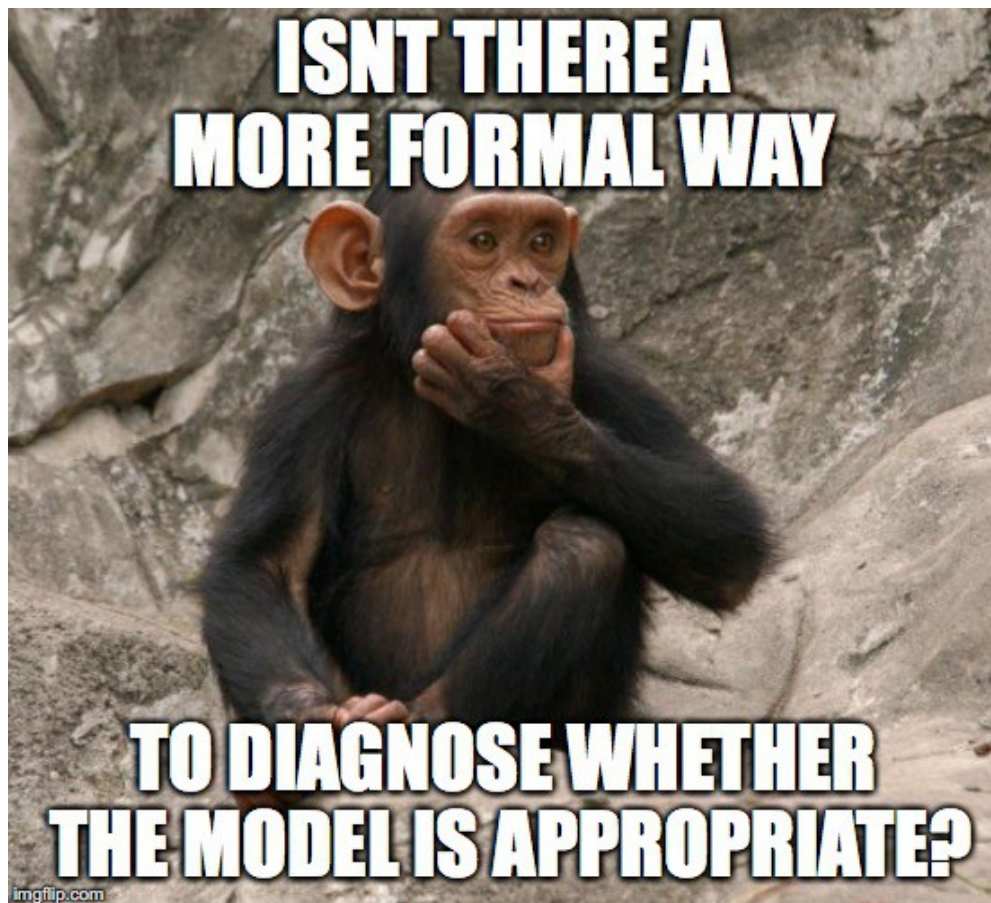We have 4 regression models with the same regression coefficients and standard errors, but the data differ strongly:

Linear model looks appropriate

Linear model looks inappropriate (non-linear)

Linear model with outlier

Linear model looks inappropriate (only 2 X-values, one very influential observation)

Anscombe 1973

Anscombe F.J. (1973) Graphs in Statistical Analysis. The American Statistician 27, 17–21.

38

Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. **Model assumptions and diagnostics I**

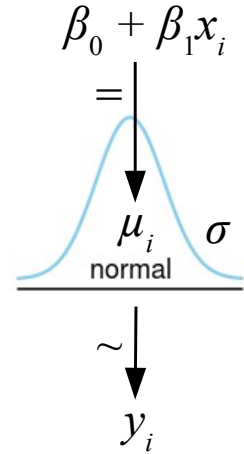8. Model assumptions and diagnostics II + wrap up

Model assumptions

Recap of model and derivation of model assumptions:

Classical model definition

Probability distribution-centric model definition

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$
$$\varepsilon \sim \text{Normal}(0, \sigma)$$

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 x_i$$
$$\varepsilon = y_i - \mu_i$$

$$\beta_0 + \beta_1 x_i$$
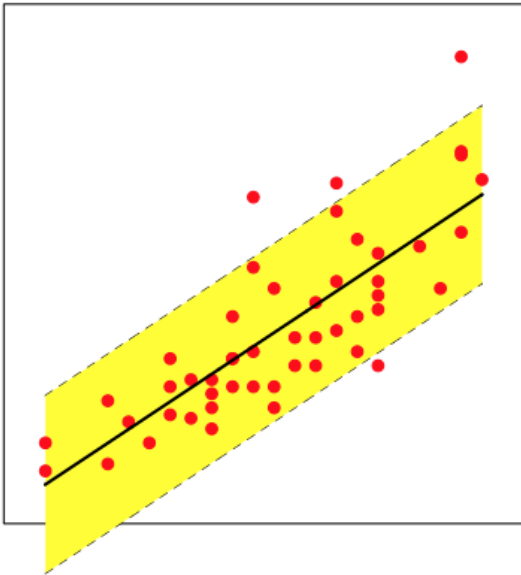$$=$$
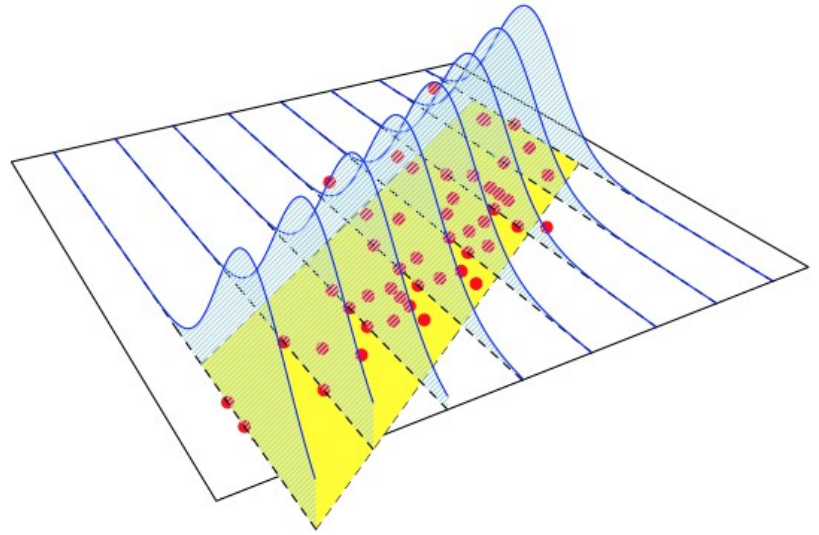$$\mu_i \quad \sigma$$
normal
$$\sim$$
$$y_i$$

- Assumptions:
  - Linear relationship
  - Normal distribution of error
  - Variance homogeneity (Homoscedasticity)
  - Independence of errors

41

Visualisation of some model assumptions

Linear relationship

Normal distribution and constant variance

http://freakonometrics.hypotheses.org/tag/poisson

In the right figure, the assumption of constant variance and normal distribution of errors around the fitted mean $\hat{y}$ for each X is visualised, which would be checked using the red observations. Constant variance is checked by plotting residuals against the corresponding fitted values $\hat{y}$. When checking the assumption of normal distribution, this is done across all observations and not for for each X as typically the number of observations per X is insufficient. The check is based on the distribution of all residuals.

Diagnostics for model assumptions of the simple linear regression model

- Assumptions:

  - Linear relationship: $\mu_i = \beta_0 + \beta_1 x_i$ → checking via model visualisation and residuals *vs.* fitted values plots

  - Normal distribution of error → checking via QQ-plots

  - Variance homogeneity: $\mathrm{Var}(Y|\,X = x)$ is constant in x → checking via residuals *vs.* fitted values plots

  - Independence of errors → checking via serial correlation plots

Although hypothesis tests (hypothesis testing will be introduced a bit later) for checking the assumptions exist, most textbooks recommend graphical diagnostics. During the R implementation, we give an example how hypothesis tests can be misleading.

The assumption of linearity can simply be checked by plotting the observations and the fitted line. See for example the Anscombe plots shown before. Residuals *vs.* fitted plots can also aid in detecting non-linearity. Later, for the case of multiple variables, we will get to know so-called *partial residual plots.*

We have already discussed QQ-plots for checking the normal distribution assumption.

We will discuss how to check the assumption of variance homogeneity (also called homoscedasticity) with residuals *vs.* fitted plots in the following.

The independence assumption is typically violated in situations of spatially or temporally structured data or for data that are nested/ hierarchically structured. Such data are not the main focus of this course. See Faraway (2015, p. 81-83) or see Plant (2012) for diagnostic tools to spot serial correlation.

We will discuss later, what can be done, if assumptions are violated.

Plant R.E. (2012) Spatial data analysis in ecology and agriculture using R. CRC 43Press, Boca Raton.

Model diagnostics: Variance homogeneity
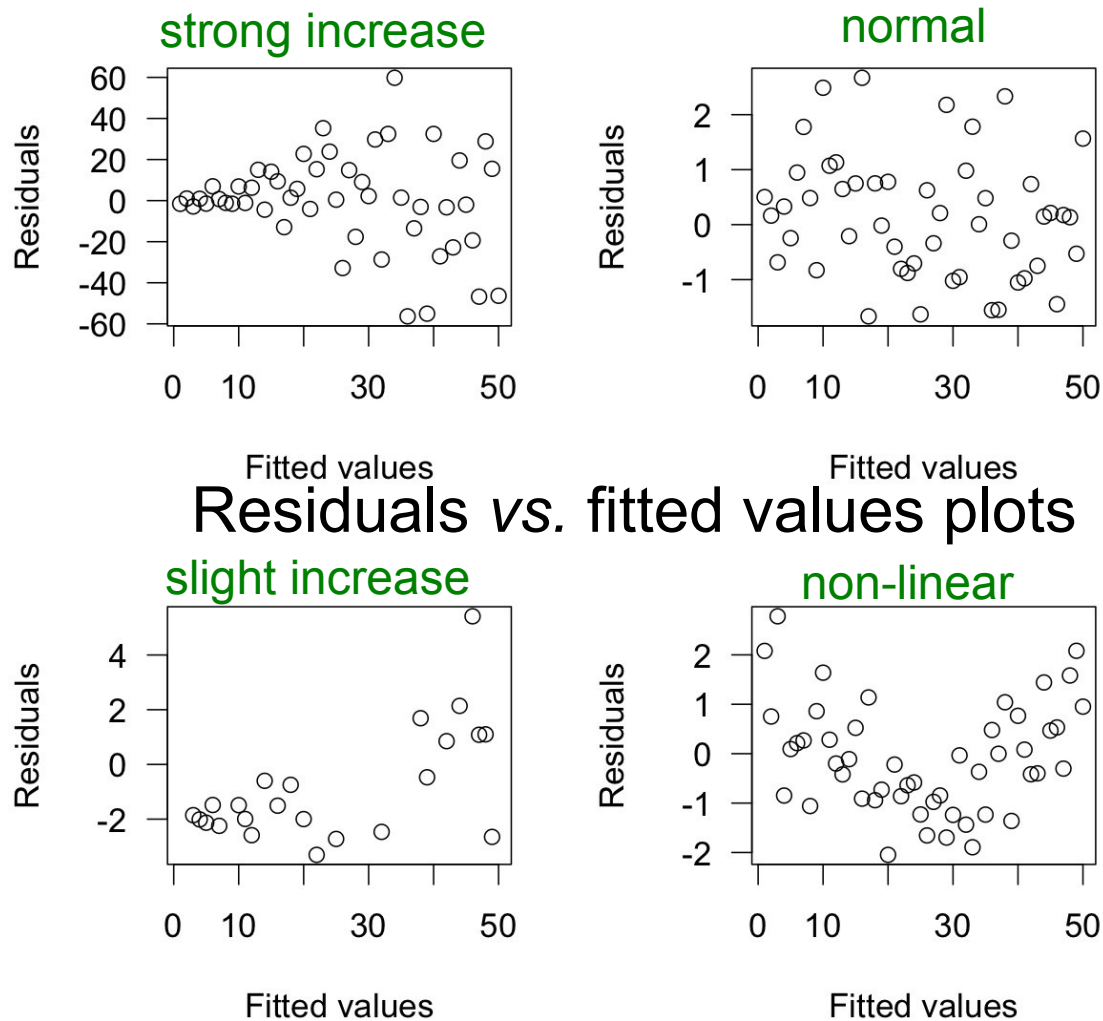
The displayed residuals *vs.* fitted values plots can be used to check whether the assumption of variance homogeneity holds. If the residuals are not randomly distributed (top right) but display patterns, this may indicate non-constant variance, also termed heterogeneous variance or heteroscedasticity, (bottom and top left) or non-linearity (bottom right).

How to deal with violation of model assumptions

- Violation of assumptions can invalidate specific or all regression estimates → results may be unreliable

- Often multiple assumptions violated simultaneously (e.g. non-linear relationship can cause non-constant error variance)

- Check whether important variables are missing from model → can lead to violation of assumptions (e.g. serial correlation, non-linear relationship)

- Data is dependent, shows serial correlation:

  - Aggregate data to achieve independence

  - Model error structure (e.g. Generalised least squares)

  - Use different model (e.g. Linear mixed effect model)

If model diagnostics show that model assumptions are not met, you should first check whether the model misses one or several important variables. A missing variable can lead to serial correlation and other issues (see Faraway 2015: 83). We will later discuss how to model multiple variables simultaneously with a linear model.

Data dependency often arises in case of temporal or spatial data. For example, repeated measurements over time from an individual organism or stream can result in serial correlation, also referred to as (temporal) *autocorrelation*. Similarly, different spatial proximities between sampling units or sites in field studies may result in serial correlation, because sampling units that are closer may be more similar. For example, when measuring the amount of rain, two sampling units that are 1 km apart will presumably be more similar in the amount of rain they receive than two units 1000 km apart. However, a detailed treatment of this issue and guidance for study design is beyond the scope of this course. A very readable introduction into the design of field studies and how to consider spatial autocorrelation is given by Fortin & Dale (2014: Chapter 1).

In case of spatial or temporal autocorrelation, the spatial and temporal structure can be incorporated into the model using generalised least squares (see chapter 4 in Zuur (2009)). More complex nested and hierarchically structured data can be modelled with mixed effect models, which will be discussed later.

Fortin M.-J. & Dale M.R.T. (2014) Spatial analysis : a guide for ecologists, 2nd edition. Cambridge Univ. Press, Cambridge.

Zuur, A. F. et al. 2009: Mixed effects models and extensions in ecology with R. Springer: New York

How to deal with violation of model assumptions

- Relationship is non-linear, discernible in model plot or residuals *vs.* fitted values plot:

  - Use different model (e.g. Generalised linear model, Generalised additive model, Random forest)

  - Transform variable(s)

- Variance is not constant (i.e. is heteroscedastic), discernible in residuals *vs.* fitted values plot:

  - Correct standard errors (e.g. Heteroscedasticity-consistent standard errors)

  - Use different model (e.g. Generalised least squares, Generalised linear model)

  - Transform variable(s)

We will discuss the generalised linear model (GLM) and random forests in detail later. Moreover, the issue of variable transformation will be discussed shortly.

Correcting the standard errors for heteroscedasticity using so-called *sandwich estimators* yields to reliable estimates of the standard errors. The estimates of the regression coefficients are not affected by the corrections and remain the same as in the original model. See Fox (2015: 305) and Matloff (2017: 135-137) for technical details on sandwich estimators. In more severe cases of heteroscedasticity, a different model that accounts for heteroscedasticity in the estimation of regression coefficients should be used. What constitutes a severe case? Fox (2015: 307) suggests that heteroscedasticity has serious effects when the ratio of the largest to smallest variance of residuals is > 4 (or > 10 if we can safely assume normal distribution of the errors). The before mentioned generalised least squares model can be used if the linearity assumption holds (see chapter 4 in Zuur (2009)). If non-constant variance is likely a result of non-linearity check whether this can be cured using a GLM.

Zuur, A. F. et al. 2009: Mixed effects models and extensions in ecology with R. Springer: New York.

How to deal with violation of model assumptions

- Error not normally distributed, discernible in QQ-plot with residuals:

  - Least important assumption, model relatively robust towards violation

  - Violation often associated with violation of other assumptions → curing other violations often achieves normality

  - If any other assumption is violated, a non-normal error distribution is particularly problematic for heavy-tailed or strongly skewed error distributions. In this case:

    - Use robust linear regression method (e.g. median regression)
    - Use different model (e.g. Generalised linear model)
    - Transform variable(s)

For an example of robust regression methods (with R code) see Matloff (2017: 236-238) and Faraway (2015: 123-130).

Be careful with transformations!
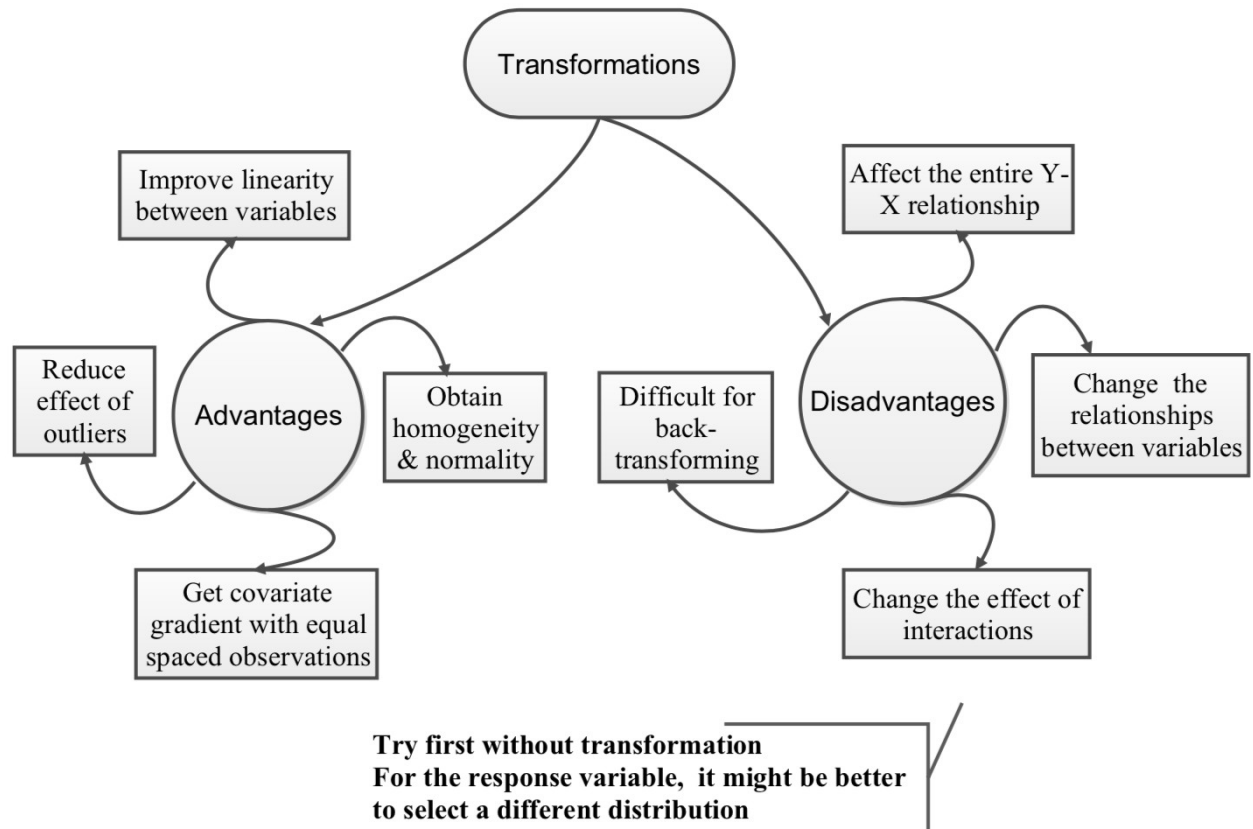
Change model (e.g. use a GLM) if other model fits better in terms of distribution or shape of relationship

Ieno & Zuur 2015: 36

In the past, most articles and text books used variable transformation or suggested to transform variables to meet the assumptions of the linear model (St-Pierre et al. 2018). However, alternative models such as the generalized linear model (GLM), which will be discussed later, are available that allow for fitting models to a wide range of data without prior transformation. For example, if the response represents count or proportion data, it can be directly modelled with a GLM. Hence, when you can model your data without transformation, you should rather reformulate the model than stick to a linear model, though this will also depend on the properties of the data. Xiao et al. (2011) give guidance on when to log-transform and use a linear model or when to use a nonlinear regression model. Szöcs & Schäfer (2015) discuss this issue with a particular emphasis on ecotoxicological data. Warton et al. (2016) give guidance on what to consider when selecting a model for count data.

Ieno E.N. & Zuur A.F. (2015) A beginner's guide to data exploration and visualization with R. Highland Statistics Ltd, Newburgh.

St-Pierre A.P., Shikon V. & Schneider D.C. (2018) Count data in biology-Data transformation or model reformation? Ecology and Evolution 8, 3077–3085.

Szöcs E. & Schäfer R. (2015) Ecotoxicology is not normal. Environmental Science and Pollution Research 22, 13990–13999.

Warton D.I., Lyons M., Stoklosa J. & Ives A.R. (2016) Three points to consider when choosing a LM or GLM test for count data. Methods in Ecology and Evolution 7, 882–890.

Xiao X., White E.P., Hooten M.B. & Durham S.L. (2011) On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. Ecology 92, 1887–1894.

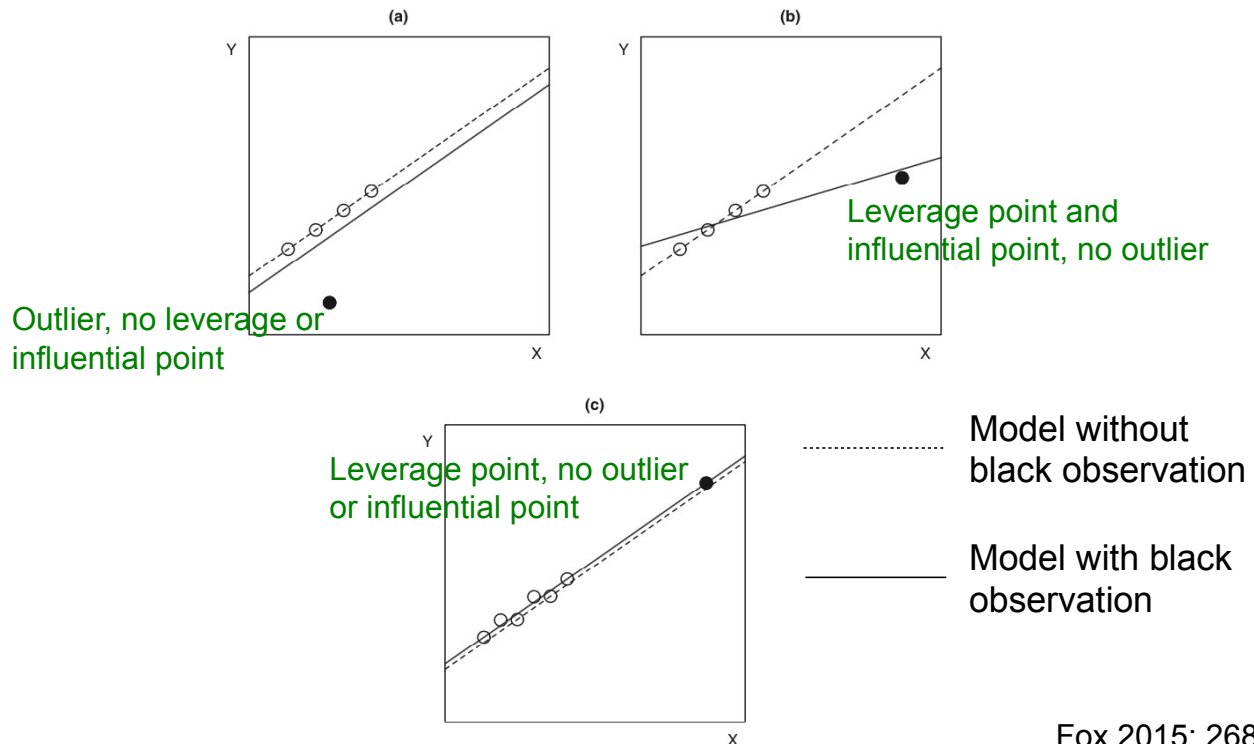# Introduction to the linear model: simple linear regression

Contents

1. Research goals and general concept

2. Simple linear regression model

3. Calculation of regression coefficients

4. Accuracy of regression coefficients and predictions

5. Confidence intervals – intro and application

6. Accuracy of overall model and sum of squares

7. Model assumptions and diagnostics I

8. **Model assumptions and diagnostics II + wrap up**

Further model diagnostics: Outliers

Some terminology:

- Outlier: Unusual observation in $Y|X$ deviating from model

- Leverage point: Unusual observation regarding distribution of X

- Influential point: Observation influencing model fit (e.g. coefficients)

(a) Outlier, no leverage or influential point

(b) Leverage point and influential point, no outlier

(c) Leverage point, no outlier or influential point

Model without black observation

Model with black observation

Fox 2015: 268

Beside checking for assumptions, model diagnostics are used to detect outliers.

In the context of the linear regression model, an outlier is an unusual observation regarding its Y value, conditional on its X value (i.e. the value for Y might not be unusual for another X). A leverage point is a predictor outlier, i.e. an observation with a value of X distant to the range of the other values.

Influential points exercise high influence on the model fit in terms of regression coefficients or the explained variance R^2. Leverage points and outliers are not necessarily influential points, but an influential point is usually at least one of these.

Diagnosing leverage: Hat values

<u>Recap</u>

Linear model in matrix form: $\qquad\qquad \hat{Y} = X\,b$

Estimate regression coefficients via: $\quad b = (X^T X)^{-1}(X^T Y)$

Substitution of b by $\ X^{-1}\hat{Y}\ $: $\qquad\qquad \hat{Y} = \underbrace{X(X^T X)^{-1}(X^T} Y)$

Hat matrix **H** (named for "putting a hat on **Y**")

So-called *hat values* h_i (calculated from **H**) summarise influence of Y_i on <u>all</u> fitted $\hat{Y} \rightarrow$ higher h, higher influence on fitted Y (but not necessarily influential point)
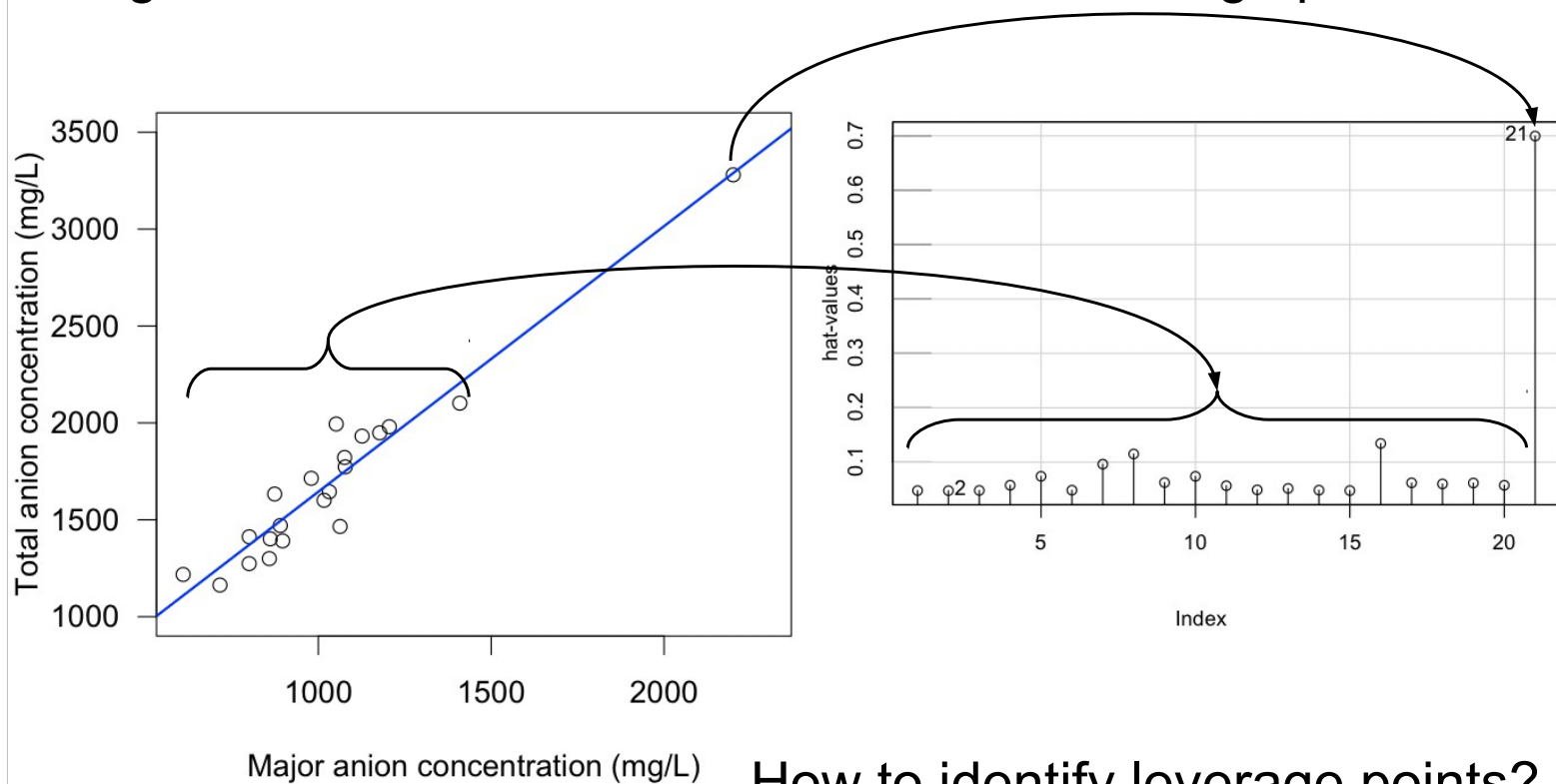
h measures distance to mean of X

For further details see Fox (2015: 270).

Diagnosing leverage: Hat values
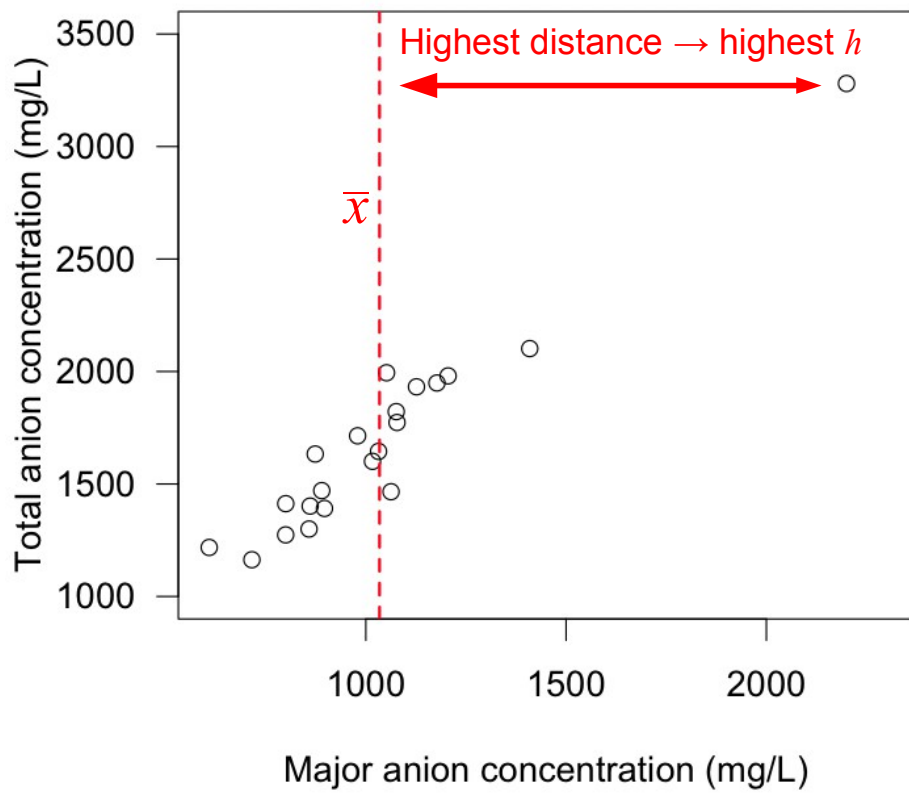
Diagnosing leverage: Hat values

High h relative to other observations $\rightarrow$ Leverage point

How to identify leverage points?

$$\bar{h} = \frac{p}{n} \quad \text{Check observations with } h > 2\frac{p}{n}$$

p = number of parameters in model (including intercept)
n = sample size, number of observations

Leverage points exercise high influence on the fitted $\hat{y}$ (but not necessarily on the model fit) because they are distant from the other x-values (see previous slides).

Faraway (2015):83 and Sheater (2009) suggest to inspect points with hat values > 2-fold the average hat value.

However, hat values do not consider the deviation from the fitted $\hat{y}$ and graphical inspection is most suitable to check whether a high leverage point is really problematic. Notwithstanding, we will learn about a tool to inspect for influential points shortly, which may be used to evaluate whether a leverage point is problematic.

A nice illustration of leverage can be found here: http://www.rob-mcculloch.org/teachingApplets/Leverage/index.html.
Move the observation with the highest x value and see how this affects

the fitted regression line.

Diagnosing outliers and influential points

Outliers:

- can be identified via residuals $e \rightarrow$ higher e, higher deviation from model

- Standardisation of residuals by their standard error and by h, because they are influenced by h:

$$\text{Standardised residual } r_i = \frac{e_i}{\sqrt{\text{MSE}(1-h_i)}}$$
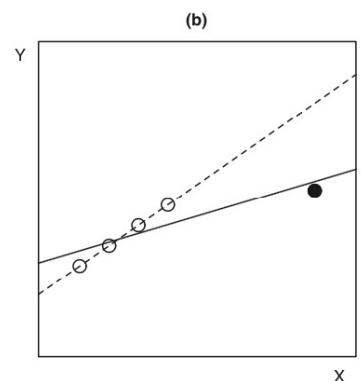
Influential points:

- Intuition: Outliers with high leverage $\rightarrow$ Cook's distance D combines standardised residuals with hat values:

$$D_i = \frac{r_i}{p} \frac{h_i}{(1-h_i)}$$

When we calculated the confidence interval for the fitted $\hat{y}$ we noticed that the confidence intervals were increasing with distance to the mean of x, which is equivalent to an increase of the CI with leverage in terms of the hat value h. Hence, when looking for residuals that are outliers, this needs consideration. Standardised residuals account for the leverage in terms of h. In addition, standardised residuals are the result of the division by the square root of the MSE (also termed residual standard error as discussed before). Remember that for normally distributed data, 95% of data fall within an interval of the mean ± 2-fold the standard error. Consequently, Sheater (2009: 60) suggests to consider standardised residuals that fall outside the interval as outliers, unless you have large data sets, where adaptation of this interval is required. Note that a non-normal error distribution can result in diagnosing many points as outliers, although these observations are rather usual for a wide-tailed non-normal probability distribution. Thus, only check for outliers, after the assumption for the error distribution has been inspected.

A related concept is that of the studentised residual t_i, which is calculated in a similar way as r_i but where observation i has been omitted from the calculation of MSE. The figure from Fox (2015: 268) shows how a leverage point (in black) influences the value of its residual: The leverage point drags the line, thereby decreasing its residual. This explains the motivation of using studentised residuals for outlier inspection. For further details see Sheater (2009: 59-61) and Fox (2015: 272-273).

(b)

Cook's distance D is an important index to diagnose influential points. It measures the influence of observations on the model fit by calculating the combined effect of leverage and of the magnitude of the residual. The higher D_i the higher the change in model fit when observation i is removed from the model. A point with a high Cooks distance tends to be either an outlier or a leverage point or both. There are different rules of thumb as to when consider a point as influential (e.g. $D > 1$ or $D > 4/n\text{-}2$), but in practice it is important to look for gaps in the values for D (Sheater 2009: 68).

How to deal with unusual observations?

- Unusual observations are model-dependent: may disappear if model specification (e.g. variables included), type of model (e.g. GLM instead of linear model) or variables (e.g. transformation) change

- Check whether values are plausible

- Unusual observations that are not influential points → not much to worry

- Check robustness of model results when removing influential observation → report (and plot) both results

- If other model assumptions are met, but model results respond strongly to removal of influential observation:
  - Use robust or quantile regression model
  - Transform data

We refer to outliers, leverage points and influential points as unusual observations.

Never use automated outlier removal! Gossip has it that the ozone hole was later detected because of the automated omission of outliers (e.g. see Faraway 2015: 89). Although this is a nice story to warn against automated omission, the story seems incorrect (Pukelsheim 1990). Nevertheless, to just run an automated statistical methods is rarely appropriate, given that every data set is unique and typically requires specific choices, also based on knowledge of the characteristics of the data.

If you omit observations from a model, report both results in a journal article, at least in the supporting/supplementary information. For an example see Schäfer et al. (2011: Figure 2).

Methods such as robust regression and quantile regression have been developed to deal with influential points (e.g. see Fox 2015: 586). They are outside the scope of this course. For an example of robust regression methods with R code see Matloff (2017: 236-238) and Faraway (2015: 123-130).

Pukelsheim F. (1990): Robustness of Statistical Gossip and the Antarctic Ozone Hole (Letter to the Editor). The IMS Bulletin 19, 540-545. Free to download at: https://www.math.uni-augsburg.de/htdocs/emeriti/pukelsheim/1990c.pdf
Schäfer R.B. et al. (2011) Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. Environmental Science & Technology 45, 1665–1672.

Flowchart for simple linear regression
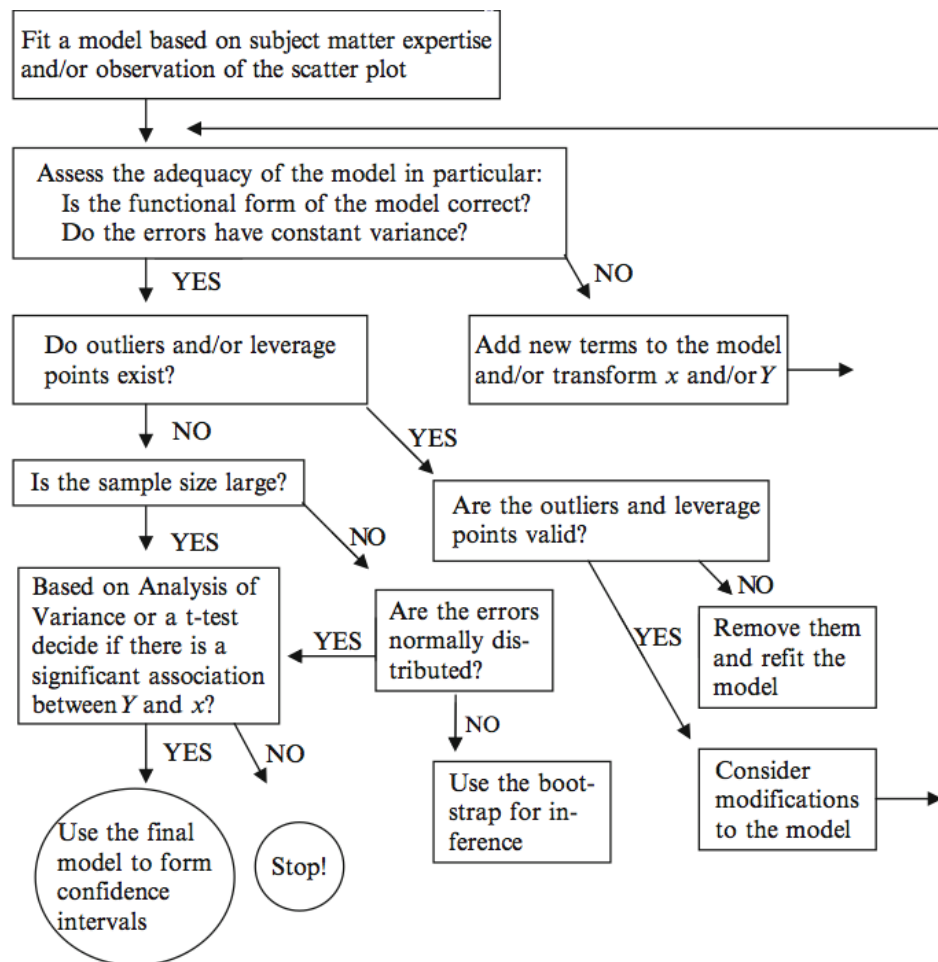
Taken from Sheather 2009: p.103

Note that this flowchart only serves the purpose of giving orientation. The suggestions may differ from those presented in this lecture. For example, if the errors do not have constant variance, the flowchart suggests the addition of new terms to the model or variable transformation. However, we have discussed in the lecture that a reformulation of the model (e.g. using a GLM) can be more appropriate. Note also that the bootstrap may not be reliable for small sample sizes, which we will discuss in the following lecture.