NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (12)

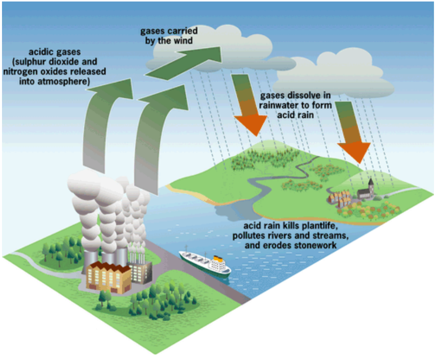
Multiple linear regression, normal equations [3.1-3.2]

Mette Langaas

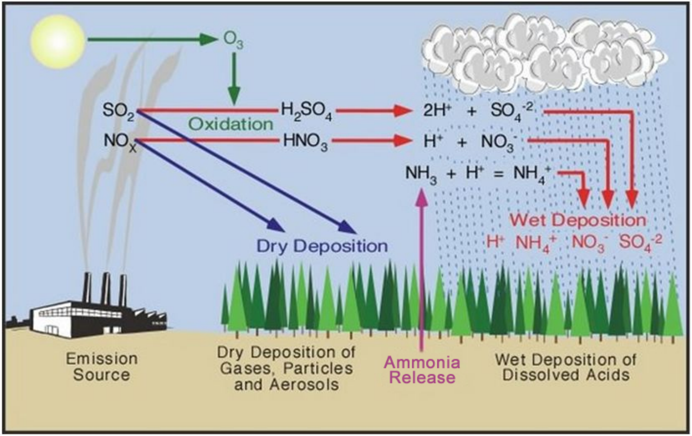
To be lectured: February 11, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Acid rain

occurs when emissions of sulfur dioxide (SO2) and oxides of nitrogen (NOx) react in the atmosphere with water, oxygen, and oxidants to form various acidic compounds. These compounds then fall to the earth in either dry form (such as gas and particles) or wet form (such as rain, snow, and fog).



Source: http://myecoproject.org/get-involved/pollution/acid-rain/



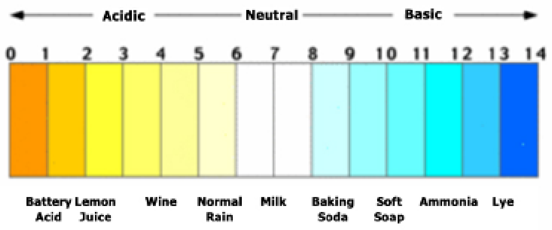
http://www.eoearth.org/view/article/149814/

Acid rain in Norwegian lakes

Measured pH in Norwegian lakes explained by content of

- x1: SO_4: sulfate (the salt of sulfuric acid),
- x2: $N0_3$: nitrate (the conjugate base of nitric acid),
- x3: Ca: calsium,
- x4: latent Al: aluminium,
- x5: organic substance,
- x6: area of lake,
- x7: position of lake (Telemark or Trøndelag),

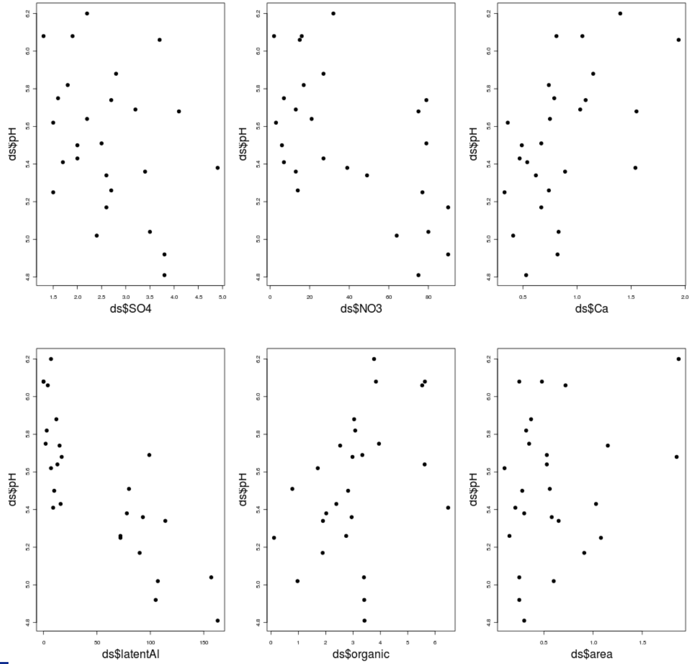
pH is a measure of the acidity of alkalinity of water, expressed in terms of its concentration of hydrogen ions. The pH scale ranges from 0 to 14. A pH of 7 is considered to be neutral. Substances with pH of less that 7 are acidic; substances with pH greater than 7 are basic.

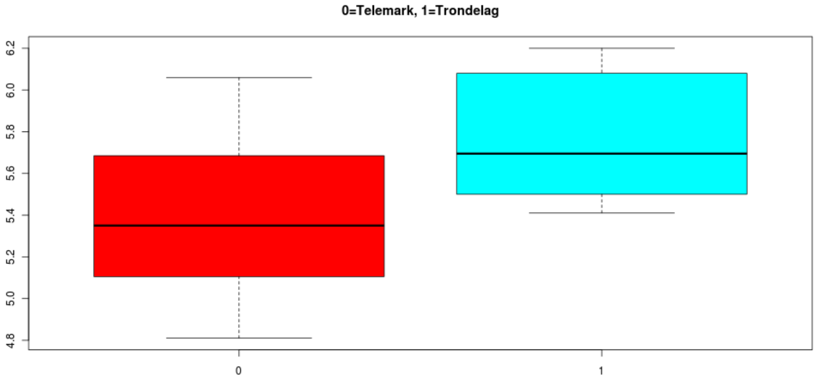


(Source: PhysicalGeography.net)

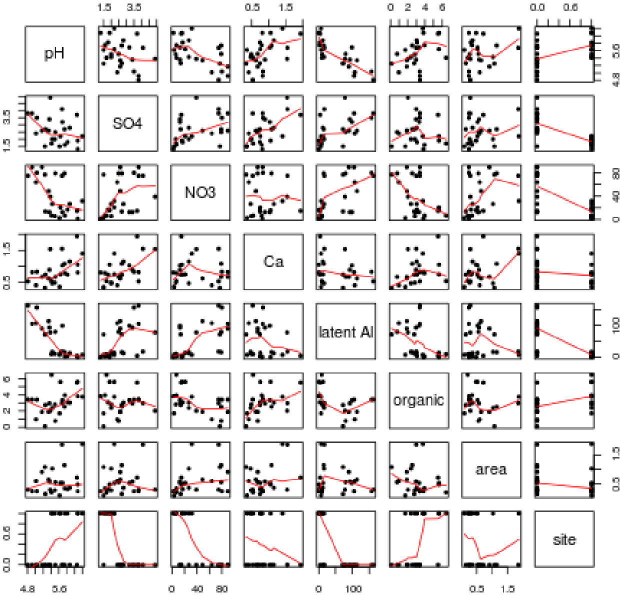
The pH scale: A value of 7.0 is considered neutral. Values higher than 7.0 are increasingly alkaline or basic. Values lower than 7.0 are increasingly acidic.

http://www.eoearth.org/view/article/149814/


ds$pH
ds$SO4
ds$NO3
ds$Ca
ds$latentAl
ds$organic
ds$area


0=Telemark, 1=Trondelag
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
0
1

Acid rain data



Part 4: Multiple linear regression

[Bingham & Fry Chapter 3]

In ch1 we studied simple linear regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{where } i = 1, \dots, n$$

$$\varepsilon_i \text{ i.i.d } N(0, \sigma^2)$$

We found $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

are the least squares (LS) and maximum likelihood (ML) estimators of α and β, and

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \underbrace{\hat{Y}_i}_{\hat{\alpha} + \hat{\beta} x_i})^2$$

is the ML estimator for σ^2.

We also saw that $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_x \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right)$ then

$$E(Y \mid X = x) = \alpha + \beta x = \left(\mu_y - \rho \frac{\sigma_y}{\sigma_x}\right) + \left(\rho \frac{\sigma_y}{\sigma_x}\right) x$$

is a linear function in x.

Now: $x_{1i} = 1 \quad \forall i = 1, .., n$ ← number of observation

to treat the intercept in the same way as the slope, and

$$Y_i = \underbrace{\beta_1 \cdot x_{i1}}_{\text{old } \alpha} + \underbrace{\beta_2 \cdot x_{i2}}_{\text{old } \beta x_i} + \dots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \dots, n$$

ε_i i.i.d $N(0, \sigma^2)$

We want vectors and matrices!

$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

Let

$$\underset{n \times 1}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

be a vector of RVs

$n = \#$ observations

$$\underset{\substack{n \times p \\ \uparrow \\ X}}{\underline{X}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & & & \vdots \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix}$$

$\uparrow$
1
1
1

is a design matrix

↑

reflecting the design of the experiment.

The x's may be set by design - as we will see in Design of Experiments (DoE) or be observed together with Y in an observational study (acid rain).

We look at X as a matrix of <u>constants</u> – not RV's

$$\underset{p\times 1}{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$ vector of regression parameters

Remark: more common to use $\beta_0, \dots, \beta_p$, but we stay with the notation of the book.

$$\underset{n\times 1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$ is a random vector of errors, where

$E(\varepsilon) = 0$ and $Var(\varepsilon_i) = \sigma^2$

and $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall\, i \neq j \Rightarrow Cov(\varepsilon) = \sigma^2 \underset{\substack{\uparrow \\ n\times n}}{I}$

The regression model:

$$\underbrace{\underset{n\times 1}{Y}}_{\text{response vector}} = \underbrace{\underset{n\times p}{X}}_{\text{design matrix}} \underbrace{\underset{p\times 1}{\beta}}_{\text{parameter vector}} + \underbrace{\underset{n\times 1}{\varepsilon}}_{\text{error vector}}, \quad \underset{n\times 1}{\varepsilon} \sim N_n(0, \sigma^2 I)$$

random vector (Y)

random vector (ε)

Questions about $\boldsymbol{X}$

1. Why do we want to assume that the design matrix $\boldsymbol{X}$ has full rank?
2. Can we find $\boldsymbol{X}^{-1}$?

Q: what are typical values of n and p?

n: 10 – 10 000 $\qquad$ p: 1 – 10

We assume that $n >> p$, i.e. the number of observations n is much larger than the number of covariates p.

And we will assume that $\underset{n \times p}{X}$ has full rank p.

Q: Why do we want to assume that X has full rank p?

We don't want to include covariates that are linear combinations of eachother – that adds no information.

Q: Can we find X^{-1}? NO. $\quad$ $\underset{n \times p}{X}$ is not quadratic when $n >> p$

$\Rightarrow$ proceed to estimate $\underline{\beta \ \& \ \sigma^2}$

Multiple linear regression model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ for } i = 1, ..., n$$

Matrix formulation:

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

$$E(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\boldsymbol{0}} \quad \text{and} \quad Cov(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\sigma^2 \boldsymbol{I}}$$

where

— $\boldsymbol{\beta}$ and σ^2 are unknown parameters and

— the design matrix $\boldsymbol{X}$ has ith row $[x_{i1}, x_{i2}, ..., x_{ip}]$.

Estimation of β

We have $Y = X\beta + \varepsilon$, with elements

$$Y_i = \underbrace{X_i}_{\text{row vector : } 1\times p} \underbrace{\beta}_{p\times 1} + \varepsilon_i \qquad \varepsilon_i \text{ i.i.d } N(0, \sigma^2)$$

and thus Y_i's are independent and $N(X_i\beta, \sigma^2)$.

ML estimation

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^2\right\}$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\Bigg\{-\frac{1}{2\sigma^2} \underbrace{\underbrace{\sum_{i=1}^{n}(y_i - X_i\beta)^2}_{(y - X\beta)^T(y - X\beta)}}_{Q(\beta)}\Bigg\}$$

Maximizing L wrt β will be the same as minimizing $Q(\beta)$ [as before...].

Remark: $e = y - Xb$

residuals $n \times 1$ — respons $n \times 1$ — estimates $n \times 1$

$$Q(b) = e^T e = (y - Xb)^T (y - Xb)$$

Task: $\frac{\partial Q(b)}{\partial b} = 0$ solve to find b.

set of p equations

$$\begin{matrix} \frac{\partial Q(b)}{\partial b_1} \\ \frac{\partial Q(b)}{\partial b_2} \\ \vdots \\ \frac{\partial Q(b)}{\partial b_p} \end{matrix}$$

Need two simple rules for derivatives wrt vectors.

$$\frac{\partial}{\partial b}\left(\underbrace{d^T b}_{\sum_{i=1}^{p} d_i b_i}\right) = \underbrace{d}_{p \times 1}$$

$$\frac{\partial}{\partial b}\left(\underbrace{b^T D b}_{\sum_{j=1}^{p}\sum_{k=1}^{p} b_j d_{jk} \cdot b_k}\right) = (D + D^T)\, b$$

if $D = D^T$

$\Rightarrow 2Db$

$$Q(b) = (y - Xb)^T (y - Xb)$$

$$= y^Ty \underbrace{- y^TXb - b^TX^Ty} + b^TX^TXb$$

$$= y^Ty - 2\underbrace{y^TX}_{a^T} b + b^T \underbrace{X^TX}_{D} b \qquad (X^TX)^T = X^TX$$

$$\frac{\partial Q(b)}{\partial b} = 0 - 2(y^TX)^T + \left(X^TX + (X^TX)^T\right)b = 0$$

$$-\cancel{2}X^Ty + \cancel{2}(X^TX)b = 0$$

$$\underline{\underline{(X^TX)\,b = X^Ty}} \qquad \text{Normal equations (NE)}$$

"$Ax = b$"

$p\times n \; n\times p$ $p\times n \; n\times 1$

$p \times p$ $(p\times 1)$ $(p\times 1)$

How can we solve this? Since X has full rank, then X^TX will also have full rank and $(X^TX)^{-1}$ exists.

$$b = (X^TX)^{-1}X^Ty$$

And the LS estimator

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

RV

Both: (X^TX) called information matrix.

Least squares estimation

- $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ with $E(\varepsilon) = \mathbf{0}$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$.
- Let $\boldsymbol{X}$ has full rank $p \leq n$.
- The Least Squares Estimate (LSE) of β is given by

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

Since X has full rank X^TX is PD i.e.

$$z^T (X^TX) z > 0 \quad \forall z \neq 0$$

Proof: book p 64-65

Near linear dependence, called multicollinearity, will make multiple linear regression numerically unstable and the interpretation of $\hat{\beta}$ will be difficult.