TMA4267 Linear Statistical Models V2017 (L9)

Part 2: Linear regression: Parameter estimation [F:3.2]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 14, 2017

The classical linear model

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

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

where

- $\boldsymbol{\beta}$ and σ^2 are unknown parameters and
- the design matrix $\boldsymbol{X}$ has full rank, with ith row $[x_{i1}x_{i2}\cdots x_{ip}]$.

Today

1. find estimator for $\boldsymbol{\beta}$,
2. find estimator for σ^2, and
3. look at two idempotent matrices $\boldsymbol{H}$ and $\boldsymbol{I}-\boldsymbol{H}$ to arrive at
4. geometric interpretation.

Rules for derivatives with respect to a vector

- Let $\boldsymbol{\beta}$ be a p-dimensional column vector of interest,
- and let $\frac{\partial}{\partial \boldsymbol{\beta}}$ denote the p-dimensional vector with partial derivatives wrt the p elements of $\boldsymbol{\beta}$.
- Let $\boldsymbol{d}$ be a p-dimensional column vector of constants and
- $\boldsymbol{D}$ be a $p \times p$ symmetric matrix of constants.

Rule 1:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{d}^T \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}(\sum_{j=1}^{p} d_j \beta_j) = \boldsymbol{d}$$

Rule 2:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T \boldsymbol{D} \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}(\sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j d_{jk} \beta_k) = 2\boldsymbol{D}\boldsymbol{\beta}$$

See Härdle and Simes (2015), page 65, Equation (2.23) and (2.24).

Two questions

Have found least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

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

and we have assumed that the rank$(\boldsymbol{X}) = p$ for $n \times p$ design matrix (where $n > p$).

- Q1: What can we say about $\boldsymbol{X}^T\boldsymbol{X}$?
- Q2: Why is the following wrong?

Using $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$,

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

The classical linear model

The model

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

is called a classical linear model if the following is true:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{I}$.
3. The design matrix has full rank $\mathrm{rank}(\boldsymbol{X}) = k + 1 = p$.

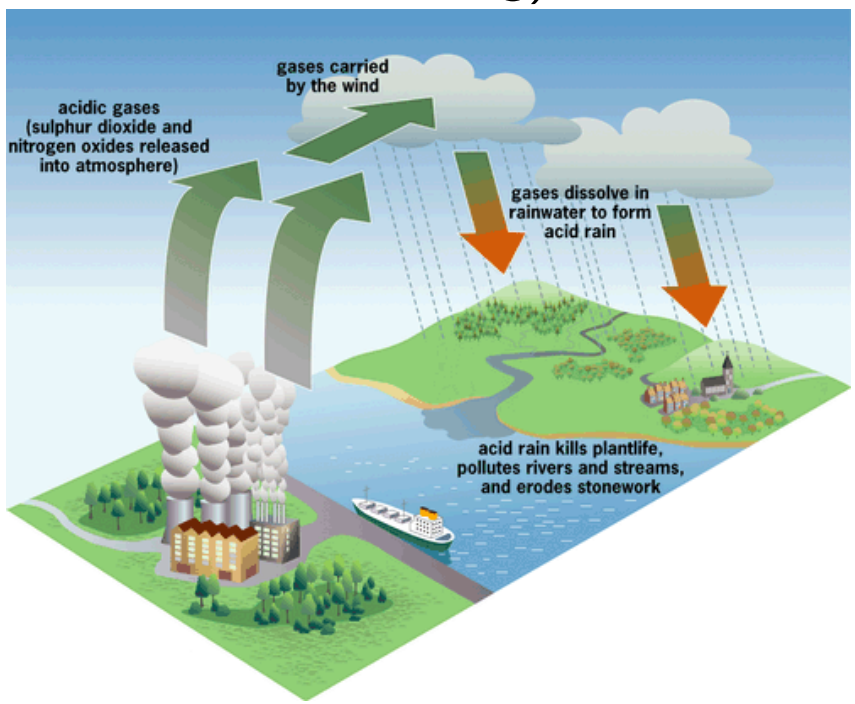
The classical *normal* linear regression model is obtained if additionally

4. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$

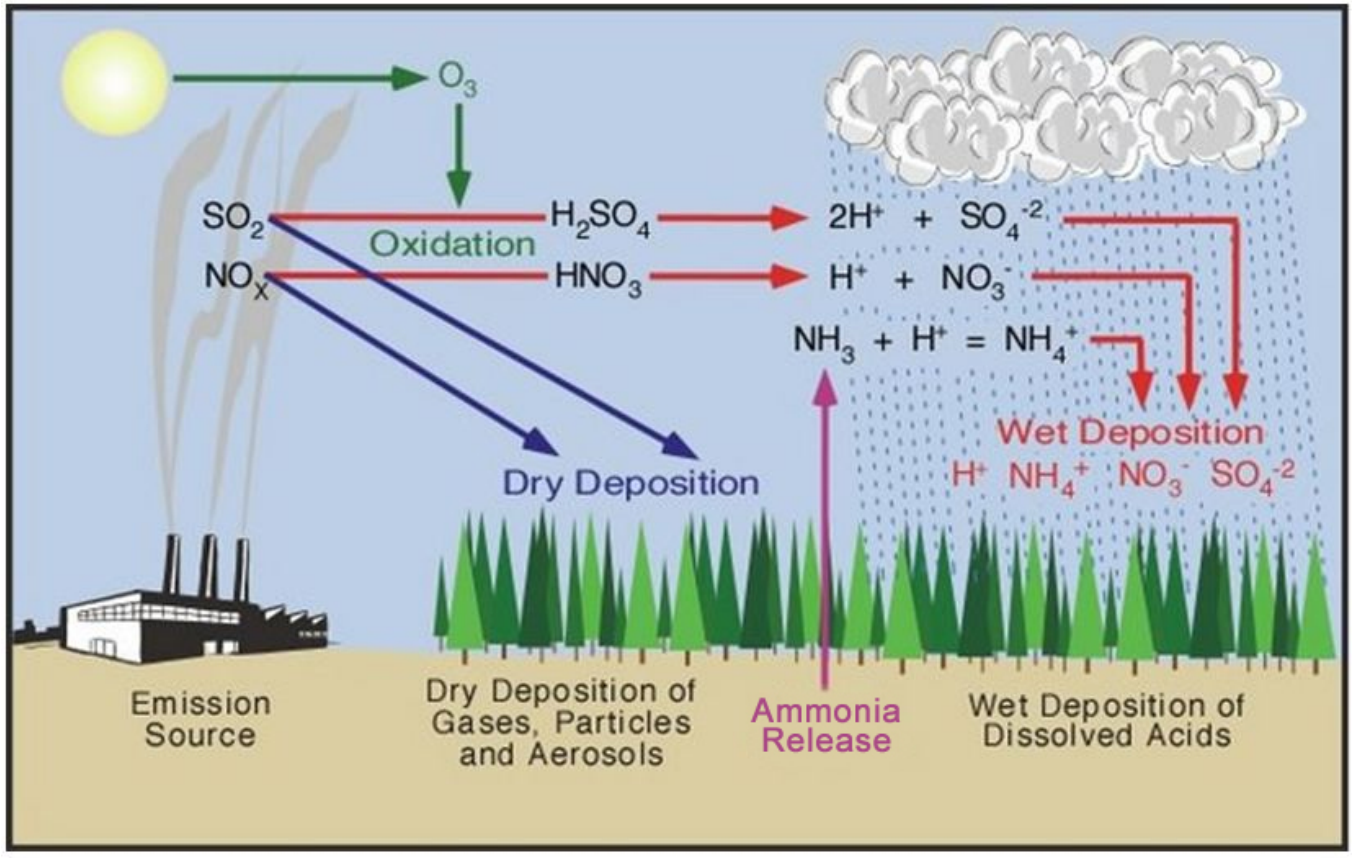
holds. For random covariates these assumptions are to be understood conditionally on $\boldsymbol{X}$.

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/



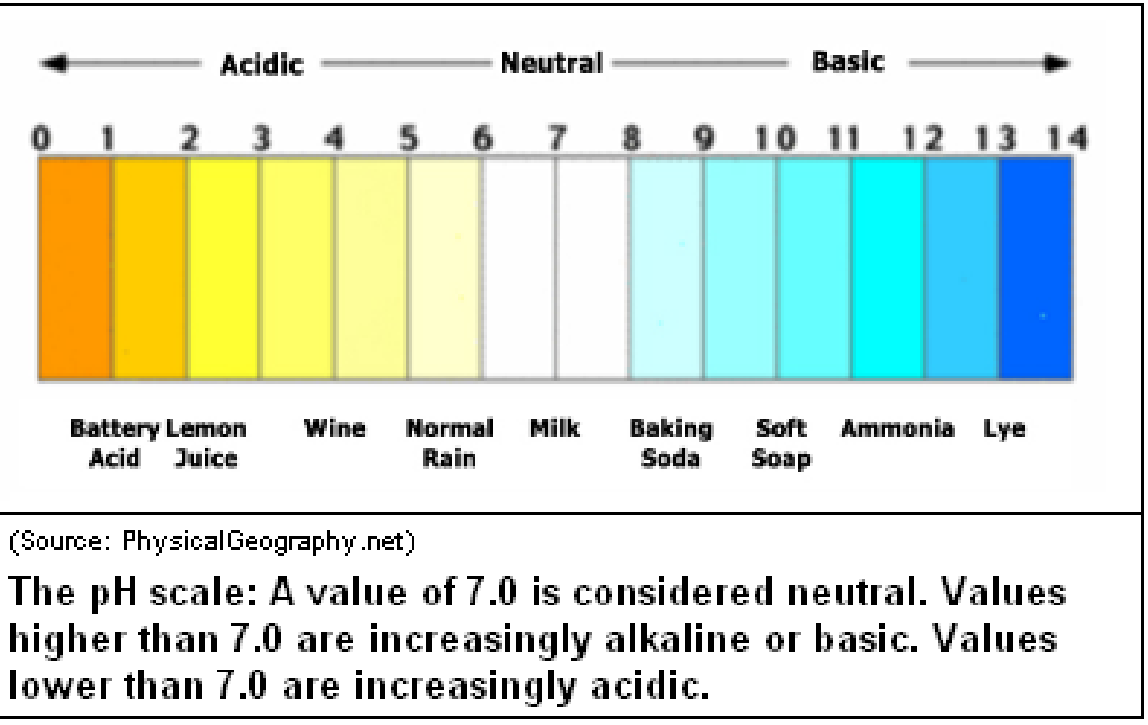
http://www.eoearth.org/view/article/149814/

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),

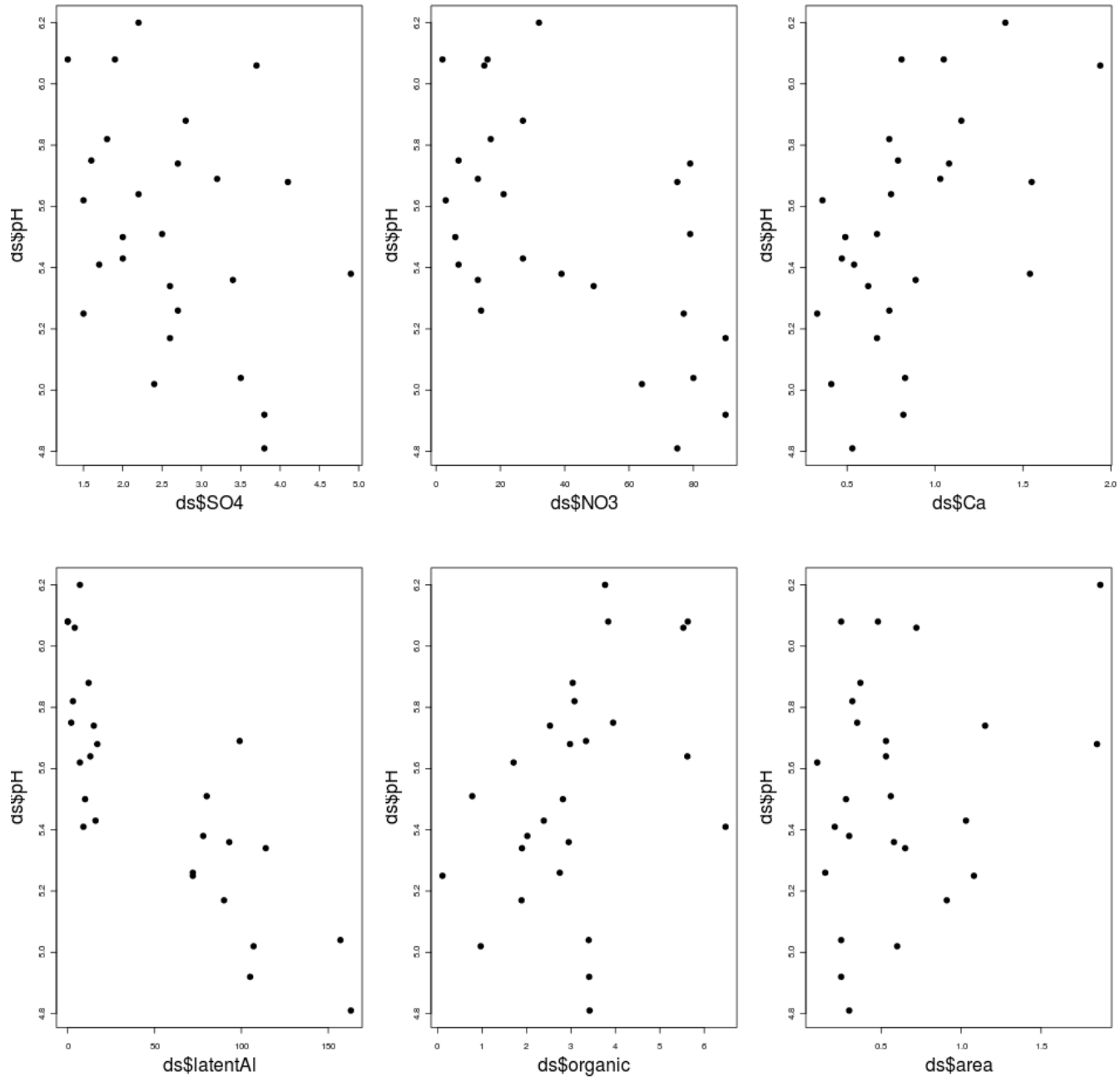
pH is a measure of the acidity of alkalinity of water, expressed in terms of its concentration of hydrogen ions. The pH scale ranges from 0 to 14. A pH of 7 is considered to be neutral. Substances with pH of less that 7 are acidic; substances with pH greater than 7 are basic.

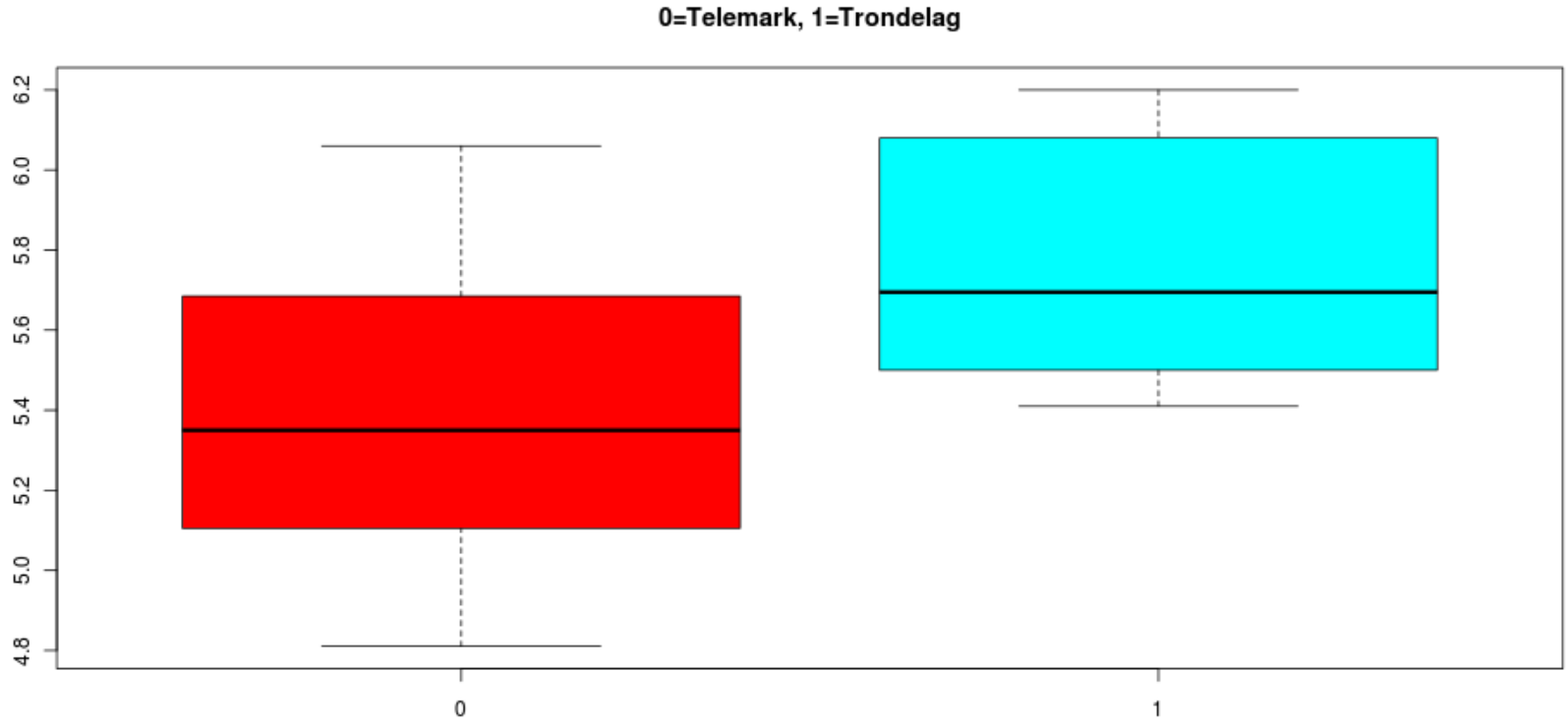


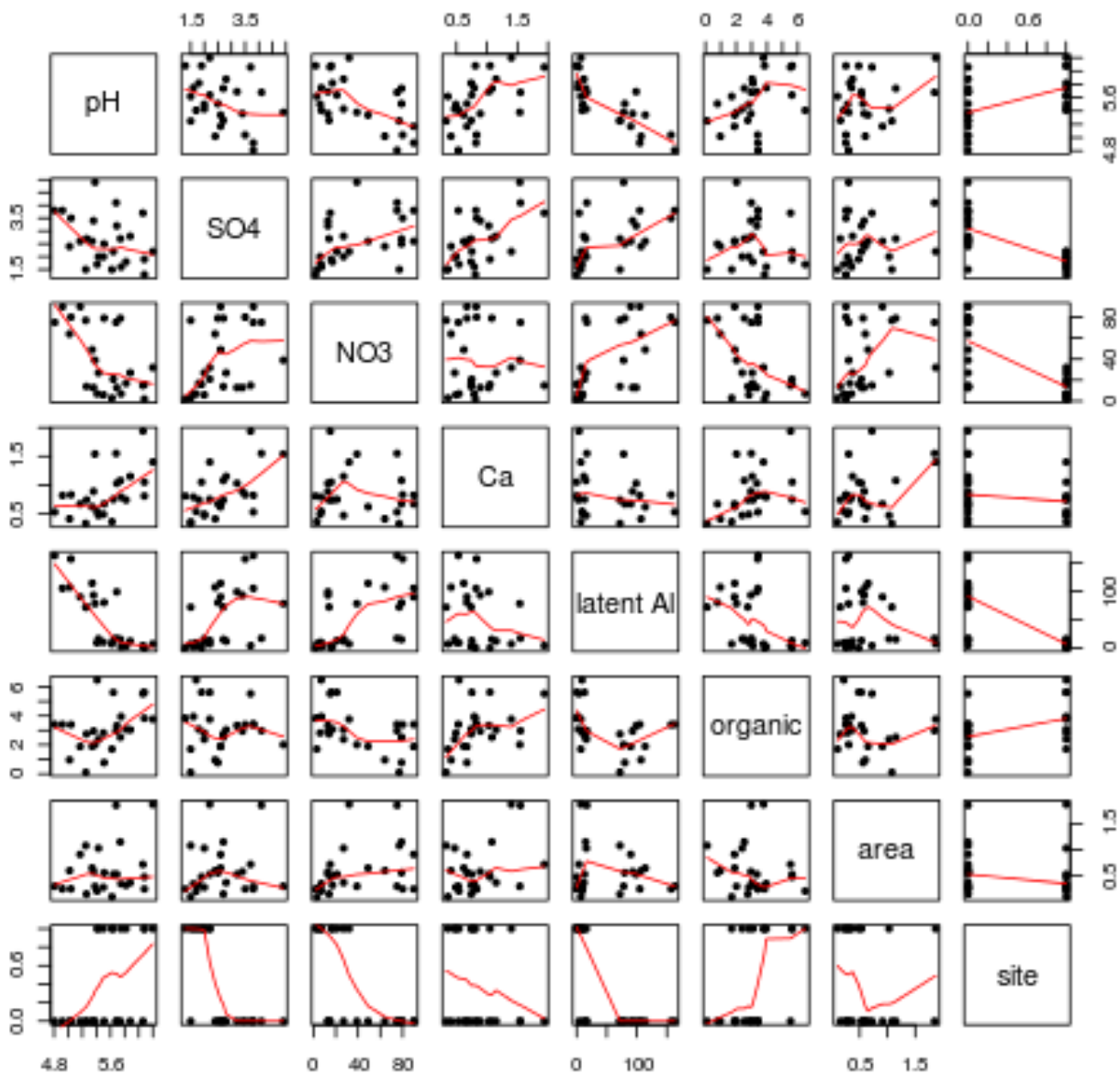
(Source: PhysicalGeography.net)

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

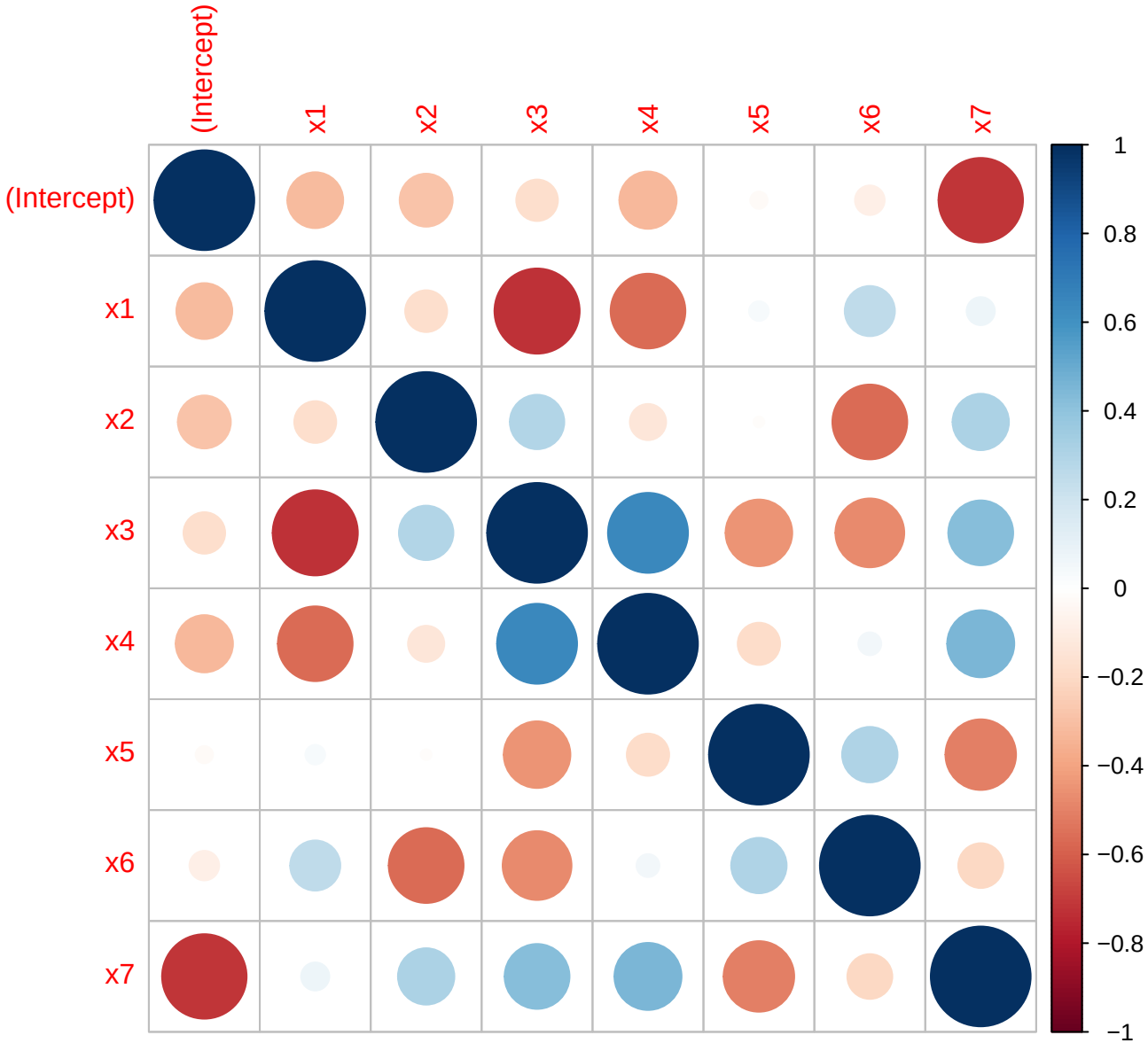

Acid rain data
pH
SO4
NO3
Ca
latent Al
organic
area
site

Output from fitting the full model in R

```
> fit=lm(y~.,data=ds)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6764334  0.1389162  40.862  < 2e-16 ***
x1          -0.3150444  0.0587512  -5.362 4.27e-05 ***
x2          -0.0018533  0.0012587  -1.472    0.158
x3           0.9751745  0.1449075   6.730 2.62e-06 ***
x4          -0.0002268  0.0010038  -0.226    0.824
x5          -0.0334242  0.0225009  -1.485    0.155
x6          -0.0039399  0.0724339  -0.054    0.957
x7           0.0888722  0.1025724   0.866    0.398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1165 on 18 degrees of freedom
Multiple R-squared:   0.93,Adjusted R-squared:  0.9027
F-statistic: 34.15 on 7 and 18 DF,  p-value: 3.904e-09
```

Question: explain how to interpret $\hat{\beta}_0$ and $\hat{\beta}_3$.


(Intercept)
x1
x2
x3
x4
x5
x6
x7
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

3.10 Asymptotic Properties of the Least Squares Estimator

1. The least squares estimator $\hat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ and the ML or REML estimator $\hat{\sigma}_n^2$ for the variance σ^2 are consistent.
2. The least squares estimator asymptotically follows a normal distribution, specifically

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{V}^{-1}).$$

That is the difference $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ normalized with $\sqrt{n}$ converges in distribution to the normal distribution on the right-hand side.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.120)

Projection matrix: definition and properties

- A matrix $\boldsymbol{A}$ is a *projection matrix* if it is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$.
- An idempotent matrix is an *orthogonal* projection matrix if, in the decomposition of a vector, $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} + (\boldsymbol{v} - \boldsymbol{A}\boldsymbol{v})$, $\boldsymbol{A}\boldsymbol{v}$ and $\boldsymbol{v} - \boldsymbol{A}\boldsymbol{v} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v}$ are always orthogonal, that is, $(\boldsymbol{A}\boldsymbol{v})^T(\boldsymbol{v} - \boldsymbol{A}\boldsymbol{v}) = 0$.
- A symmetric projection matrix is orthogonal.
- The eigenvalues of a projection matrix are 0 and 1.
- If a $(n \times n)$ symmetric projection matrix $\boldsymbol{A}$ has rank r then r eigenvalues are 1 and $n - r$ are 0.
- The trace and rank of a symmetric projection matrix are equal: $tr(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.

Results so far

- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

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

- Restricted maximum likelihood estimator for $\boldsymbol{\sigma}^2$:

$$\hat{\boldsymbol{\sigma}^2} = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{\text{SSE}}{n-p}$$

- Projection matrices: idempotent, symmetric/orthogonal:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$
$$\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

with important connection:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$$
$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{I} - \boldsymbol{H}\boldsymbol{Y}$$

Results from Mathematics 3

Best approximation theorem

The vector $\hat{\boldsymbol{Y}}$ in the column space of $\boldsymbol{X}$ that makes $|| \boldsymbol{Y} - \hat{\boldsymbol{Y}} ||$ as small as possible, is the orthogonal projection of $\boldsymbol{Y}$ on the column space of $\boldsymbol{X}$.

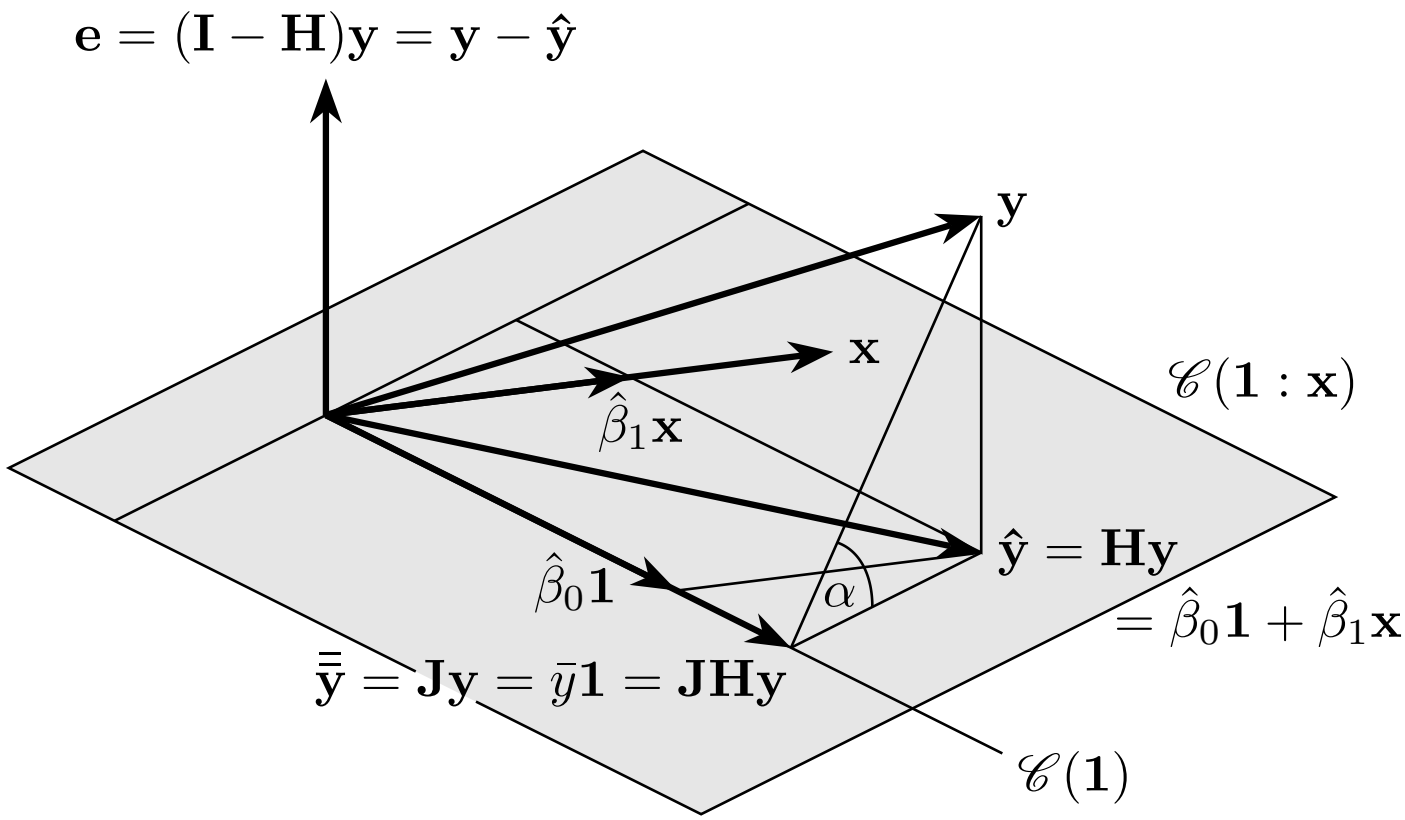
Orthogonal decomposition

We want $\hat{\beta}$ to minimize $|| \boldsymbol{Y} - \hat{\boldsymbol{Y}} ||= (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})$ (least squares principle).

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\beta}$, so $\boldsymbol{X}\hat{\beta}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$. $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ projects onto the column space of $\boldsymbol{X}$. Observe: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.

This is equivalent to observing that $\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.
$\hat{\varepsilon} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, and $\boldsymbol{I} - \boldsymbol{H}$ projects onto the space orthogonal to the column space of $\boldsymbol{X}$. Observe: $(\boldsymbol{I}\text{-}\boldsymbol{H})\boldsymbol{X}=\boldsymbol{0}$

That is, $\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$ is orthogonal to all columns of $\boldsymbol{X}$, so $\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \boldsymbol{0}$ and $\boldsymbol{X}^T\boldsymbol{X}\hat{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$.



Putanen, Styan and Isotalo: Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty, Figure 8.3.

3.7 Geometric Properties of the Least Squares Estimator

The method of least squares has the following geometric properties:

1. The predicted values $\hat{\boldsymbol{y}}$ are orthogonal to the residuals $\hat{\boldsymbol{\varepsilon}}$, i.e., $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$.
2. The columns $\boldsymbol{x}^j$ of $\boldsymbol{X}$ are orthogonal to the residuals $\hat{\boldsymbol{\varepsilon}}$, i.e., $(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0$ or $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$.
3. The average of the residuals is zero, i.e.,

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

4. The average of the predicted values $\hat{y}_i$ is equal to the average of the observed response y_i, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}.$$

5. The regression hyperplane runs through the average of the data, i.e.,

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k.$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.112)

Alternative summery of Geometry of Least Squares

- Mean response vector: $E(\boldsymbol{Y}) = \boldsymbol{X}\beta$
- As β varies, $\boldsymbol{X}\beta$ spans the model plane of all linear combinations. I.e. the space spanned by the columns of $\boldsymbol{X}$: the column-space of $\boldsymbol{X}$.
- Due to random error (and unobserved covariates), $\boldsymbol{Y}$ is not exactly a linear combination of the columns of $\boldsymbol{X}$.
- LS-estimation chooses $\hat{\beta}$ such that $\boldsymbol{X}\hat{\beta}$ is the point in the column-space of $\boldsymbol{X}$ that is closes to $\boldsymbol{Y}$.
- The residual vector $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is perpendicular to the column-space of $\boldsymbol{X}$.
- Multiplication by $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ projects a vector onto the column-space of $\boldsymbol{X}$.
- Multiplication by $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ projects a vector onto the space perpendicular to the column-space of $\boldsymbol{X}$.

Today

- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

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

  has mean $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
- For the normal model: $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.
- Asymptotic properties of the least squares estimator: normality.
- Orthogonal projection matrices $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ with geometric interpretation.

Next time: properties of residuals and $\hat{\sigma}^2$, confidence intervals and hypothesis testing for regression coefficients.