TMA4267 Linear Statistical Models V2017 [L7]

Part 2: Linear regression [F p73-86]
Model definition [F3.1], Parameters and residuals [F3.1.1], Model check [F3.1.2]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 7, 2017

Part 2: Linear regression

Part 2: Linear regression

- Fahrmeir et al (2013): Regression. Chapter 3.1, 3.2, 3.4 and required parts of 3.5 and Appendix B.

Part 3: Hypothesis testing and analysis of variance

- Fahrmeir et al (2013): Regression. Chapter 3.3 and required parts of 3.5 and Appendix B.
- Härdle et al (2015): Applied Multivariate Statistical Analysis. Chapter 8.1.1. (ANOVA).
- A short note on multiple testing (to be written).

File TMA4267Part2and3.pdf available from course www-page.

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study

B. M. Nes, I. Janszky, U. Wisløff, A. Støylen, T. Karlsen (2012) in Scandinavian Journal of Medicine and Science in Sports.

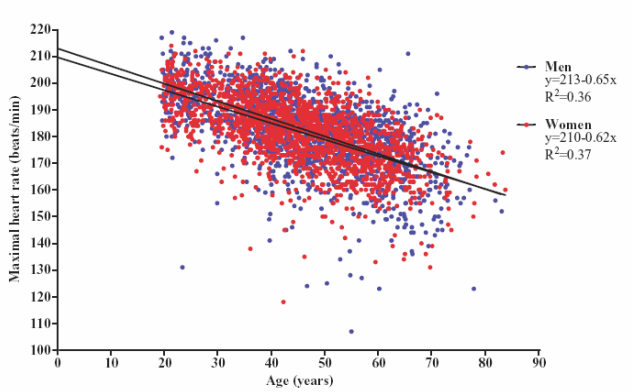
- *HRmax describes the highest heart rate achieved by a subject exercising to exhaustion and is verified by a plateau of heart rate despite increasing workload. In the literature, HRmax commonly refers to the peak heart rate at termination of a graded maximal exercise test.*
- *However, in clinical settings, a maximal exercise test is not always feasible and there is a need to predict HRmax from age prior to testing to be able to adequately assess heart rate response and relative intensity of effort at submaximal levels.*

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study

- *HRmax at a given age is frequently estimated by the "220 - age" formula.*
- *The aim of the present study was to develop a new prediction formula for HRmax through analysis of HRmax measured at VO2peak in a diverse population of 4635 healthy subjects and compare this formula with three commonly used prediction formulas. Furthermore, we wanted to investigate the relationship between HRmax and gender, physical activity status, BMI, and objectively measured aerobic fitness.*

Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study - Statistical procedures

- Only subjects that fulfilled the criteria of a maximal test, with registered maximal heart rate (HRmax), were included in the analysis (n = 3320).
- General linear modeling was used to determine the effect of age on HRmax. HRmax was entered as the dependent variable and age as the independent variable. Nonlinearity of the relationship between age and HRmax was investigated by including polynomial terms to the regression model.
- In a subsequent analysis, the effects of gender, BMI, physical activity status, and maximal oxygen uptake were examined by entering these factors as independent variables in addition to age. In further subsequent models, interaction terms were included as well to assess effect modification.
- The continuous variables were checked for normality, homogeneity of variances, and heteroscedasticity of the residuals.



Nes et al (2012): Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study. $n = 3320$ individuals.

Munich Rent Index data set

described in Fahrmeir et al (2013) on pages 19-20.

```
> library("gamlss.data")
> ds=rent99
> dim(ds)
[1] 3082    9
> colnames(ds)
[1] "rent"  "rentsqm"  "area"  "yearc"    "location" "bath"
[7] "kitchen"  "cheating" "district"
```

```
> summary(ds)
      rent           rentsqm            area           yearc
 Min.   :  40.51   Min.   : 0.4158   Min.   : 20.00   Min.   :1918
 1st Qu.: 322.03   1st Qu.: 5.2610   1st Qu.: 51.00   1st Qu.:1939
 Median : 426.97   Median : 6.9802   Median : 65.00   Median :1959
 Mean   : 459.44   Mean   : 7.1113   Mean   : 67.37   Mean   :1956
 3rd Qu.: 559.36   3rd Qu.: 8.8408   3rd Qu.: 81.00   3rd Qu.:1972
 Max.   :1843.38   Max.   :17.7216   Max.   :160.00   Max.   :1997
 location bath     kitchen  cheating    district
 1:1794   0:2891   0:2951   0: 321   Min.   : 113
 2:1210   1: 191   1: 131   1:2761   1st Qu.: 561
 3:  78                               Median :1025
                                      Mean   :1170
                                      3rd Qu.:1714
                                      Max.   :2529
```


The classical linear model

The model

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

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

1. $\mathrm{E}(\varepsilon) = \mathbf{0}$.
2. $\mathrm{Cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2\boldsymbol{I}$.
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. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$

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

Conditional mean and covariance

If we believe that the vector with elements Y and $\boldsymbol{X}$ are multivariate normal $N_{k+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we may look at the partition

$$\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_X} \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} \right)$$

The conditional distributions of the components are (multivariate) normal, with conditional mean and variance of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ are

$$\begin{aligned} \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mu_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X) \\ \mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \Sigma_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \end{aligned}$$

Observe: mean is linear in $\boldsymbol{x}$ and variance independent of $\boldsymbol{x}$.

Model assumptions for the classical linear model [F:3.1.2]

What are our model assumptions, how can we spot violations and what can we do to amend the violations.

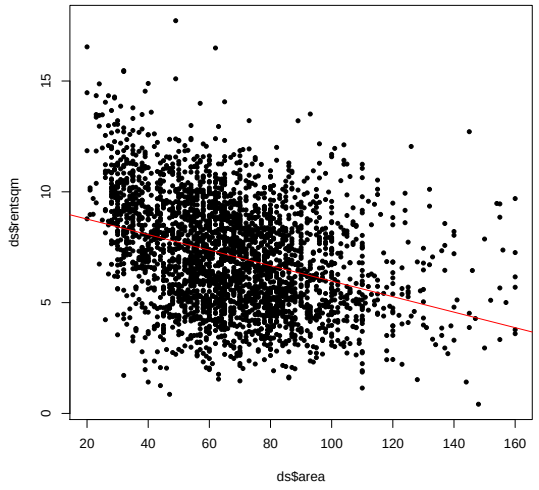
1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. Homoscedastic error variance: $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

We mainly use plots to assess this (more on model fit in F:3.4 Model choice and variable seletion)

- Covariate vs response (for each covariate)
- Covariate vs error (when we have simulated data and know the truth)
- Covariate vs residual (estimated error),
- Predicted response vs residual (to be popular later).

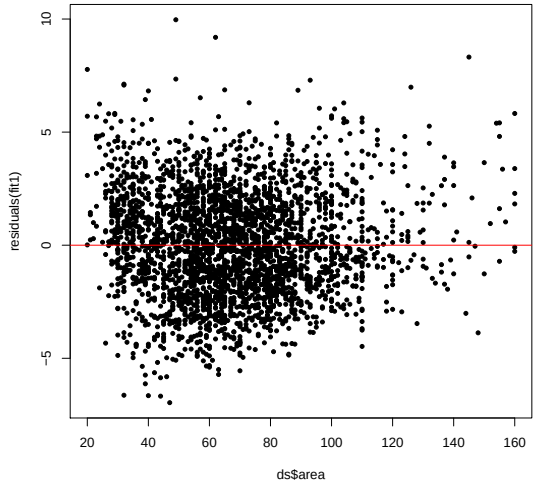
Linearity of covariates: Covariate vs. response

Munich Rent Index: area vs rentsqm



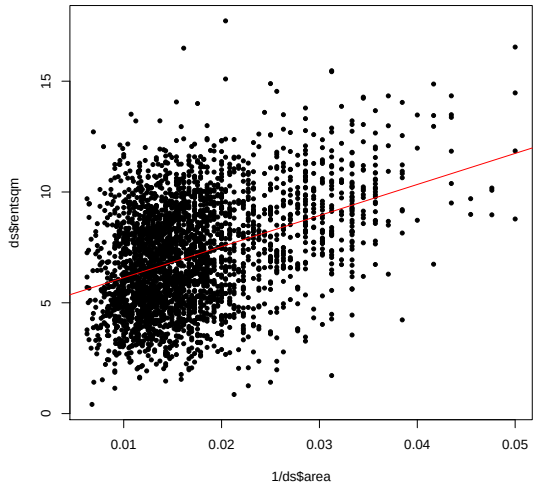
Linearity of covariates: Covariate vs. residual (residual plot)

Munich Rent Index: area vs residual



Linearity of covariates: Transformed covariate vs. response

Munich Rent Index: 1/area vs rentsqm



3.2 Modeling Nonlinear Covariate Effects Through Variable Transformation

If the continuous covariate z has an approximately nonlinear effect $\beta_1 f(z)$ with known transformation f, then the model

$$y_i = \beta_0 + \beta_1 f(z_i) + \ldots + \varepsilon_i$$

can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \varepsilon_i,$$

where $x_i = f(z_i) - \bar{f}$. By subtracting

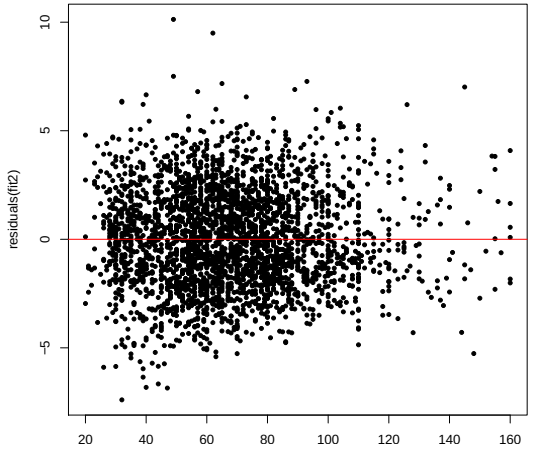
$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(z_i),$$

the estimated effect $\hat{\beta}_1 x$ is automatically centered around zero. The estimated curve is best interpreted by plotting $\hat{\beta}_1 x$ against z (instead of x).

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.94)

Linearity of covariates: Transformed covariate vs. residual (residual plot)

Munich Rent Index: 1/area vs residual



3.3 Modeling Nonlinear Covariate Effects Through Polynomials

If the continuous covariate z has an approximately polynomial effect $\beta_1 z + \beta_2 z^2 + \ldots + \beta_l z^l$ of degree l, then the model

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_l z_i^l + \ldots + \varepsilon_i$$

can be transformed into the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_l x_{il} + \ldots + \varepsilon_i,$$

where $x_{i1} = z_i$, $x_{i2} = z_i^2$, ..., $x_{il} = z_i^l$.
The centering (and possibly orthogonalization) of the vectors $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})'$, $j = 1, \ldots, l$, to $\boldsymbol{x}^1 - \bar{\boldsymbol{x}}_1, \ldots, \boldsymbol{x}^l - \bar{\boldsymbol{x}}_l$ with the mean vector $\bar{\boldsymbol{x}}_j = (\bar{x}_j, \ldots, \bar{x}_j)'$ facilitates interpretation of the estimated effects. A graphical illustration of the estimated polynomial is a useful way to interpret the estimated effect of z.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.95)

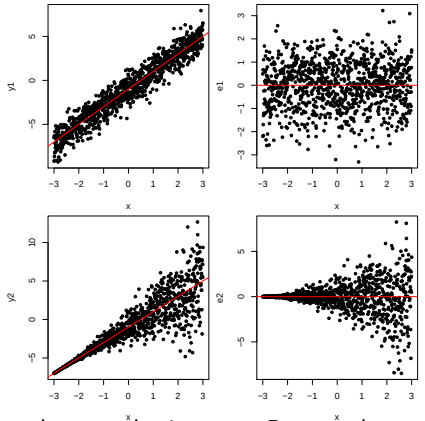
Homoscedastic errors

```
n=1000
x=seq(-3,3,length=n)
beta0=-1
beta1=2
xbeta=beta0+beta1*x
sigma=1
e1=rnorm(n,mean=0,sd=sigma)
y1=xbeta+e1
ehat1=residuals(lm(y1~x))
plot(x,y1,pch=20)
abline(beta0,beta1,col=1)
plot(x,e1,pch=20)
abline(h=0,col=2)
```


Heteroscedastic errors

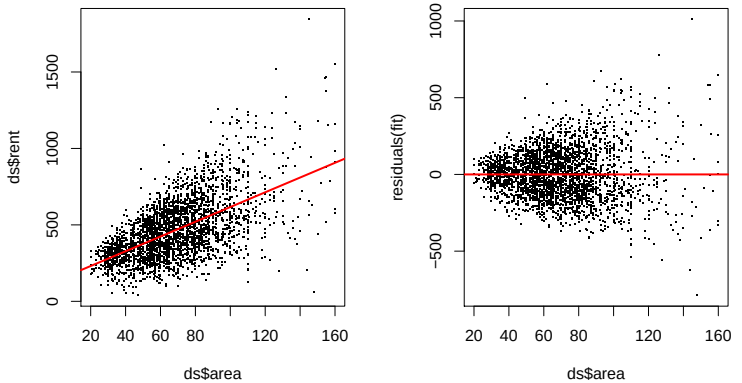
```
sigma=(0.1+0.3*(x+3))^2
e2=rnorm(n,0,sd=sigma)
y2=xbeta+e2
ehat2=residuals(lm(y2~x))
plot(x,y2,pch=20)
abline(beta0,beta1,col=2)
plot(x,e2,pch=20)
abline(h=0,col=2)
```


Homo- and heteroscedastic errors



Top: homoscedastic errors. Bottom: heteroscedastic errors. Right: x vs y. Left: x vs error. Example from Fahrmeir et al (2013): Regression. Springer. (p.79). R code from TMA4267 lectures tab.

Homoscedastic errors?



Left: area vs rent, right: area vs residuals. Fahrmeir et al (2013): Regression. Springer. (p.80). R code from TMA4267 lectures tab.

Today

- Normal linear model: implication for Y.
- Model parameters $\boldsymbol{\beta}, \sigma^2$, parameter estimators $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$, residuals $\hat{\varepsilon} = Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.
- Model assumptions.
- Next: covariates- how to include in linear regression, and then parameter estimation.

PART 2:
LINEAR REGRESSION

TMA4267 L7
09.02.2017

Model definition [FF31.0]

Y = variable of primary interest (response, dependent variable)

$x_1, x_2, \ldots, x_k$ = regressors, explanatory variables, independent variables, covariates

Assumptions:

$$Y = \underbrace{f(x_1, x_2, \ldots, x_k)}_{\text{systematic component}} + \underbrace{\varepsilon}_{\text{error term}}$$

1) Systematic component is a linear combination of the covariates.

$$f(x_1, x_2, \ldots, x_k) = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{\text{multiple}}$$

simple

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad f(x) = x^T\beta$$

$p \times 1$, $p = k+1$

2) Additive error $Y = x^T\beta + \varepsilon$

Restrictive? Maybe → transformations?

Data and design matrix

We collect independent data (Y_i, x_i) for $i = 1, \ldots, n$ (pairs)

response

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$(1, x_{i1}, x_{i2}, \ldots, x_{ik})^T$

$$X_{n\times p} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}$$

← individual/observational unit

↑ covariate

Design matrix. We will assume that $n >> p$.

and $\text{rank}(X) = p$

n = number of observations

p = number of covariates + 1 (intercept)

Q: What can make $\text{rank}(X) < p$?

Ex: Munich rent index: $n = 3082$

Y = rent or rent pr sq.m → $Y = \begin{bmatrix} 526 \\ 9.41 \\ \vdots \end{bmatrix}$

$$X = \begin{bmatrix} 1 & 80 & 1970 & 1 & 0 & 0 & 0 & 916 \\ 1 & 91 & 1941 & 1 & 0 & 0 & 1 & 206 \\ \vdots & \vdots & \vdots & 2 & 1 & 1 & 1 & \vdots \\ 1 & \vdots & \vdots & 3 & \vdots & 1 & \vdots & \vdots \\ & & & & & 1 & 1 & \\ & & & & & & 2 & \end{bmatrix}$$

area; year of construction; location (both, with bath, with c. heating); district

X is chosen so that we have a good model.

The classical linear model

$Y_i = x_i^T\beta + \varepsilon_i$ (one person)

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

ε: unobserved random vector

1) $E(\varepsilon) = 0$

2) $\text{Cov}(\varepsilon) = E(\varepsilon\varepsilon^T) - \underbrace{E(\varepsilon)}_{0}\underbrace{E(\varepsilon)^T}_{0} = \sigma^2 I$

$\text{Var}(\varepsilon_i) = \sigma^2$ for all i ← homoscedastic error

$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ ← uncorrelated errors

3) The design matrix has rank $\text{rank}(X) = k+1 = p$

4) If we in addition assume that $\varepsilon \sim N_n(0, \sigma^2 I)$ then we have a normal linear regression model.

What does the imply for the distribution of Y?

$Y \sim N_n$ since $Y = \underbrace{\mathbb{X}\beta}_{\text{constant}} + \underbrace{\varepsilon}_{\sim N_n}$, and

$E(Y) = E(\mathbb{X}\beta + \varepsilon) = \mathbb{X}\beta + \underbrace{E(\varepsilon)}_{0} = \mathbb{X}\beta$

$Cov(Y) = Cov(\mathbb{X}\beta + \varepsilon) = 0 + Cov(\varepsilon) = \sigma^2 I$

$$Y \sim N_n(\mathbb{X}\beta, \sigma^2 I)$$

The covariates $\mathbb{X}$ may be regarded as random variables, and then the assumptions (1)+(2) are made conditional on $\mathbb{X} = x$, so
$E(\varepsilon \mid \mathbb{X} = x) = 0$ and $Cov(\varepsilon \mid \mathbb{X} = x) = \sigma^2 I$

If we instead assume that

$$\begin{bmatrix} Y \\ \mathbb{X}_1 \\ \mathbb{X}_2 \\ \vdots \\ \mathbb{X}_n \end{bmatrix} \sim N_{n+1} \Rightarrow \begin{array}{l} E(Y \mid \mathbb{X} = x) = \text{linear in } x \\ Var(Y \mid \mathbb{X} = x) = \text{not dependent on } x \end{array}$$

Be aware: don't mix errors ε (unobserved) with residuals $\hat{\varepsilon}$ ("observed").

The residuals will be used to assess model assumptions as proxies for the errors.

Model assumption [FS 1.2]

1) Linearity of covariates $Y = \mathbb{X}\beta + \varepsilon$

$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ ok
$\beta_0 + \beta_1 x_{i1}^2 + \varepsilon_i$ ok
$\beta_0 + \beta_1 \log(x_{i1}) + \varepsilon_i$ ok
$\beta_0 + \beta_1 \sin(\beta_2 x_{i1}) + \varepsilon_i$ not ok

If the relationship between Y and x_1 is nonlinear $\Rightarrow$ ok to use polynomial (or similar) in x_1. More advanced: nonparametric regression

2) Homoscedastic error variance: $Cov(\varepsilon) = \sigma^2 I$

Need to check that $Var(\varepsilon)$ does not vary systematic across observations

Look at covariate vs residuals → trend? fan out, fan in.
Solution if problem: If we knew $Cov(\varepsilon) = \Sigma$

Model parameters, estimates and residuals [FS 1.1]

$Y = \mathbb{X}\beta + \varepsilon$, $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$

The model parameters are β ($p \times 1$), σ^2
(the unknown)

We will develop estimators:
$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$ (LS) by least squares and maximum likelihood.
$\hat{\sigma}^2 = \frac{1}{n-p}(Y - \mathbb{X}\hat{\beta})^T(Y - \mathbb{X}\hat{\beta})$ (ML) by restricted maximum likelihood (REML)

Further: Y is a random vector with mean $\mathbb{X}\beta$, and estimator for $E(Y) = \mathbb{X}\beta$ is $\hat{Y} = \mathbb{X}\hat{\beta}$.

The error ε is a random vector with $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$, but ε is not observed.

$Y = \underbrace{\mathbb{X}\beta}_{\text{unknown}} + \varepsilon$ ← unobserved
observed, observed

Our best guess for the error is the residual vector $(\hat{\varepsilon}, e)$

$\hat{\varepsilon} = Y - \hat{Y} = Y - \mathbb{X}\hat{\beta}$

So, the residuals can be calculated, and we may think of the residuals as predictions of the errors

We may use a so-called general linear model, with weighted least squares (Rec Ex 3. P4), but Σ is in general unknown.

Remaining: 3) uncorrelated errors
4) additive error.

TMA4267 Linear Statistical Models V2017 (L8)

Part 2: Linear regression: Modelling the effects of covariates [F:3.1.3] Parameter estimation: Estimator for β [F:3.2.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 10, 2017

Model assumptions for the classical linear model [F:3.1.2]

What are our model assumptions, how can we spot violations and what can we do to amend the violations.

1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$
2. Homoscedastic error variance: $Var(\varepsilon_i) = \sigma^2$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$

We mainly use plots to assess this (more on model fit in F:3.4 Model choice and variable seletion)

- Covariate vs response (for each covariate)
- Covariate vs error (when we have simulated data and know the truth)
- Covariate vs residual (estimated error),
- Predicted response vs residual.

The classical linear model

The model

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

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

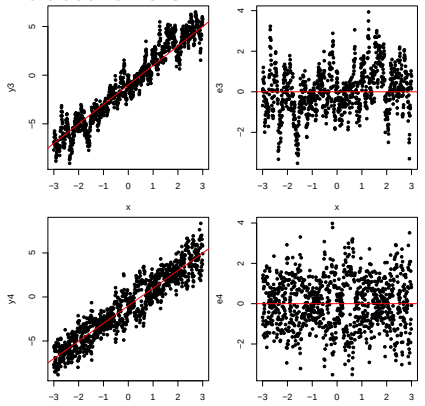
1. $\mathrm{E}(\varepsilon) = \boldsymbol{0}$.
2. $\mathrm{Cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2\boldsymbol{I}$.
3. The design matrix has full rank $\mathrm{rank}(\boldsymbol{X}) = k + 1 = p$.

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

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

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

Uncorrelated errors?



Top: positively autocorrelated errors. Bottom: negatively correlated errors. Right: x vs y. Left: x vs error. Example from Fahrmeir et al (2013): Regression. Springer. (p.81). R code from TMA4267 lectures tab.

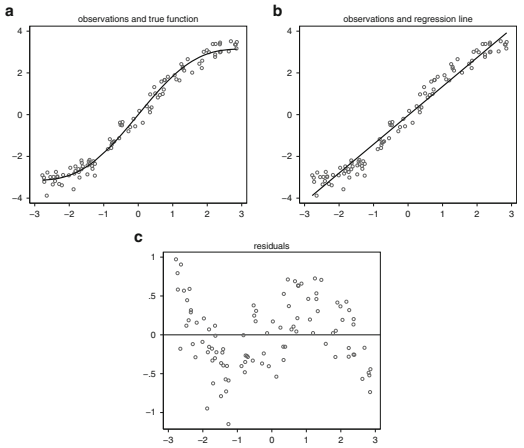
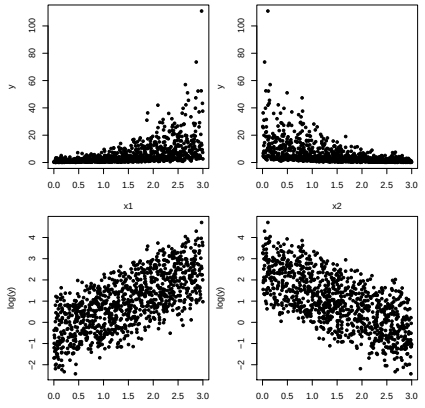


Fig. 3.4 Illustration for correlated residuals when the model is misspecified: Panel (**a**) displays (simulated) data based on the function $\mathrm{E}(y_i \mid x_i) = \sin(x_i) + x_i$ and $\varepsilon_i \sim \mathrm{N}(0, 0.3^2)$. Panel (**b**) shows the estimated regression line, i.e., the nonlinear relationship is ignored. The corresponding residuals can be found in panel (**c**)

Fahrmeir et al (2013): Regression. Springer. (p.82)

Multiplicative errors



Top: x1 and x2 vs y. Bottom: x1 and x2 vs log(y). Example from Fahrmeir et al (2013): Regression. Springer. (p.85). R code from TMA4267 lectures tab.

Multiplicative errors

```
x1=runif(n,0,3)
x2=runif(n,0,3)
e=rnorm(n,0,0.4)
y=exp(1+x1-x2+e)
plot(x1,y,pch=20)
plot(x2,y,pch=20)
plot(x1,log(y),pch=20)
plot(x2,log(y),pch=20)
```


Covariates - how to include in the linear regression?

1. Continuous covariates: as is, transformed or using polynomials.
2. Categorical covariates: dummy variable or effect coding.
3. Interactions between covariates.

Munich rent index data

```
> colnames(ds)
[1] "rent" "rentsqm" "area" "yearc" "location" "bath"
[7] "kitchen"  "cheating" "district"
> apply(ds[,1:4],2,summary)
           rent rentsqm   area yearc
Min.      40.51  0.4158  20.00  1918
1st Qu.  322.00  5.2610  51.00  1939
Median   427.00  6.9800  65.00  1959
Mean     459.40  7.1110  67.37  1956
3rd Qu.  559.40  8.8410  81.00  1972
Max.    1843.00 17.7200 160.00  1997
> unlist(apply(ds[,5:8],2,table))
location.1 location.2 location.3  bath.0  bath.1 kitchen.0
      1794       1210         78    2891     191    2951
 kitchen.1 cheating.0 cheating.1
       131        321       2761
```


Linear coding

```
> fit1=lm(rentsqm~as.numeric(location),data=ds)
> summary(fit1)
Call:
lm(formula = rentsqm ~ as.numeric(location), data = ds)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.54390    0.12368  52.911  < 2e-16 ***
as.numeric(location)  0.39312    0.08016   4.904 9.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.427 on 3080 degrees of freedom
Multiple R-squared:  0.007748,Adjusted R-squared:  0.007425
F-statistic: 24.05 on 1 and 3080 DF,  p-value: 9.878e-07
```


How to code categorical covariates: rentsqm vs location with linear coding

- Location average=1, good=2 and top=3, and regression model
$$\text{rentsqm}_i = \beta_0 + \beta_1 \text{location}_i + \varepsilon_i$$
- Parameter estimate: $\hat{\beta}_1 = 0.39$. What does that mean?
  - Flat of average location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$
  - Flat of good location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 2$
  - Flat of top location: $\widehat{\text{rentsqm}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3$
- What is the difference in predicted rentsqm between top and good location, and between good and average location?
- So, the difference between a top and a good location is the same as the difference between good and average. Is this what we want?

rentsqm vs location with dummy variable coding

$$\text{aloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not average} \\ 1 & \text{location}_i \text{ is average} \end{cases}$$

$$\text{gloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not good} \\ 1 & \text{location}_i \text{ is good} \end{cases}$$

$$\text{tloc}_i = \begin{cases} 0 & \text{location}_i \text{ is not top} \\ 1 & \text{location}_i \text{ is top} \end{cases}$$

$$\text{rentsqm}_i = \beta_0 + \beta_1 \text{aloc}_i + \beta_2 \text{gloc}_i + \beta_3 \text{tloc}_i + \varepsilon_i$$

- Write down the design matrix for this regression model, when we have 1794 flats with average location, 1210 with good and 78 with top location.
- What is the rank of this design matrix?
- Is there a problem, and a solution?

3.4 Dummy Coding for Categorical Covariates

For modeling the effect of a covariate $x \in \{1, \ldots, c\}$ with c categories using dummy coding, we define the $c - 1$ dummy variables

$$x_{i1} = \begin{cases} 1 & x_i = 1, \\ 0 & \text{otherwise}, \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c - 1, \\ 0 & \text{otherwise}, \end{cases}$$

for $i = 1, \ldots, n$, and include them as explanatory variables in the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i.$$

For reasons of identifiability, we omit one of the dummy variables, in this case the dummy variable for category c. This category is called reference category. The estimated effects can be interpreted by direct comparison with the (omitted) reference category.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.97)

Dummy coding via contr.treatment

```
> contrasts(ds$location)=contr.treatment(3)
> fit2=lm(rentsqm~location,data=ds)
> summary(fit2)
Call:
lm(formula = rentsqm ~ location, data = ds)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.95654    0.05728 121.456  < 2e-16 ***
location2    0.31570    0.09025   3.498 0.000475 ***
location3    1.21579    0.28060   4.333 1.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.426 on 3079 degrees of freedom
Multiple R-squared:  0.008867,Adjusted R-squared:  0.008223
F-statistic: 13.77 on 2 and 3079 DF,  p-value: 1.109e-06
```


Effect coding via contr.sum

```
> contrasts(ds$location)=contr.sum(3)
> fit3=lm(rentsqm~location,data=ds)
> summary(fit3)
Call:
lm(formula = rentsqm ~ location, data = ds)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.46704    0.09638  77.477  < 2e-16 ***
location1   -0.51050    0.10189  -5.010 5.75e-07 ***
location2   -0.19479    0.10445  -1.865   0.0623 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.426 on 3079 degrees of freedom
Multiple R-squared:  0.008867,Adjusted R-squared:  0.008223
F-statistic: 13.77 on 2 and 3079 DF,  p-value: 1.109e-06
```


Response: birth weight

Covariates: glucose level of mother and BMI of mother.

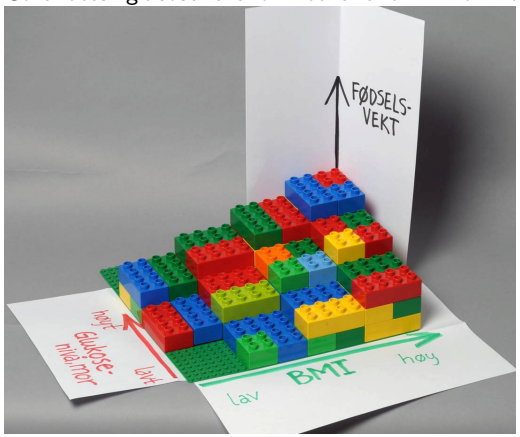


Figure from Kathrine Frey Frøslie.

Response: birth weight

Covariates: glucose level of mother and BMI of mother - with interaction.

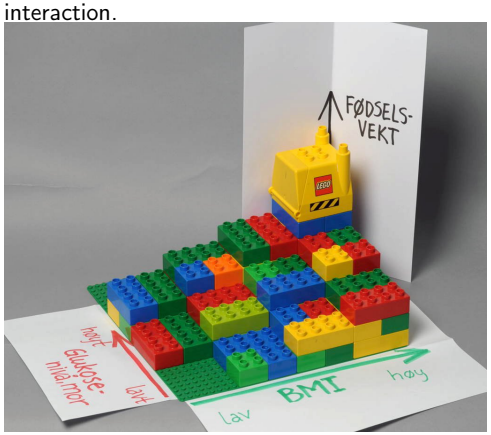


Figure from Kathrine Frey Frøslie.

Today

- Model assessment: residual plots.
- Covariates: how to include in linear regression?
- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$.

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

The classical linear model

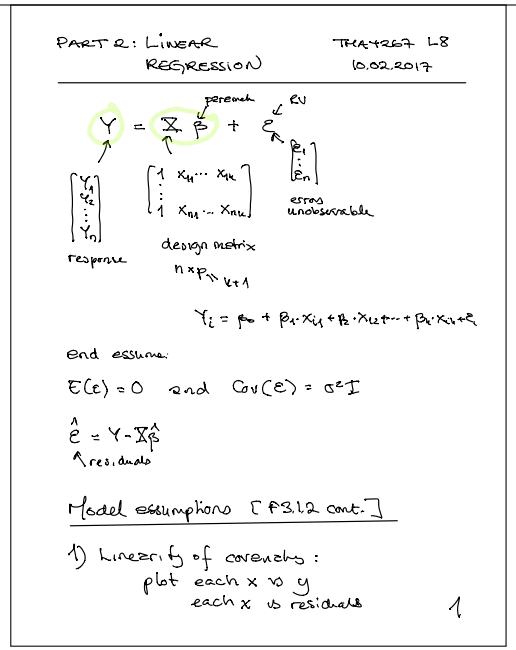
$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

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

where

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

Next: find the estimator $\hat{\boldsymbol{\beta}}$.


PART 2: LINEAR REGRESSION

TMA4267 L8
10.02.2017

response

design matrix
$n \times p_{+1}$

error
unobservable

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$

End assume:

$E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$

$\hat{\varepsilon} = Y - X\hat{\beta}$
↑ residuals

Model assumptions [F3.12 cont.]

1) Linearity of covariates:
plot each x vs y
each x vs residuals

1

2) Homoscedastic error var.enco

Plot x vs $\hat{\varepsilon}$
$\hat{y}$ vs $\hat{\varepsilon}$

3) Uncorrelated errors: $Cov(\varepsilon_i, \varepsilon_j) = 0$

If data comes from measurements in time or space errors may be autocorrelated.

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + u_i \qquad u_i \sim N(0,1)$$

(ε_{i-1}: popular time series model)

But, autocorrelation may also be due to misspecification of the model, e.g. a missing (unobserved) covariate, or by modelling a linear instead of a nonlinear relationship.

4) Additivity of errors: $Y = X\beta + \varepsilon$

$$Y = \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon\}$$
$$= \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdots \exp(\beta_k x_k) \cdot \exp(\varepsilon)$$

has multiplicative errors.
Transforming the model (Y) using ln gives additive errors

$$X = \underset{3082 \times 4}{\begin{bmatrix} \beta_0 & aloc & gloc & tloc \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}} \quad \begin{matrix} \text{average} \cdot 1794 \\ \text{good} \cdot 1210 \\ \text{top} \cdot 78 \end{matrix}$$

Problem: $rank(X) = 3$, not unique solution
How to solve:
a) not include intercept
b) omit one of the dummy variables, the omitted category is called the reference category

Ex: let average be omitted: contr.treatment
$rentsqm_i = \beta_0 + \beta_2 gloc_i + \beta_3 tloc_i + \varepsilon_i$

average loc$_i$: $\widehat{rentsqm} = \hat{\beta}_0$
good : $\widehat{rentsqm} = \hat{\beta}_0 + \hat{\beta}_2$
top : $\widehat{rentsqm}$ $\hat{\beta}_0 + \hat{\beta}_3$

c) add a restriction: sum-to-zero
$\sum_{j=1}^{3} \beta_j = 0$ R. contr.sum

Effect coding (important in Part 3&4)
We have $X_1 = \begin{cases} 1 & \text{average} \\ 2 & \text{good} \\ 3 & \text{top} \end{cases}$

Covariates: how to include in the linear regression [F3.1.3]

1) Continuous x_1
* Y vs x_1 linear
* transform x_1 ($\ln x_1$, $\frac{1}{x_1}$, $\sqrt{x_1}$)
* use polynomial in x_1.

2) Categorical ← nominal (green, red, blue) ← district
ordinal (good, average, top) ← location

Ex: Y = rentsqm
X_1 = location (1: average, 2: good, 3: top)
$\hat{\beta}_1 = 0.39$ → the effect of good location is twice the effect of average location

a) Linear coding

b) Dummy variable coding
Ex: location → aloc 1 ← 1794, gloc 1 ← 1210, tloc 1 ← 78 (0 otherwise), 3082
(see slide)

$$z_1 = \begin{cases} 1 & \text{if } x_1 = 1 \\ -1 & \text{if } x_1 = 3 \\ 0 & \text{else } (x_1 = 2) \end{cases} \qquad z_2 = \begin{cases} 1 & \text{if } x_1 = 2 \\ -1 & \text{if } x_1 = 3 \\ 0 & \text{else } (x_1 = 1) \end{cases}$$

$y_i = \alpha_0 + \alpha_1 z_1 + \alpha_2 \cdot z_2 + \varepsilon$

$\alpha_3 = -\alpha_1 - \alpha_2$.

3) Interactions

Is the effect (on Y) of a change in x_1 dependent on the value of another covariate x_2?

Lego: Y = birth weight child
x_1 = glucose level of mother
x_2 = BMI of mother ($\frac{kg}{m^2}$)

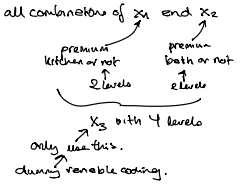
Figure 1: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Figure 2: High glucose will have a different effect on birth weight when BMI is low compared to when BMI is high. ⇒ we have an interaction between x_1 and x_2.

a) Continuous x_1 and x_2
Simplest solution: $Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}}_{f(x_1, x_2)} + \varepsilon_i$

many complex solutions possible

b) categorical: may do the same as for continuous. Easiest solution: define new variable with all combinations of x_1 and x_2



Parameter estimation [F3.2]

Estimator for β [F3.2.1]

$Y = X\beta + \varepsilon$

1) Maximum likelihood

If $\varepsilon \sim N_n(0, \sigma^2 I)$ then $Y \sim N_n(X\beta, \sigma^2 I)$

Alt 1: $Y_1, Y_2, \dots, Y_n$ independent

$E(Y_i) = x_i^T\beta$, $Var(Y_i) = \sigma^2$

2) Least squares: minimize $LS(\beta)$ wrt β

$LS(\beta) = (y - X\beta)^T (y - X\beta)$

i) $LS(\beta) = y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$

observe: $\underbrace{y^TX\beta}_{\text{scale}} = \beta^TX^Ty$

$LS(\beta) = y^Ty - 2y^TX\beta + \beta^TX^TX\beta$

ii) To minimize $LS(\beta)$ wrt β we may solve

$$\frac{\partial LS(\beta)}{\partial \beta} = 0$$

$$\begin{pmatrix} \frac{\partial LS(\beta)}{\partial \beta[1]} \\ \frac{\partial LS(\beta)}{\partial \beta[2]} \\ \vdots \\ \frac{\partial LS(\beta)}{\partial \beta[p]} \end{pmatrix}$$

"Need" two rules for derivatives wrt vector:

Rule 1: $\frac{\partial}{\partial \beta}\underset{1\times p}{(d^T\beta)}\underset{p\times 1}{} = \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{p} d_i\beta_i\right) = d$

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

$\{d_{jk}\}$

$= (D + D^T)\beta$ and $2D\beta$ when $D = D^T$.

$$\rightarrow f(y_i; \underbrace{x_i^T\beta}_{\mu}, \sigma) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\cdot e^{-\frac{1}{2\sigma^2}(y_i - \underbrace{\mu}_{x_i^T\beta})^2}$$

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

$f(y)$ joint density of $Y_1, \dots, Y_n$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right\}$$

homework

$$\text{(*)} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\underbrace{(y - X\beta)^T(y - X\beta)}_{LS(\beta)}\right\}$$

maximizing L wrt β is the same as minimizing $LS(\beta)$ wrt β

↑ with respect to

Alt 2: $Y \sim N_n(\mu, \Sigma)$

$$f(y; \mu, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}[\det(\Sigma)]^{-\frac{1}{2}}$$

$$\exp\left\{-\frac{1}{2}(y - \mu)^T\Sigma^{-1}(y - \mu)\right\}$$

Homework: $\mu = X\beta$, $\Sigma = \sigma^2 I$ $\Rightarrow$ get the same $L(\beta, \sigma^2)$ as (*)

Homework: $\frac{\partial}{\partial \beta} LS(\beta)$ with these two rules!

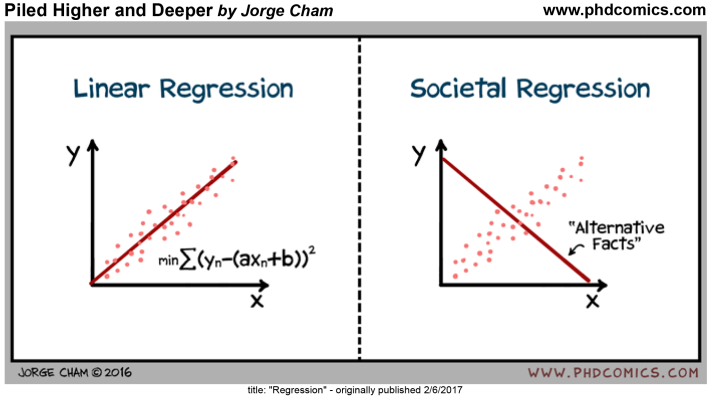
TMA4267 Linear Statistical Models V2017 (L9)

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

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 14, 2017



title: "Regression" - originally published 2/6/2017

The classical linear model

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

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

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 $n > p$).

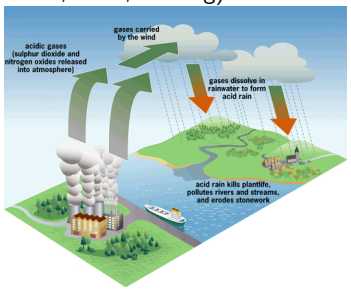
- Q1: What can we say about $\boldsymbol{X}^T\boldsymbol{X}$?
- Q2: Why is the following wrong?

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

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

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/

The classical linear model

The model

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

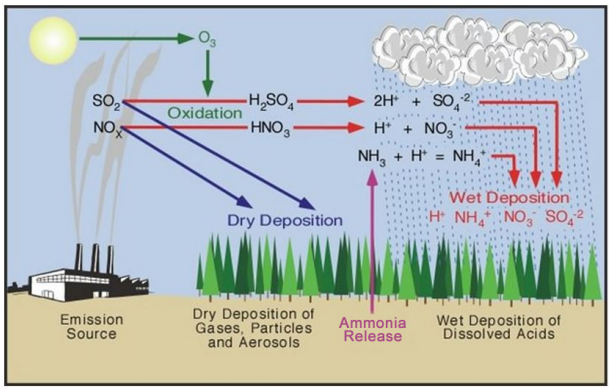
is called a classical linear model if the following is true:

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

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

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

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



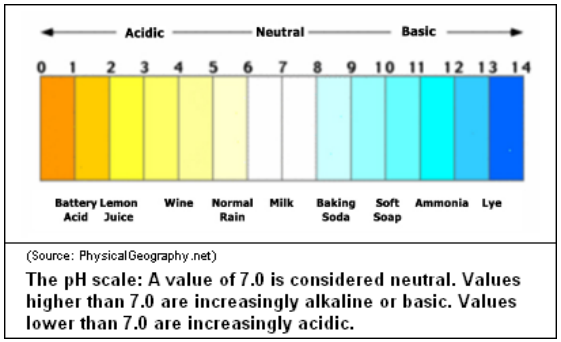
http://www.eoearth.org/view/article/149814/

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),

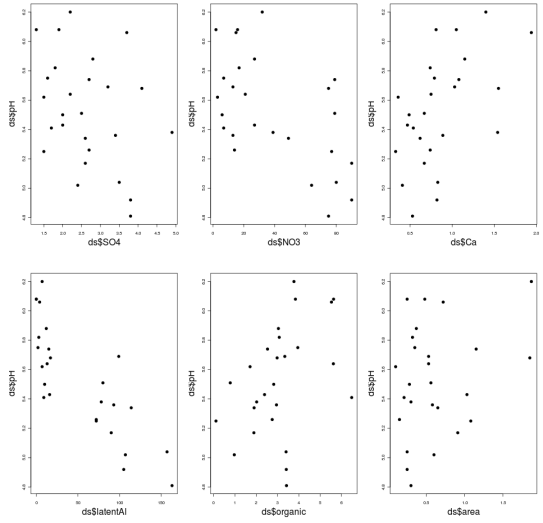
pH is a measure of the acidity of alkalinity of water, expressed in terms of its concentration of hydrogen ions. The pH scale ranges from 0 to 14. A pH of 7 is considered to be neutral. Substances with pH of less that 7 are acidic; substances with pH greater than 7 are basic.

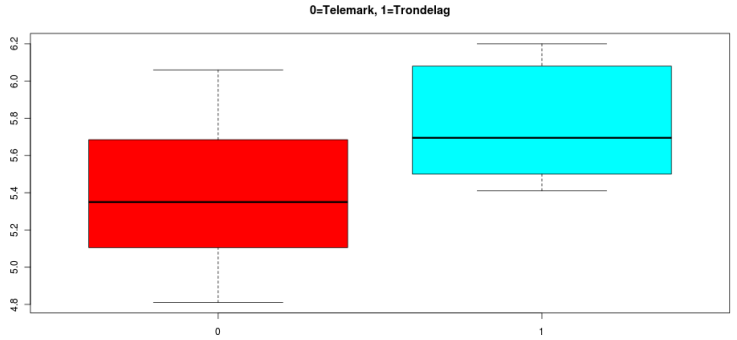


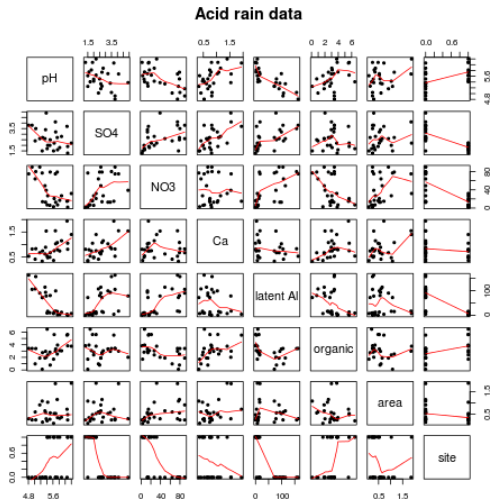
(Source: PhysicalGeography.net)

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/



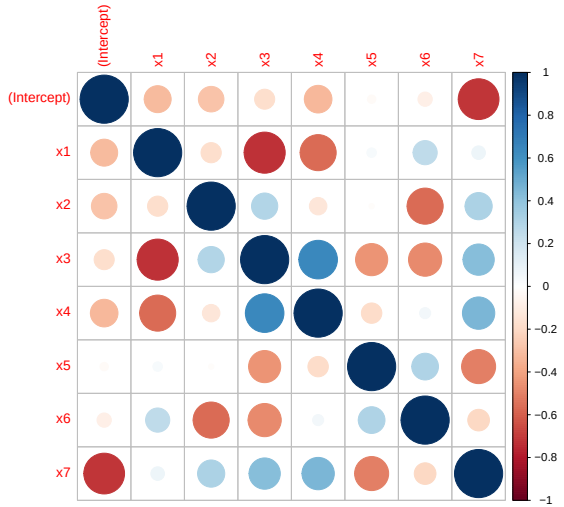




Output from fitting the full model in R

```
> 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 $\hat{\beta}_3$.



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, $\boldsymbol{A}^2 = \boldsymbol{A}$.

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, $(\boldsymbol{A}\boldsymbol{v})^T(\boldsymbol{v} - \boldsymbol{A}\boldsymbol{v}) = 0$.

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.

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.

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, $\boldsymbol{v} = \boldsymbol{Av} + (\boldsymbol{v} - \boldsymbol{Av})$, $\boldsymbol{Av}$ and $\boldsymbol{v} - \boldsymbol{Av} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v}$ are always orthogonal, that is, $(\boldsymbol{Av})^T(\boldsymbol{v} - \boldsymbol{Av}) = 0$.
- 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.

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 σ^2:

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

- 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:

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

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, $\boldsymbol{v} = \boldsymbol{Av} + (\boldsymbol{v} - \boldsymbol{Av})$, $\boldsymbol{Av}$ and $\boldsymbol{v} - \boldsymbol{Av} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v}$ are always orthogonal, that is, $(\boldsymbol{Av})^T(\boldsymbol{v} - \boldsymbol{Av}) = 0$.
- 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 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 $\boldsymbol{X}$.

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

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

This is equivalent to observing that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

This is equivalent to observing that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.

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

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

This is equivalent to observing that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.

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

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

This is equivalent to observing that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.
$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, 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, $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to all columns of $\boldsymbol{X}$, so $\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ and $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

Results from Mathematics 3

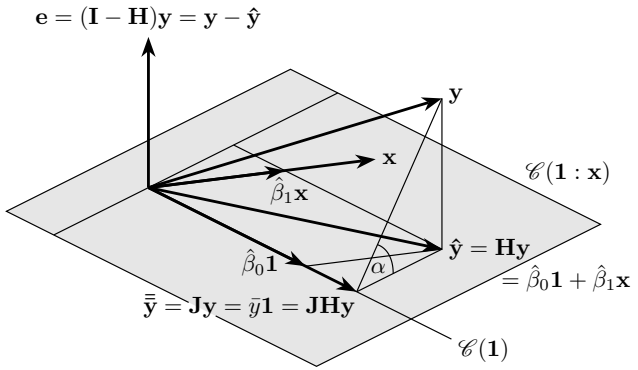
Best approximation theorem
The vector $\hat{\boldsymbol{Y}}$ in the column space of $\boldsymbol{X}$ that makes $\| \boldsymbol{Y} - \hat{\boldsymbol{Y}} \|$ as small as possible, is the orthogonal projection of $\boldsymbol{Y}$ on the column space of $\boldsymbol{X}$.

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

The column space of $\boldsymbol{X}$ consists of vectors of the form $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, so $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 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 $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.

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



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: $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$
- As $\boldsymbol{\beta}$ varies, $\boldsymbol{X}\boldsymbol{\beta}$ 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 $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 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.

Parameter estimation in linear regression

TMA4267 L9
14.02.2017

$\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$, $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$

Estimate β by minimizing $LS(\beta)$

(i) $LS(\beta) = (y - X\beta)^T(y - X\beta) \leftarrow \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$

$= y^Ty - 2y^TX\beta + \beta^TX^TX\beta$

(ii) $\frac{\partial LS(\beta)}{\partial \beta} = 0$

$\frac{\partial}{\partial \beta}(y^Ty - 2y^TX\beta + \beta^TX^TX\beta) = 0$

Rules (see slides)

$a^T = -2y^TX$

$b = X^TX$

$0 - 2X^Ty + 2X^TX\beta = 0$

$X^Ty = X^TX\beta$ normal equations

iii) Solving the normal equations (substitute β with $\hat{\beta}$)

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

ii) min or max

$$\frac{\partial^2}{\partial \beta^2} LS(\beta) = \frac{\partial}{\partial \beta}\left(-2X^T y - 2X^T X\beta\right)\Big|_{\beta=\hat{\beta}}$$

$$= 2X^T X$$

If this matrix has only positive eigenvalues this will be the minimum.

Questions on slide:

X rank p : $X^T X$ $p \times p$ matrix
symmetric
positive definite
inverse exists

If $v^T X^T X v > 0$ for all $v \neq 0$ then $X^T X$ positive definite.

$v^T X^T X v = (Xv)^T (Xv) \geq 0$

Ex: Acid rain : fit in R using lm

$\hat{\beta}_0$ = "Estimate.Intercept" = 5.67
= estimate of the pH in a lake when $x_1 = 0, x_2 = 0, \ldots, x_7 = 0$.
(Telemark)

$\hat{\beta}_3$ = 0.975
α ↑ If x_3 increase by one unit, and all the other covariates are kept constant, then we predict that pH will increase by $\hat{\beta}_3 = 0.975$.

St.Error: $\widehat{SD}(\hat{\beta})$

Assume that $v^T X^T X v = 0$ then $Xv = 0$.
If X has full rank then $Xv = 0$ only has $v = 0$ as solution.
⇒ then $X^T X$ must be positive definite.

$\hat{\beta} = (X^T X)^{-1} X^T Y$ is the least squares estimator of β. If we assume a normal linear model then $\hat{\beta}$ is also the maximum likelihood estimator

Properties of $\hat{\beta}$

$\hat{\beta} = \underbrace{(X^T X)^{-1} X^T}_{C} Y$ (RV) and $E(Y) = X\beta$, $Cov(Y) = \sigma^2 I$

C constants

$Y = X\beta + \varepsilon$, $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$

Find $E(\hat{\beta})$ and $Cov(\hat{\beta})$:

$E(\hat{\beta}) = E(CY) = C\underbrace{E(Y)}_{X\beta} = \underbrace{(X^T X)^{-1} X^T X}_{I}\beta = \beta$

unbiased

$Cov(\hat{\beta}) = Cov(CY) = C\underbrace{Cov(Y)}_{\sigma^2 I} C^T$

$= (X^T X)^{-1} X^T \sigma^2 I \left[(X^T X)^{-1} X^T\right]^T$

$= \sigma^2 (X^T X)^{-1} \underbrace{X^T X (X^T X)^{-1}}_{I}$ ($X^T X$ is symmetric, $(X^T X)^{-1}$ is symmetric)

$= \sigma^2 (X^T X)^{-1}$

In a normal model: $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$

Ex: Acid rain : What is $Var(\hat{\beta}_3)$?

$\widehat{SD}(\hat{\beta}_3) = 0.144$: $\sqrt{\hat{\sigma}^2 \cdot \underbrace{(X^T X)^{-1}_{4,4}}_{\text{ISS}}}$

Need estimator for σ.

Last information on $\hat{\beta}$: From part 1:

$$(\hat{\beta}-E(\hat{\beta}))^T Cov(\hat{\beta})^{-1}(\hat{\beta}-E(\hat{\beta})) \sim \chi^2_p$$

$$\frac{1}{\sigma^2}(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta) \sim \chi^2_p$$

$$Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

$$Cov(\hat{\beta})^{-1} = \frac{1}{\sigma^2}(X^TX)$$

Estimator for σ^2 [F3.2.2]

$Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

$$L(\beta,\sigma^2) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\}$$

$$l(\beta,\sigma^2) = \ln(L(\beta,\sigma^2))$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)$$

$$\frac{\partial l}{\partial \sigma^2} = 0 \iff \frac{d(\ln x)}{x} = \frac{1}{x},\ \frac{dx^{-1}}{dx} = -\frac{1}{x^2}$$

$$0 = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}(y-X\beta)^T(y-X\beta) = 0$$

Predicted values and residuals [F3.2.1]

$E(Y) = X\beta$, so $\widehat{E(Y)} = X\hat{\beta} = \hat{Y}$ ← prediction

$\hat{\beta} = (X^TX)^{-1}X^TY$

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H}Y = HY$$

$H = X(X^TX)^{-1}X^T$ is called the "hat matrix" for putting the hat on Y.

Residuals: $\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I-H)Y$

$= IY$

Observe (see also RecEx3.P3) that

H is symmetric
is idempotent $H^2 = H$
has rank p ⇒ show this

$(I-H)$ is also symmetric and idempotent, and $\text{rank}(I-H) = n-p$. ⇒ show this

⇒ Work with RecEx3.P3 - and be ready for supervision Thurs. kl.15 at Sonja.

$$\frac{n}{\sigma^2} = \frac{1}{\sigma^4}(y-X\beta)^T(y-X\beta)$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(Y-X\hat{\beta})^T(Y-X\hat{\beta}) = \frac{1}{n}\hat{\varepsilon}^T\hat{\varepsilon}$$

$\hat{\varepsilon}$

$\hat{\varepsilon}^T\hat{\varepsilon}$ = sums of squares of errors SSE

But, this estimator is rarely used, because it is biased (use transformula part 1 to find the mean)

However: $\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\hat{\varepsilon}$ is unbiased

REML ↑restricted found by maximizing

$$L(\sigma^2) = \int_p L(\beta,\sigma^2)\,d\beta$$

TMA4295 Statistical inference: more on this.

Ex: Acid rain $\widehat{SD}(\hat{\beta}_0) = \sqrt{(X^TX)^{-1}[4,4]\cdot\hat{\sigma}^2}$

$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\hat{\varepsilon}$ $\hat{\sigma} = 0.1165$

$\hat{\sigma}$ is residual standard error in printout

TMA4267 Linear Statistical Models V2017 (L10)

Part 2: Linear regression: Parameter estimation [F:3.2], Properties of residuals and distribution of estimator for error variance Confidence interval and hypothesis for one regression coefficient

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 17, 2017

Today

1. Properties for residuals (from the hat matrix), leading to properties for $\hat{\sigma}^2$,
2. Then, confidence interval and hypothesis test for regression coefficient.

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

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

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

Results so far

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

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

with mean $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

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

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

- Projection matrices: idempotent, symmetric/orthogonal:

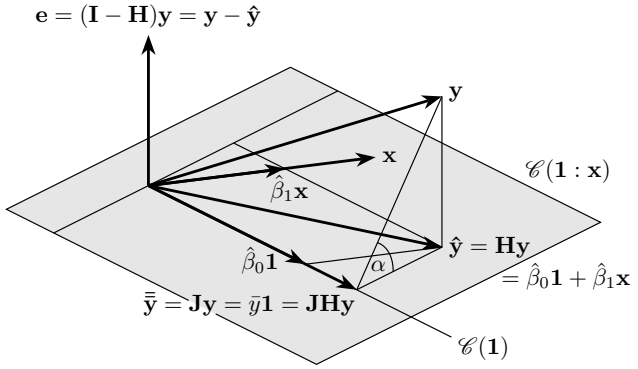
$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

projects onto column space of $\boldsymbol{X}$

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

projects onto space orthogonal to column space of $\boldsymbol{X}$

with important connection: predictions $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ and residuals $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$



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

Quadratic forms [F:B3.3, Theorem B.2]

Random vector $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, symmetric constant matrix $\boldsymbol{A}$.

- Quadratic form: $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
- The "trace-formula": $\mathrm{E}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

Then, let $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{R}$ is a symmetric and idempotent matrix with rank r.

$$\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X} \sim \chi^2_r$$

Now, also $\boldsymbol{S}$ is a symmetric and idempotent matrix with rank s, and $\boldsymbol{R}\boldsymbol{S} = \boldsymbol{0}$.

$$\frac{s\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}}{r\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}} \sim F_{r,s}$$

Properties: $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

- 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 $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
- In addition $\hat{\boldsymbol{\beta}}$ is best linear unbiased estimator (BLUE), that is, among all unbiased estimator it has minimum variance in each component. (More in TMA4295 Statistical Inference.)
- For the normal model: $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.
- Restricted maximum likelihood estimator for $\boldsymbol{\sigma}^2$:

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

- For the normal model

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

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),

Random sample of $n = 26$ lakes.

Output from fitting the full model in R

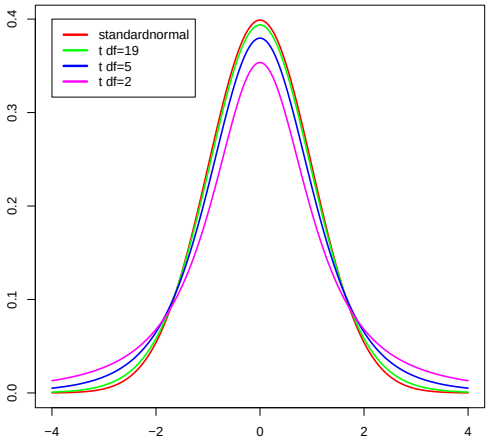
```
> fit=lm(y~.,data=ds)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6764334  0.1389162  40.862  < 2e-16 ***
x1          -0.3150444  0.0587512  -5.362 4.27e-05 ***
x2          -0.0018533  0.0012587  -1.472    0.158
x3           0.9751745  0.1449075   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
```

W. S. Gosset alias Student

Historisk: Student-t fordelingen

- W.S. Gosset (1876-1937) was employed by the Guinness Brewing Company of Dublin.
- Sample sizes available for experimentation in brewing were necessarily small, and Gosset knew that a correct way of dealing with small samples was needed.
- He consulted Karl Pearson (1857-1936) of Universiy College in London about the problem. Pearson told him the current state of knowledge was unsatisfactory.
- The following year Gosset undertook a course of study under Pearson. An outcome of his study was the publication in 1908 of Gosset's paper on "The Probable Error of a Mean," which introduced a form of what later became known as Student's t-distribution.
- Gosset's paper was published under the pseudonym "Student."
- The modern form of Student's t-distribution was derived by R.A. Fisher and first published in 1925.

t-distribution



DEF: t-distribution

Let Z be a standard normal random variable and V a chi-squared random variable with parameter ν (degrees of freedom). If Z and V are independent, the distribution of the random variable T

$$T = \frac{Z}{\sqrt{V/\nu}}$$

has probability density function

$$h(t) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}}(1+\frac{t^2}{\nu})^{-(\nu+1)/2}$$

for $-\infty < t < \infty$. This distribution is called the (Student) $t-$distribution with ν degrees of freedom.

- $\mathrm{E}(T) = 0$ if $\nu \geq 2$.
- $\mathrm{Var}(T) = \frac{\nu}{\nu-2}$ if $\nu \geq 3$.

Are $\hat{\boldsymbol{\beta}}$ and SSE are independent?

Independence – from Part 1:
Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent iff $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$.

We have:

- $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$
- $\boldsymbol{AY} = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and
- $\boldsymbol{BY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.
- Now $\boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{B}^T = \sigma^2\boldsymbol{A}\boldsymbol{B}^T = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$
- since $\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{X} - \boldsymbol{HX} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}$.
- We conclude that $\hat{\boldsymbol{\beta}}$ is independent of $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$,
- and, since SSE=function of $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$: SSE=$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$,
- then $\hat{\boldsymbol{\beta}}$ and SSE are independent.

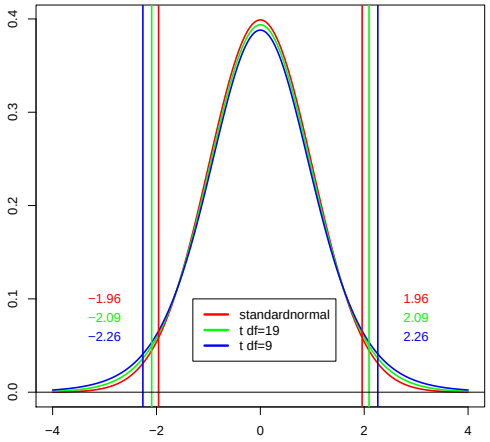
Kritiske verdier i t-fordelingen

$P(T > t_{\alpha,\nu}) = \alpha$

$\nu \backslash \alpha$	.150	.100	.075	.050	.025	.010	.005	.001	.0005
1	1.963	3.078	4.165	6.314	12.706	31.821	63.657	318.309	636.619
2	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.327	31.599
3	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.215	12.924
4	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850
21	1.063	1.323	1.494	1.721	2.080	2.518	2.831	3.527	3.819
22	1.061	1.321	1.492	1.717	2.074	2.508	2.819	3.505	3.792
23	1.060	1.319	1.489	1.714	2.069	2.500	2.807	3.485	3.768
24	1.059	1.318	1.487	1.711	2.064	2.492	2.797	3.467	3.745
25	1.058	1.316	1.485	1.708	2.060	2.485	2.787	3.450	3.725
26	1.058	1.315	1.483	1.706	2.056	2.479	2.779	3.435	3.707
27	1.057	1.314	1.482	1.703	2.052	2.473	2.771	3.421	3.690
28	1.056	1.313	1.480	1.701	2.048	2.467	2.763	3.408	3.674
29	1.055	1.311	1.479	1.699	2.045	2.462	2.756	3.396	3.659
30	1.055	1.310	1.477	1.697	2.042	2.457	2.750	3.385	3.646
35	1.052	1.306	1.472	1.690	2.030	2.438	2.724	3.340	3.591
40	1.050	1.303	1.468	1.684	2.021	2.423	2.704	3.307	3.551
50	1.047	1.299	1.462	1.676	2.009	2.403	2.678	3.261	3.496
60	1.045	1.296	1.458	1.671	2.000	2.390	2.660	3.232	3.460
80	1.043	1.292	1.453	1.664	1.990	2.374	2.639	3.195	3.416
100	1.042	1.290	1.451	1.660	1.984	2.364	2.626	3.174	3.390
120	1.041	1.289	1.449	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.036	1.282	1.440	1.645	1.960	2.326	2.576	3.090	3.291

4

Quantiles and critical values: N og t: $\alpha/2 = 0.025$



Acid rain in R

```
ds=read.table("https://www.math.ntnu.no/emner/
TMA4267/2017v/acidrain.txt",header=TRUE)
fit=lm(y~.,data=ds)
> confint(fit)
                    2.5 %       97.5 %
(Intercept)  5.384581378  5.9682854281
x1          -0.438476153 -0.1916126966
x2          -0.004497716  0.0007911594
x3           0.670735075  1.2796138706
x4          -0.002335625  0.0018820903
x5          -0.080696921  0.0138484550
x6          -0.156117992  0.1482381575
x7          -0.126624544  0.3043688780
```

P-values: `http://www.statistrikk.no/wp-content/uploads/2017/02/nerdekort.jpg`

Today

- Distribution of SSE/σ^2 is chisquared $(n-p)$.
- Independence of $\hat{\beta}$ and SSE.
- Inference about β components can be performed using the t-distribution

PART 2: LINEAR REGRESSION — THURSDAY L10 — 11.02.2017

[FS2]

Distribution of $\hat{\varepsilon}$ (residuals)

Residuals: $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = Y - \underbrace{X(X^TX)^{-1}X^T}_{H}Y$

$= Y - HY = (I-H)Y$

$E(\hat{\varepsilon}) = E((I-H)Y) = (I-H)\underbrace{E(Y)}_{X\beta}$ ($E(\varepsilon)=0$, $Y = X\beta + \varepsilon$, $E(Y) = X\beta$, $Cov(Y) = Cov(\varepsilon) = \sigma^2 I$)

$= \underbrace{(I-H)X\beta}_{\text{project onto the space orth. to column space of } X} = 0$

or $(I-H)X\beta = (X - \underbrace{HX}_{X})\beta = 0$

$HX = X(X^TX)^{-1}X^TX = X$

$Cov(\hat{\varepsilon}) = Cov((I-H)Y) = (I-H)\underbrace{Cov(Y)}_{\sigma^2 I}(I-H)^T$

$= \sigma^2(I-H)I(I-H) = \sigma^2(I-H)$

Assume $\varepsilon \sim N_n(0, \sigma^2 I) \Rightarrow \hat{\varepsilon} \sim N_n(0, \sigma^2(I-H))$

NB: $\text{rank}(H) = p$, $\text{rank}(I-H) = n-p$, which means $(I-H)^{-1}$ does not exist and we use the singular version of the normal pdf.

Distribution of SSE and $\hat{\sigma}^2$

$SSE = \hat{\varepsilon}^T\hat{\varepsilon} = Y^T(I-H)(I-H)Y$

($\sum_{i=1}^n (y_i - \hat{y})^2$, $\hat{\varepsilon} = (I-H)Y$)

$SSE = Y^T(I-H)Y$

remember $\text{rank}(I-H) = n-p$.

RecEx3.P3 looks at the distribution of $\frac{1}{\sigma^2}Y^T(I-H)Y = \frac{SSE}{\sigma^2}$ by using the result on quadratic forms from Part 1 (see slides)

$Y \sim N_n(X\beta, \sigma^2 I)$

$Y^* = \frac{1}{\sigma}(Y - X\beta) \sim N_n(0, I)$

$Y^{*T}(I-H)Y^* = \frac{1}{\sigma^2}Y^T(I-H)Y$

$(I-H)(Y - X\beta)$

$(I-H)Y - \underbrace{(I-H)X\beta}_{0}$

$\hat{\varepsilon}^T\hat{\varepsilon}$

$\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$

χ^2_{n-p}

Now: $\hat{\sigma}^2 = \frac{1}{n-p}SSE \Leftrightarrow SSE = (n-p)\hat{\sigma}^2$

$V = \frac{SSE}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

$E(V) = n-p$ $\quad Var(V) = 2(n-p)$

Is $\hat{\sigma}^2$ an unbiased estimator?

$E(\hat{\sigma}^2) = E(\frac{1}{n-p}\underbrace{SSE}_{\sigma^2 \cdot V}) = \frac{1}{n-p}E(\sigma^2 V)$

$= \frac{\sigma^2}{n-p}\underbrace{E(V)}_{n-p} = \sigma^2$ unbiased.

This is true when we assume $\varepsilon \sim N$.

If we do not assume $\varepsilon \sim N$, then we can use the trace-formula

$E(SSE) = E(Y^T(I-H)Y)$ ($E(Y) = X\beta$, $Cov(Y) = \sigma^2 I$)

$= tr((I-H)\sigma^2 I) + (X\beta)^T\underbrace{(I-H)X\beta}_{0}$

$= (n-p)\sigma^2 + 0$

$E(\hat{\sigma}^2) = E(\frac{SSE}{n-p}) = \sigma^2$

Inference about one β_j

Ex: Acid rain β_1 = effect of SO_4 on pH of lake

$\hat{\beta}_1 = -0.315$

$SD(\hat{\beta}_1) = \sqrt{\sigma^2 [(X^T X)^{-1}]}$ [arrangement corresp. to SO_4 x_1]

$\hat{SD}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]}$

$\Downarrow$

St.Error $\Rightarrow$ 0.0587 in printout

$\hat{\sigma}$: "Residual standard error" = 0.1165

$\Downarrow \sqrt{\frac{SSE}{n-p}}$ $n = 26$, $p = 8$ $\}$ $n-p = 18$

"on 18 degrees of freedom".

To find a confidence interval for β_j – or to test hypotheses about β_j we need to know the distribution of a statistic involving $\hat{\beta}_j$ and β_j – with no other unknown parameters.

$$\hat{\beta}_j \sim N_1(\beta_j, \sigma^2 \underbrace{(X^T X)^{-1}_{[j,j]}}_{c_{jj}})$$

We have:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\sigma} \sim N(0,1)$$

and $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\sigma}}{\sqrt{\frac{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}}{n-p}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-p}$$

$\hat{\beta}_j$ and $\hat{\sigma}^2$ need to be independent in order that this holds. $\Rightarrow$ Rec Ex 3.P3 + slides

$\Downarrow$

Use $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-p}$ for inference.

a) Find a 95% confidence interval (CI) for β_j

and $\hat{\sigma}^2 = \frac{SSE}{n-p}$ when $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$.

Then:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}} \sim N(0,1)$$

($\hat{\beta}_j - \beta_j$: $E(\hat{\beta}_j)$; $\sqrt{c_{jj}\sigma^2}$: $SD(\hat{\beta}_j)$)

$\hat{SD}(\hat{\beta}_j) = \sqrt{c_{jj}} \cdot \hat{\sigma}$ and $\hat{\beta}_j$ and $\hat{\sigma}^2$ are independent (to be shown)

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-p}$$

General result:

$$\frac{N(0,1)}{\sqrt{\frac{\chi^2_q}{q}}} \sim t_q$$

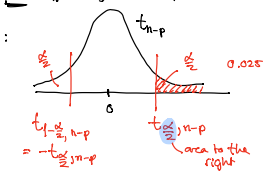
b) Test:

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

at significance level 0.05.

Ex: Acid rain β_1 = effect of SO_4 on pH.

a) 95% CI: $(1-\alpha)\cdot 100$



$$P\left(\underbrace{-t_{\frac{\alpha}{2},n-p}}_{0.025} < T_j < \underbrace{t_{\frac{\alpha}{2},n-p}}_{0.025}\right) = \underbrace{1-\alpha}_{0.95}$$

$\uparrow$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$$

L = lower, U = upper

$$P\Big(\underbrace{\hat{\beta}_j - t_{\frac{\alpha}{2},n-p}\cdot\sqrt{c_{jj}}\,\hat{\sigma}}_{\hat{\beta}_L} < \beta_j < \underbrace{\hat{\beta}_j + t_{\frac{\alpha}{2},n-p}\sqrt{c_{jj}}\,\hat{\sigma}}_{\hat{\beta}_U}\Big) = 1-\alpha$$

Ex: $\hat{\beta}_1 = -0.515$, $\sqrt{\hat{V}_{jj}\hat{\sigma}^2} = 0.058$, $n=25$, $p=8$, $t_{0.025,17} = 2.1$

$-0.515 \pm 2.1 \cdot 0.058$

$= (-0.44, -0.197)$

How do you interpret this interval?

→ Strong belief (95%) that β_j is in interval:

We see that 0 is not in the interval – what does this mean? ⇒ Reject $H_0: \beta_0 = 0$ vs $H_1: \beta_j \neq 0$ at sign. level 5%

6

TMA4267 Linear Statistical Models V2017 (L11)

Part 2: Linear regression:
Parameter estimation [F:3.2] and model selection [F:3.4]
Hypothesis test for one regression coefficient
Studentized and standardized residuals
decomposition of variability and signficance of regression
R^2, SPSE=Expected squared prediction error

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 21, 2017

Today

1. Hypothesis testing for β_j.
2. Residuals: standardized (or studentized) preferred.
3. Decomposition of variability: SST=SSR+SSE, and significance of regression.
4. R^2 gives the proportion of variability explained by the regression model. and will never decrease if new covariates are added to the model.
5. Model choice considerations.
6. SPSE: Expected squared prediction error.

The classical linear model

The model

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

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

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

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

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

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

Properties for the normal linear model

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

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

  with $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.
- Restricted maximum likelihood estimator for $\boldsymbol{\sigma}^2$:

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

  with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- Statistic for inference about β_j, c_{jj} is diagonal element j of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$$

Acid rain in Norwegian lakes

Measured pH in Norwegian lakes explained by content of

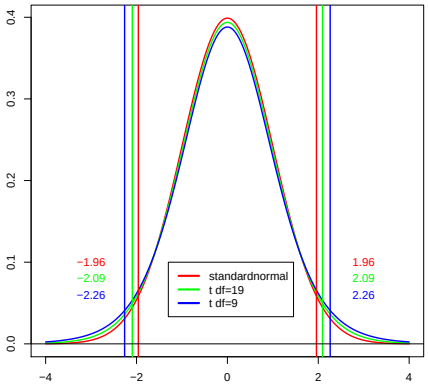
- x1: SO_4: sulfate (the salt of sulfuric acid),
- x2: $N0_3$: nitrate (the conjugate base of nitric acid),
- x3: Ca: calsium,
- x4: latent Al: aluminium,
- x5: organic substance,
- x6: area of lake,
- x7: position of lake (Telemark or Trøndelag),

Random sample of $n = 26$ lakes.

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
```


Quantiles and critical values: N og t: $\alpha/2 = 0.025$



In R: specify area to the left, but our notation gives area to the right. Fahrmeir et al: notation with area to the left.

Properties of the residuals

- Residuals (raw): $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.
- with mean $\mathrm{E}(\hat{\varepsilon}) = \boldsymbol{0}$ and covariance matrix $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
- In the normal model $\varepsilon \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and then also the vector of residuals are normal, but with heteroscedastic variances and non-zero covariances.
- *Standardized residuals:* divide (raw) residuals by estimated standard deviation.
- *Studentized residuals*: leave-one-out version.
- Studentized residuals are compared with the normal distribution to assess normality of the error term.

3.12 Overview of Residuals

Ordinary Residuals

The residuals are given by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \quad i = 1, \ldots, n.$$

Standardized Residuals

The standardized residuals are defined by

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

where h_{ii} is the ith diagonal element of the hat matrix.

Studentized Residuals

The studentized residuals are defined by

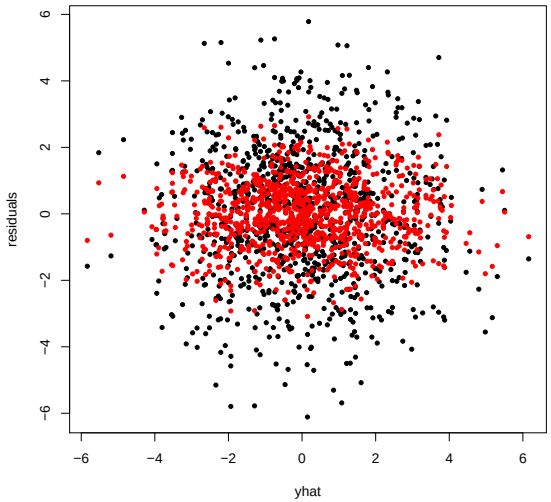
$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i\left(\frac{n - p - 1}{n - p - r_i^2}\right)^{1/2}.$$

The studentized residuals are used to verify model assumptions and to discover outliers (see Sect. 3.4.4).

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.126)

Simulating data and checking residuals

```
n=1000
beta=matrix(c(0,1,1/2,1/3),ncol=1)
set.seed(123)
x1=rnorm(n,0,1); x2=rnorm(n,0,2); x3=rnorm(n,0,3)
X=cbind(rep(1,n),x1,x2,x3)

y=X%*%beta+rnorm(n,0,2)
fit=lm(y~x1+x2+x3)
yhat=predict(fit)
summary(fit)
ehat=residuals(fit); estand=rstandard(fit); estud=rstudent(fit)
plot(yhat,ehat,pch=20)
points(yhat,estand,pch=20,col=2)
#points(yhat,estud,pch=20,col=5)
```

Black: raw residuals, red: standardized residuals (identical to studentized here).

Examination of model assumptions

1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$
2. Homoscedastic error variance: $Cov(\varepsilon) = \sigma^2 \boldsymbol{I}$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$
5. Assumption of normality: $\varepsilon \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

Plotting residuals

1. Plot the residuals, r_i^* against the predicted values, $\hat{y}_i$.
   - Dependence of the residuals on the predicted value: wrong regression model?
   - Nonconstant variance: transformation or weighted least squares is needed?
2. Plot the residuals, r_i^*, against predictor variable or functions of predictor variables. Trend suggest that transformation of the predictors or more terms are needed in the regression.
3. Assessing normality of errors: QQ-plots and histograms of residuals. As an additional aid a test for normality can be used, but must be interpreted with caution since for small sample sizes the test is not very powerful and for large sample sizes even very small deviances from normality will be labelled as significant.
4. Plot the residuals, r_i^*, versus time or collection order (if possible). Look for dependence or autocorrelation.

Volume of a tree

Data for 31 trees of a certain kind in a national park in the US are given below. Three variables are measured for each tree. These are:

- D: The diameter of the tree measured in inches 1.5 m above ground level
- H: The height of the tree measured in feet.
- V: The volume of the tree measured in cubic feet.

Obs.	D	H	V	Obs.	D	H	V
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Volume of a tree

- If one wants to measure the volume of a tree the tree has to be cut down.
- But, height and diameter can be measured without cutting down the tree.
- Of interest: develop a model that can be used to estimate the tree volume from the height and diameter.

As an illustration assume we want to fit a linear model with V as response and D and H as covariates. What is the R^2 of this model?

Comment: if we start with the volume of a cylinder (area of circle times height) we may suggest a different regression model (on the log scale). Which model?

Volume: height and diameter

```
fit <- lm(Volume~.,data=ds)
summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```


Volume of a tree: IQ of lumberjack added

- We want to add the IQ of the lumberjack that cut down the tree as a covariate in the model.
- This should for obvious reasons not be a good predictor for the volume of the tree.
- To mimic this situation we simulate new data to resemble the IQ of different lumberjacks by drawing data from the normal distribution with mean 100 and standard deviation 16, and since we have 31 trees we simulate 31 observations.
- Q: will the R^2 of this new model be higher than the R^2 of the previous model?

Volume: height and diameter – and IQ of lumberjack

```
set.seed(123) # reproducible results
iq <- rnorm(31,100,16)
fit2 <- lm(Volume~Height+Diameter+iq,data=ds)
summary(fit2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.03399   10.20868  -5.979 2.24e-06 ***
Height        0.34099    0.13176   2.588   0.0154 *
Diameter      4.72507    0.26906  17.561 2.68e-16 ***
iq            0.02704    0.04678   0.578   0.5681
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.929 on 27 degrees of freedom
Multiple R-squared:  0.9486,Adjusted R-squared:  0.9429
F-statistic: 166.1 on 3 and 27 DF,  p-value: < 2.2e-16
```


Acid rain in Norwegian lakes

Data on $n = 26$ lakes, with

- y: measured pH in lake,
- 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),

We would like to use a regression model with pH of the lake as the response. Should we fit a model will all 7 covariates, or choose a subset?

Simulated data (Fahrmeir et al: Fig 3.17)

True model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Known that the model is polynomial in nature, but not up to which degree.
Try to fit polynomial also with higher order terms.

New: in addition to the data set to be used to fit the regression (called *training set*) also a data set to assess the model fit is present (called a *validation* set).

Mean Squared Error (MSE) is a scaled version of the SSE, that is $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

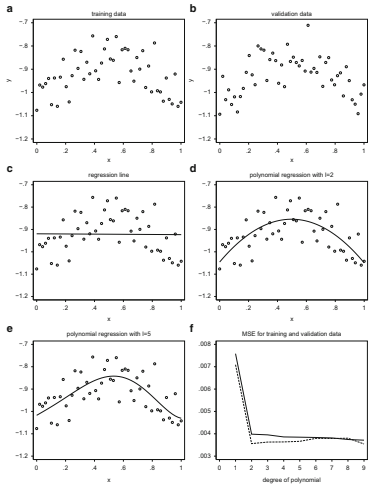


Fig. 3.17 Simulated training data y_i [panel (**a**)] and validation data y_i^* [panel (**b**)] based on 50 design points $x_i, i = 1, \ldots, 50$. The true model used for simulation is $y_i = -1 + 0.3x_i + 0.4x_i^2 - 0.8x_i^3 + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, 0.07^2)$. Panels (**c**-**e**) show estimated polynomials of degree $l = 1, 2, 5$ based on the training set. Panel (**f**) displays the mean squared error $\mathrm{MSE}(l)$ of the fitted values in relation to the polynomial degree (*solid line*). The *dashed line* shows $\mathrm{MSE}(l)$, if the estimated polynomials are used to predict the validation data y_i^*

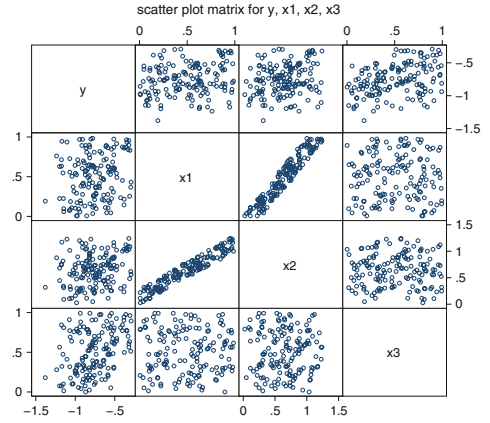
Figure from our text book: Fahrmeir et al (2013): Regression. Springer. (p.140)

Simulated data (Fahrmeir et al: Fig 3.18, Tab3.3, Tab3.4)

True model:

$$Y \sim N(-1 + 0.3x_1 + 0.2x_3, 0.2^2)$$

where also $x_2 = x_1 + u$ is observed ($u \sim$ uniform in 0,1). The variables x_1 and x_3 are uncorrelated.



ig. 3.18 Scatter plot matrix for the variables y, x_1, x_2, and x_3

Figure from our text book: Fahrmeir et al (2013): Regression. Springer. (p.141)

Table 3.3 Results for the model based on covariates x_1, x_2, and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.970	0.047	−20.46	<0.001	−1.064	−0.877
x_1	0.146	0.187	0.78	0.436	−0.224	0.516
x_2	0.027	0.177	0.15	0.880	−0.323	0.377
x_3	0.227	0.052	4.32	<0.001	0.123	0.331

Table 3.4 Results for the correctly specified model based on covariates x_1 and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.967	0.039	−24.91	<0.001	−1.042	−0.889
x_1	0.173	0.055	3.17	0.002	0.065	0.281
x_3	0.226	0.052	4.33	<0.001	0.123	0.330

Table from our text book: Fahrmeir et al (2013): Regression. Springer. (p.142)

Irrelevant and/or missing covariates in the regression

Irrelevant : variables that are included in the regression but should not have been.

missing : variables that are not included, but should have been.

Two subsets of covariates (Exam V2014 Problem 4b)

Classical linear model with identically normally distributed random errors, $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$, but now look at misspecification of $\mathrm{E}(\boldsymbol{Y})$. Suppose that the true model is

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon, \\ \varepsilon &\sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \end{aligned} \tag{1}$$

where we have partitioned the design matrix into two parts $\boldsymbol{X}_1$ ($n \times p_1$) and $\boldsymbol{X}_2$ ($n \times p_2$) and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown p_1- and p_2-dimensional vectors of regression coefficients ($p = p_1 + p_2$).

Two subsets of covariates (cont.)

Assume that we ignore the covariates in $\boldsymbol{X}_2$ and fit the model

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\delta}, \\ \boldsymbol{\delta} &\sim N_n(\boldsymbol{0}, \tau^2\boldsymbol{I}). \end{aligned} \tag{2}$$

Here $\boldsymbol{\alpha}_1$ is used in place of $\boldsymbol{\beta}_1$ to emphasize that $\boldsymbol{\alpha}_1$ (and estimates thereof) will in general be different from $\boldsymbol{\beta}_1$ in the true model. The least squares estimator for model (2) is $\hat{\boldsymbol{\alpha}_1} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$.

Two subsets of covariates (cont.)

Find the expected value and covariance matrix of $\hat{\boldsymbol{\alpha}}_1$ under the true model.

$$\mathrm{E}(\hat{\boldsymbol{\alpha}}_1) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$$

We see that the bias term for $\hat{\boldsymbol{\alpha}}_1$ is $(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$. When is the bias term equal to zero?

$$\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1) = \sigma^2(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}$$

Observe, $\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1)$ is not dependent on $\boldsymbol{\beta}_2$.

Missing covariates: findings

Bias : The estimator for the (true) covariates (in the model) is only unbiased if the true and missing covariates are uncorrelated (orthogonal design) in the data.

Variance : The variance of the estimator for the true covariates may be smaller based on the model with the missing covariates (than for the correctly specified model), and even the sum of the bias^2 and the variance may better for the model with the missing variables. So the sparse model may be better on overall (even though it is biased).

Irrelevant covariates included: findings

Bias : The estimator for the true covariates are unbiased, also if irrelevant covariates are included.

Variance : The model with the irrelevant covariants have larger variance for the true covariates, compared with the model without the irrelevant covariates. So, again sparse model is the best.

Irrelevant and/or missing covariates in the regression

Irrelevant : variables that are included in the regression but should not have been.

missing : variables that are not included, but should have been.

Conclusion in book: the model should not contain irrelevant covariates, and we should aim for a sparse model.

Law of parsimony

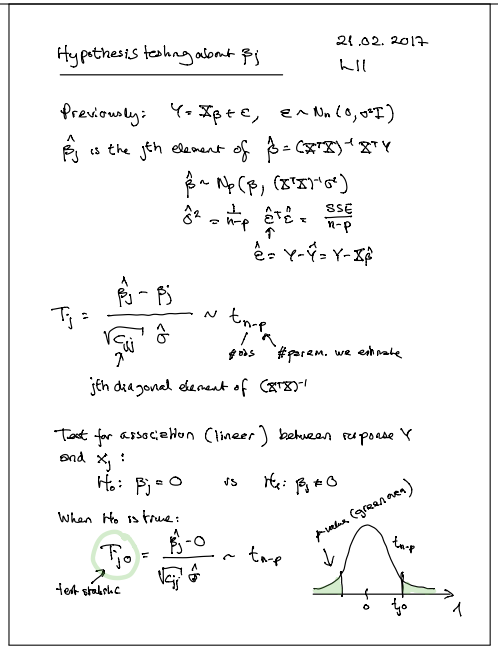
If two models are not very different – then always choose the simplest one

Today

- T-test for significance of one regression coefficient.
- Residuals: standardized (or studentized) preferred.
- Significance of regression based on F-test with SSR/(p-1) divided by SST/(n-1).
- R^2 gives the proportion of variability explained by the regression model.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

  and will never decrease if new covariates are added to the model.
- Model selection: want to choose the model that minimize the expected squared prediction error.



Hypothesis testing about β_j — 21.02.2017 — L11

Previously: $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

$\hat{\beta}_j$ is the jth element of $\hat{\beta} = (X^T X)^{-1} X^T Y$

$\hat{\beta} \sim N_p(\beta, (X^T X)^{-1}\sigma^2)$

$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\hat{\varepsilon} = \frac{SSE}{n-p}$

$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$

$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$ (df's; # param. we estimate)

jth diagonal element of $(X^T X)^{-1}$

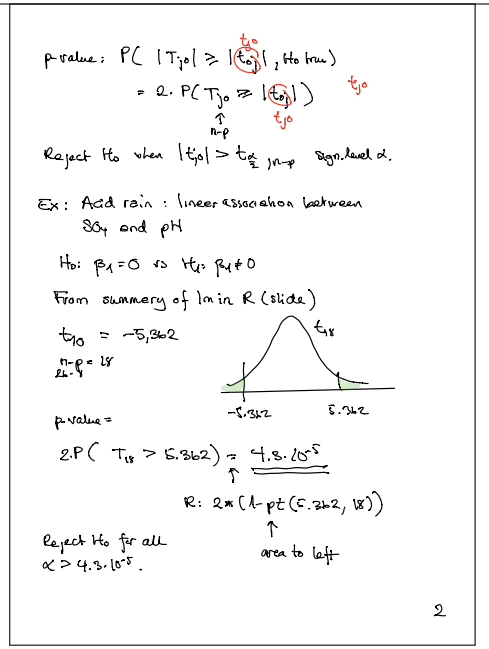
Test for association (linear) between response Y and x_j:

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

When H_0 is true:

$T_{j0} = \frac{\hat{\beta}_j - 0}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$ (test statistic)

p-value (greater area)



p-value: $P(|T_{j0}| \geq |t_{oj}|, H_0 \text{ true}) = 2 \cdot P(T_{j0} \geq |t_{oj}|)$ (t_{j0}, t_{oj}, t_{j0}; $n-p$)

Reject H_0 when $|t_{j0}| > t_{\frac{\alpha}{2}, n-p}$ Sign. level α.

Ex: Acid rain: linear association between SO_4 and pH

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

From summery of lm in R (slide)

$t_{10} = -5,362$, $n-p = 18$

p-value = $2P(T_{18} > 5.362) = 4.3 \cdot 10^{-5}$

R: 2*(1-pt(5.362, 18))

Reject H_0 for all $\alpha > 4.3 \cdot 10^{-5}$.

area to left

2

Residuals (again)

$\hat{\varepsilon} = Y - \hat{Y}$, $\hat{\varepsilon} \sim N_n(0, \sigma^2(I-H))$

R: residuals(fit)

The residuals have heteroscedastic variance
$Var(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii})$ and $Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(0-h_{ij})$
(can in general be $\neq 0$, but in most cases experience shows that ≈ 0)

Standardized residuals:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

will be (approx.) homoscedastic.

R: rstandard(fit)

Studentized residuals: Fitting the model to all obs. except i to make r_i^*. ← see slide

⇓

Has studentized!

R: rstudent(fit)

See example on slide for r_i vs $\hat{\varepsilon}_i$

Ex: Acid rain, full model (all covariates available)
$R^2 = 0.73$, 93%

Is the regression significant?

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs

H_1: at least one $\beta_j \neq 0$ $j = 1, \ldots, k$

Test statistics: $F = \frac{SSR/k}{SSE/(n-p)} \sim F_{k, n-p}$ ($SST = SSR + SSE$)

↑ prove this in Part B in a general setting

Ex: Acid rain:

F-observed: 84.15
$k = 7$, $n-p = 18$ (26 − 8)

$P(F_{7,18} > 84.15) = 3.7 \cdot 10^{-7}$ (p-value)

Analysis of variance decomposition and R^2

$y_1, \ldots, y_n$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \cdots =$$

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

- $\sum_{i=1}^n (y_i - \bar{y})^2$: sum of squared total — SST
- $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$: sum of squared errors — SSE
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$: sum of sq regression — SSR (explained by regression)

With vectors and matrices: $\hat{\varepsilon}^T\hat{\varepsilon}$

$$Y^T(I - \tfrac{1}{n}11^T)Y = Y^T(I-H)Y + Y^T(H - \tfrac{1}{n}11^T)Y$$

SST SSE SSR

This is used to define:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2: coefficient of determination
- relative proportion of total variability explained by the regression

Ex: Volume of tree and the lumberjack

Big model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

Small model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

$R^2_{big} \geq R^2_{small}$ ← since $\hat{\beta}_3$ is found to minimize $SSE = \hat{\varepsilon}^T\hat{\varepsilon}$, thus maximize $R^2 = 1 - \frac{SSE}{SST}$

$R^2_{big} = R^2_{small}$ if $\hat{\beta}_3 = 0$

if $\hat{\beta}_3 \neq 0$ the $SSE_{big} < SSE_{small}$ and $R^2_{big} > R^2_{small}$.

R^2 will always increase (or stay unchanged) when a new covariate is added to the model.

Next: more on choosing a good model, and then

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

← penalizing adding many covariates

is one criterion to use instead of R^2 for model selection.

Model choice and variable selection [F3.4]

Question 1: Is a full model (all available covariates fitted) the best model?

good interpretability — good for future predictions

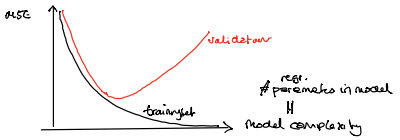
Data is divided into

Training set — estimate model parameters

Validation set — evaluate model fit

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Now calculate the MSE on the training and on the validation set and plot as a function of model complexity:



7

Answer 1: No, this may lead to overfitting = fitting the trend + the noise!

⇒ so, what can we do instead?

8

TMA4267 Linear Statistical Models V2017 (L12)

Part 2: Linear regression: Model selection [F:3.4] Transformation and Taylor expansion Quiz

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 24, 2017

What is the "best" model?

Acid rain in Norwegian lakes, data on $n = 26$ lakes, with

- y: measured pH in lake,
- 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),

Topic: choosing the "best" linear regression model!

- First, debunk popular strategies (based on simulations studies were we knew the "true" model):
  - Popular 1: fit all available covariates.
    Problem: overfitting (=fitting trends and noise).
  - Popular 2: fit all available covariates, then remove the insignificant ones (=those β_j where $H_0 : \beta_j = 0$ is not rejected).

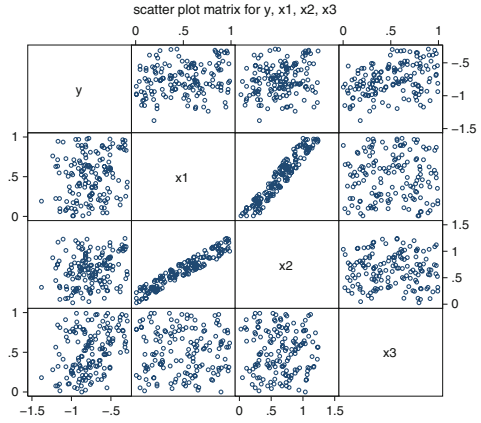


Fig. 3.18 Scatter plot matrix for the variables y, x_1, x_2, and x_3

Figure from our text book: Fahrmeir et al (2013): Regression. Springer. (p.141)

Simulated data (Fahrmeir et al: Fig 3.18, Tab3.3, Tab3.4)

True model:

$$Y \sim N(-1 + 0.3x_1 + 0.2x_3, 0.2^2)$$

where also $x_2 = x_1 + u$ is observed ($u \sim$ uniform in 0,1). The variables x_1 and x_3 are uncorrelated.

Table 3.3 Results for the model based on covariates x_1, x_2, and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.970	0.047	−20.46	<0.001	−1.064	−0.877
x_1	0.146	0.187	0.78	0.436	−0.224	0.516
x_2	0.027	0.177	0.15	0.880	−0.323	0.377
x_3	0.227	0.052	4.32	<0.001	0.123	0.331

Table 3.4 Results for the correctly specified model based on covariates x_1 and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.967	0.039	−24.91	<0.001	−1.042	−0.889
x_1	0.173	0.055	3.17	0.002	0.065	0.281
x_3	0.226	0.052	4.33	<0.001	0.123	0.330

Table from our text book: Fahrmeir et al (2013): Regression. Springer. (p.142)

Topic: choosing the "best" linear regression model!

- First, debunk popular strategies (based on simulations studies were we knew the "true" model):
  - Popular 1: fit all available covariates.
    Problem: overfitting (=fitting trends and noise).
  - Popular 2: fit all available covariates, then remove the insignificant ones (=those β_j where $H_0 : \beta_j = 0$ is rejected).
    Problem: may also remove important covariates that are correlated with unimportant ones - but insignificant because being masked by the unimportant ones.
- Study of irrelevant and missing covariates:

  Irrelevant : variables that are included in the regression but should not have been (IQ of lumberjack)

  missing : variables that are not included, but should have been (omitting height in the tree volum example)

  Conclusion in book: the model should not contain irrelevant covariates, and we should aim for a sparse model.

  Take home message is the "Law of parsimony": *If two models are not very different – then always choose the simplest one.*

All models are wrong?

A model is a simplification or approximation of reality and hence will not reflect all of reality.

George Box noted that "all models are wrong, but some are useful". While a model can never be "truth"a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless.

Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.

Expected squared prediction error (SPSE)

Possible criterion we want to minimize: SPSE.
Definition (j, M, ... given in classnotes)

$$\text{SPSE} = \sum_{j=1}^{J} \text{E}((Y_j - \hat{Y}_{jM})^2)$$

can be written as:

$$\text{SPSE} = \sum_{j=1}^{J} \text{E}((Y_j - \hat{Y}_{jM})^2) = n\sigma^2 + |M|\sigma^2 + \sum_{j=1}^{J}(\mu_{jM} - \mu_j)^2$$

Problem: Not useful on practise since μ_j and σ^2 are unknown.
Plan: Find a way to estimate SPSE and then choose the model M with the minimum SPSE!

How to estimate SPSE?

$$\text{SPSE} = \sum_{j=1}^{J} \text{E}((Y_j - \hat{Y}_{jM})^2)$$

Assume we have fitted a model M with $|M|$ regression parameters.

1. Use new (independent) data – if available (seldom the case):

$$\widehat{SPSE} = \sum_{j=1}^{J}(Y_j - \hat{Y}_{jM})^2$$

2. Cross-validation: mimic new data by dividing data into k folds (popular is $k = n$ and $k = 10$). In a for-loop let $j = 1, \ldots, k$, and use all folds except fold j to estimate regression parameter, and use the jth fold to calculated the $\widehat{SPSE}$. Sum across folds.

Choose the model M that minimizes the $\widehat{SPSE}$.

Cross-validation (5-fold)

Will be taught in TMA4300 Computational statistics and will be a backbone in TMA4268 Statistical Learning.
`http://blog-test.goldenhelix.com/wp-content/uploads/2015/04/B-fig-1.jpg`

How to estimate SPSE?

$$\text{SPSE} = \sum_{j=1}^{J} \text{E}((Y_j - \hat{Y}_{jM})^2)$$

Assume we have fitted a model M with $|M|$ regression parameters.

3. Use existing data (only): It can be shown that $\text{E}(\widehat{SPSE}) = SPSE - 2 \mid M \mid \sigma^2$ when used on the same data that was used to make the prediction, so a better estimate for existing data is
$$\widehat{SPSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{iM})^2 + 2|M|\hat{\sigma}^2 = SSE + 2|M|\hat{\sigma}^2$$
where $\hat{\sigma}^2$ is the same for all models M, and is often estimated using the most complex model under study.
4. Other criteria: all have the same form; a first term based on SSE (or R^2) for model M, and a second term penalizing the model complexity.

Choose the model M that minimizes the $\widehat{SPSE}$.

For models with the same model complexity – easy solution: SSE

Estimators for SPSE to be used on the same data as to be used for estimating the model parameters have the same form; a first term based on SSE (or R^2) for model M, and a second term penalizing the model complexity.
If we consider two models with the same model complexity then SSE can be used to choose between these models.

Acid rain (1). Best subset

For $1, \ldots, 7$ covariates: fit all possible models, and report the model with the smallest SSE (given below) for each value for the model complexity. Explain what you see! How many models have been searched for each model complexity?

```
regfit.full=regsubsets(y~.,data=ds)
sumreg <- summary(regfit.full)
Subset selection object
Call: regsubsets.formula(y ~ ., data = ds)
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " "*" " " " "
5  ( 1 ) "*" "*" "*" " " "*" " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

Names: x1: SO_4, x2: $N0_3$, x3: Ca, x4: latent Al, x5: organic substance, x6: area of lake, x7: position of lake (Telemark or Trøndelag).

Popular model choice criteria

R^2 adjusted (corrected)

Mallows' C_p

Akaike Information Criterion (AIC)

Bayesian Information Criterion (BIC)

NB: there is no overall best choice for criterion - all of these are used.

R^2 adjusted (corrected)

$\hat{Y}_i$ is from fitting the regression model M.
Remember, for a regression model (with intercept) we have the SST=SSR+SSE.

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

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2_{\text{adj}} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-p}(1-R^2)$$

Choose the model with the *largest* R^2_{adj}.

"All" statistical software outputs this automatically! However, Fahrmeir et al (2013) believes that the penalty $n - p$ is too small.

Happiness ($n = 39$)

Are love and work the important factors determining happiness?

- y, `happiness`. 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

Data taken from library faraway, data set happy.

Happy

```
> allreg=regsubsets(happy~.,data=happy)
> sumreg <- summary(allreg)
> sumreg
Subset selection object
Call: regsubsets.formula(happy ~ ., data = happy)
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         money sex love work
1  ( 1 ) " "   " " "*"  " "
2  ( 1 ) " "   " " "*"  "*"
3  ( 1 ) "*"   " " "*"  "*"
4  ( 1 ) "*"   "*" "*"  "*"
```

	money	sex	love	work	N	p	R^2	R^2_{adj}
1	0.014				1	0.000747	7.3	4.8
2		−0.130			1	1	0.1	−2.6
3			2.270		1	8.35e-24	61.5	60.5
4				0.990	1	1.36e-13	29.1	27.2
5	0.016	−0.508			2	0.0504	8.8	3.8
6	0.009		2.206		2	8.77e-19	64.5	62.5
7	0.012			0.961	2	3.68e-10	34.6	31.0
8		−0.277	2.279		2	5.55e-18	62.0	59.9
9		0.610		1.079	2	3.48e-09	31.2	27.4
10			1.959	0.511	2	5.75e-20	68.1	66.3
11	0.011	−0.536	2.209		3	9.49e-16	66.2	63.3
12	0.011	0.305		1.009	3	1.84e-07	35.1	29.5
13	0.009		1.902	0.504	3	2.63e-17	70.9	68.4
14		0.108	1.944	0.530	3	2.22e-16	68.1	65.4
15	0.010	−0.149	1.919	0.476	4	9.89e-15	71.0	67.6

Intercept included, $N = p - 1$, p-value for significance of regression.
$R^2 = 1 - \frac{SSE}{SST}$, $R^2_{\text{adj}} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}}$. Which model to prefer?

Mallows' C_p

$\hat{Y}_i$ is from fitting regression model M.
Mallows is the name of a person.

$$C_p = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\hat{\sigma}^2} - n + 2|M|$$

Minimizing Cp gives the same optimal model as minimizing $\widehat{SPSE}$.

See Exam V2015 Problem 3 for an in depth explanation of the theory behind Mallow's Cp.

AIC

Akaike information criterion – one of the most widely used.
Designed for likelihood-based inference.

For a normal regression model:

$$\text{AIC} = n\ln(\hat{\sigma}^2) + 2(|M| + 1)$$

Choose the model with the minimum AIC.

BIC

Bayesian information criterion.

For a normal regression model:

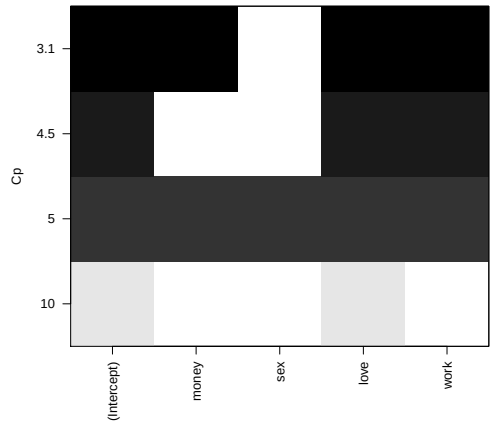
$$\text{BIC} = n\ln(\hat{\sigma}^2) + \ln(n)(|M| + 1)$$

Choose the model with the minimum BIC.

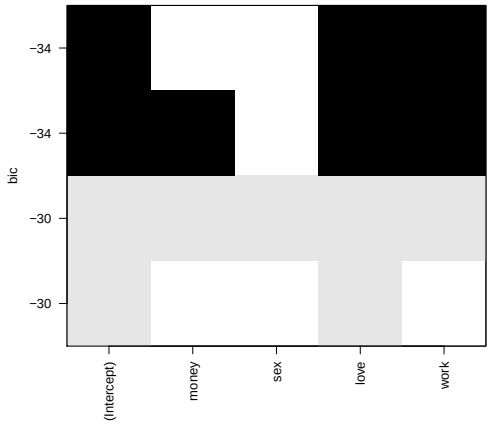
AIC and BIC are motivated in very different ways, but the final result for the normal regression model is very similar.

BIC has a larger penalty than AIC ($\log(n) vs. 2$), and will often give a smaller model (=more parsimonious models) than AIC.

Happy: Mallows' C_p



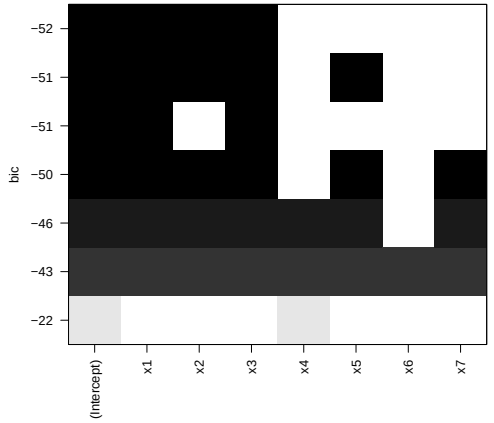
Happy: BIC



Acid rain (2)

```
Call: regsubsets.formula(y ~ ., data = ds)
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " "*" " " " "
5  ( 1 ) "*" "*" "*" " " "*" " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
# to mimic test set:
which.max(sumreg$adjr2) #5
which.min(sumreg$cp) #3
which.min(sumreg$bic) #3
# so, model 3 or 5 is suggested for us
# model 3: x1+x2+x3
# model 5: x1+x2+x3+x5+x7
```


Acid rain, BIC,



Practical use of the model criteria

- All subset selection: use smart "leaps and bounds" algorithm, works fine for number of covariates in the order of 40.
- Forward selection: choose starting model (only intercept), then add one new variable at each step - selected to make the best improvement in the model selection criteria. End when no improvement is made.
- Backward elimination: : choose starting model (full model), then remove one new variable at each step - selected to make the best improvement in the model selection criteria. End when no improvement is made.
- Stepwise selection: combine forward and backward.

Acid rain (3): stepAIC

```
> all=lm(happy~.,data=happy)
> stepAIC(all)
Start:  AIC=9.08
happy ~ money + sex + love + work

        Df Sum of Sq    RSS    AIC
- sex    1     0.142 38.229  7.221
<none>               38.087  9.076
- money  1     3.782 41.869 10.768
- work   1     6.386 44.473 13.122
- love   1    47.272 85.359 38.549

Step:  AIC=7.22
happy ~ money + love + work

        Df Sum of Sq    RSS    AIC
<none>               38.229  7.221
- money  1     3.723 41.952  8.846
- work   1     8.410 46.639 12.976
- love   1    47.742 85.971 36.828

Call:
lm(formula = happy ~ money + love + work, data = happy)

Coefficients:
(Intercept)        money         love         work
  -0.185936     0.008959     1.901709     0.503602
```


Acid rain (4): Forward

```
regfitF=regsubsets(y~.,data=ds,method="forward")
sumregF <- summary(regfitF)
Selection Algorithm: forward
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) " " " " "*" "*" " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" "*" " " " " " "
5  ( 1 ) "*" "*" "*" "*" "*" " " " "
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
which.max(sumregF$adjr2)#5
which.min(sumregF$cp) #3
which.min(sumregF$bic) #3
```


Acid rain (5): Backward

```
regfitB=regsubsets(y~.,data=ds,method="backward")
sumregB <- summary(regfitB)
Selection Algorithm: backward
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) "*" " " " " "*" " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " "*" " "
5  ( 1 ) "*" "*" "*" " " " " "*" " "
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
which.max(sumregB$adjr)#5
# backward finds same as best subset
which.min(sumregB$cp) #3
```

Model diagnosis

- Influential observations and outliers: impact of specific observations on model fit.
- Collinearity analysis: Highly correlated variables cause imprecise estimation of the regression parameters. (Why? Look at diagonal elements of $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and look back to Problem 2 in the start of this lecture.)
- Examination of model assumptions: residual plots!

Influential observations– and outliers

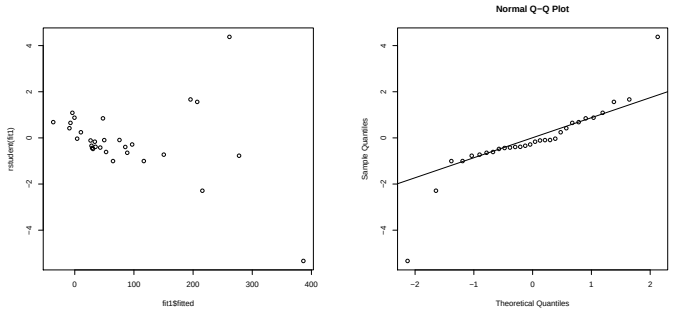
- Observations that significantly affect inferences drawn from the data are said to be influential.
- The leverage, h_{ii}, associated with the ith datapoint measures "how far the ith observation is from the other $n-1$ observations".
- Methods for assessing influential observations may be be based on change in $\boldsymbol{\beta}$ estimate when observations are deleted.
- Always investigate possible causes of an influential observation (if possible).
- Cook's distance can be used to identify influential observations.
- Robust methods (median,quantile regression) can be useful.

Want to understand more? Read for yourself in Fahrmeir et al (2013): p 160-166.

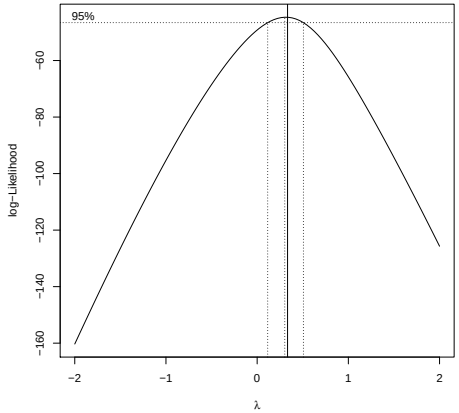
Transformations

- Multiplicative or additive model?
- Box–Cox transform with profile likelihood.
- Stabilizing the variance.

Galapagos islands, Model A, Exam V2014 Problem 2

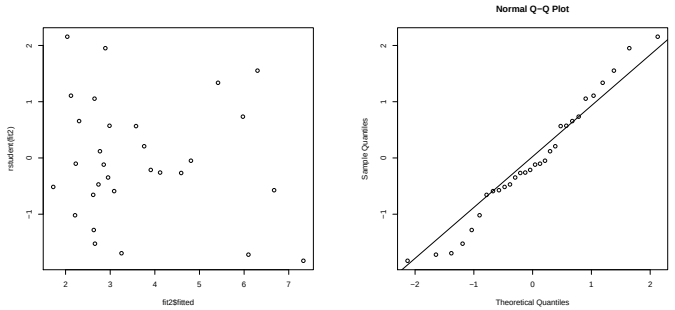


Box–Cox plot



Box–Cox transformation plot based on Model A for the Galapagos data set, RecEx4. Line at $x = 1/3$.

Galapagos islands, Model B, Exam V2014 Problem 2



Approximation of E and Var for nonlinear functions

- Have RV X, with mean $\mathrm{E}(X) = \mu$ and some variance $\mathrm{Var}(X)$.
- Want to look at a nonlinear function of X, called $g(X)$.
- Aim: find an approximation to $\mathrm{E}(g(X))$ and $\mathrm{Var}(g(X))$.
- And, the same for two RVs X_1 and X_2 with $g(X_1, X_2)$.

Example ln of BMI

Looking at residual plots from a regression model the conclusion was to analyse data of *BMI* on the natural logarithmic scale. After a regression model was fitted the predicted value for the ln(BMI) for a specific combination of the covariates was found to be 3.2151 with an estimated standard deviation of 0.1656. Use approximate methods to arrive at an estimate of the predicted value and estimated standard deviation on the original scale, $\mathrm{kg/m^2}$, and not on the logarithmic scale.

$\mathrm{E}(g(X)$ and $\mathrm{Var}(g(X))$

- Let $g(X)$ be a general function. When is $\mathrm{E}(g(X)) = g(\mathrm{E}(X))$?
  - When $g(X)$ is a linear function of X.
- What can we do if this is not the case?
  - We can calculate $\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$ when X is continuous, or a version thereof in the discrete case,
  - or if g is monotone we can use the transformations formula to find the distribution of $Y = g(X)$ and then calculate $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$, if possible.
- What if we only know $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ and not $f(x)$?
  - Use a Taylor series approximation of $g(X)$ around $g(\mu)$. g need to be differentiable.

Univariate function

First order Taylor approximation of $g(X)$ around μ.

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu)$$

This leads to the following approximations:

$$\mathrm{E}(g(X)) \approx g(\mu)$$
$$\mathrm{Var}(g(X)) \approx [g'(\mu)]^2\mathrm{Var}(X)$$

Treatment of tennis elbow (exam TMA4255 V2012, 3b)

The term *tennis elbow* is used to describe a state of inflammation in the elbow, causing pain. This injury is common in people who play racquet sports, however, any activity that involves repetitive twisting of the wrist (like using a screwdriver) can lead to this condition. The condition may also be due to constant computer keyboard and mouse use.
In a randomized clinical study the aim was to compare three different methods for treatment of tennis elbow,

- A: physiotherapy intervention,
- B: corticosteroid injections and
- C: wait-and-see (the patients in the wait-and-see group did not get any treatment but was told to use the elbow as little as possible).

Treatment of tennis elbow (cont.)

We will look at the short-term effect of treatment by studying measurements at 6 weeks. All patients participating in the study only had one affected arm.
We will look at the outcome measure called *pain-free grip force*. This was measured by a digital grip dynamometer and normalized to the grip force of the unaffected arm. A pain-free grip force of 100 would mean that the affected and the unaffected arm performed equally good.

Summary statistics for each of the treatment groups.

Treatment	Sample size	Average	Standard deviation
A (physiotherapy)	63	70.2	25.4
B (injection)	65	83.6	22.9
C (wait-and-see)	60	51.8	23.0
Total	188	69.0	

Example 2: Exam TMA4255 V2012 3d (fraction)

Let μ_A be the expected pain-free grip force for a population where the physiotherapy intervention treatment is used to treat tennis elbow, and μ_C be the expected pain-free grip force for a population where the wait-and-see treatment is used. Define the relative difference between these two expected values as

$$\gamma = \frac{\mu_A - \mu_C}{\mu_C}.$$

This can be interpreted as the expected relative gain by using physiotherapy instead of wait-and-see. Based on two independent random samples of size n_A and n_C from the physiotherapy and wait-and-see treatment groups, respectively, suggest an estimator, $\hat{\gamma}$, for γ.
Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

Bivariate function: first order Taylor

X_1 is a RV with $\mu = \mathrm{E}(X_2)$ and X_2 is a RV with $\mu_2 = \mathrm{E}(X_2)$.
Let g be a bivariate function of X_1 and X_2, and define

$$g_1'(\mu_1, \mu_2) = \frac{\partial g(x_1, x_2)}{\partial x_1} \,|_{x_1=\mu_1, x_2=\mu_2}$$
$$g_2'(\mu_1, \mu_2) = \frac{\partial g(x_1, x_2)}{\partial x_2} \,|_{x_1=\mu_1, x_2=\mu_2}$$

First order Taylor approximation:

$$g(X_1, X_2) \approx g(\mu_1, \mu_2) + g_1'(\mu_1, \mu_2)(X_1 - \mu_1) + g_2'(\mu_1, \mu_2)(X_2 - \mu_2)$$

Bivariate function: first order Taylor

$$\mathrm{E}(g(X_1, X_2)) \approx g(\mu_1, \mu_2)$$
$$\mathrm{Var}(g(X_1, X_2)) \approx [g_1'(\mu_1, \mu_2)]^2 \mathrm{Var}(X_1) + [g_2'(\mu_1, \mu_2)]^2 \mathrm{Var}(X_2) + 2 \cdot g_1'(\mu_1, \mu_2) \cdot g_2'(\mu_1, \mu_2) \mathrm{Cov}(X_1, X_2)$$

Multivariate version

From *Tabeller og formler i statistikk*.

Rekkeutvikling

En første ordens Taylorutvikling av funksjonen $g(X_1, \ldots, X_n)$ omkring $g(\mu_1, \ldots, \mu_n)$, der $\mathrm{E}(X_i) = \mu_i$, $i = 1, \ldots, n$, gir approksimasjonene

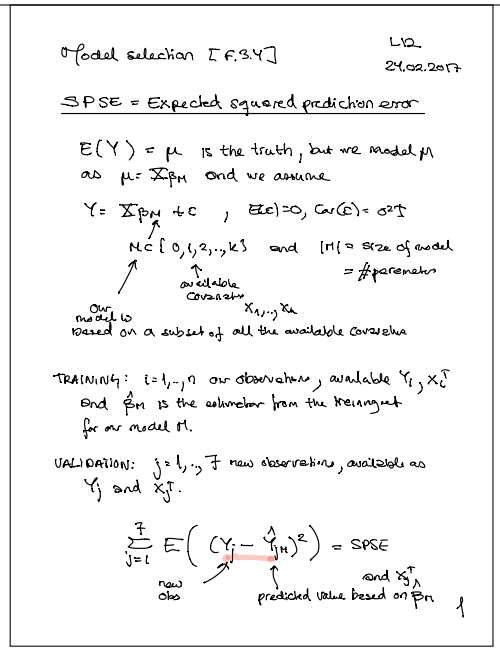
$$\mathrm{E}[g(X_1, \ldots, X_n)] \approx g(\mu_1, \ldots, \mu_n),$$

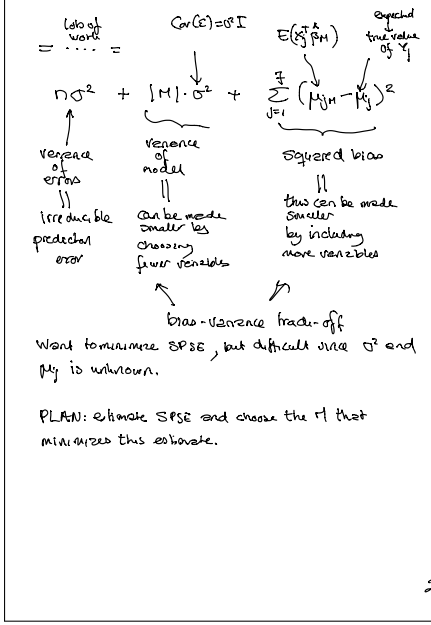
$$\mathrm{Var}[g(X_1, \ldots, X_n)] \approx \sum_{i=1}^{n} \left(\frac{\partial g(\mu_1, \ldots, \mu_n)}{\partial \mu_i} \right)^2 \mathrm{Var}(X_i) + 2 \sum_{i>j} \frac{\partial g}{\partial \mu_i} \frac{\partial g}{\partial \mu_j} \mathrm{Cov}(X_i, X_j).$$

Today

- Choosing between models of *equal model complexity*: choose the model with the minimum SSE.
- Choosing between models of *different model complexity*: Model selection based on penalized criteria (Mallows Cp, R^2_{adj},AIC and BIC). Try out on RecEx4 and Compulsory Exercise 2.
- BoxCox transformation: see RecEx4.
- Work for for yourself: Taylor solution to E and Var of nonlinear function, useful when you want to look at transformations of the data or functions of parameter estimates.

Summary of Part 2 in Kahoot!





Finding the best model ← all subsets method

(smallest SPSE)

1) Have k covariates that might be used

$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$

How many possible models can I make (want to have intercept)?

$2 \cdot 2 \cdot 2 \cdots 2 = 2^k$ possible models

Fit all possible 2^k models.

2) For $M = \{0, 1, 2, \ldots, k\}$ choose the model with the smallest SSE.

Ex: Acid rain: $k = 7 \Rightarrow$ total $2^7 = 128$ possible models

Complexity	Search method	best model
$\|M\| = 1$:	$\binom{7}{1} = 7$	x_4 (Al)
$\|M\| = 2$:	$\binom{7}{2} = 21$	x_1, x_3
$\|M\| = 3$:	$\binom{7}{3} = 35$	x_1, x_2, x_3
$\|M\| = 4$:	$\binom{7}{4} = 35$	x_1, x_2, x_3, x_5
$\|M\| = 5$:	$\binom{7}{5}$	x_1, x_2, x_3, x_5, x_7
$\|M\| = 6$:	$\binom{7}{6}$	$x_1, x_2, x_3, x_4, x_5, x_7$
$\|M\| = 7$:	$\binom{7}{7} = 1$	$x_1, x_2, x_3, x_4, x_5, x_6, x_7$
	2^7	

Ex: Happy: BIC best model = love + work

Homework: slide 24 Acid rain

Transformation of response and predictors might improve the fit of the regression model.

The BoxCox transform

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

(g_λ: class of function)

For $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$ the best value of λ on based on maximizing the likelihood

$$l(\lambda) = -\frac{n}{2} \ln\left(\frac{SSE_\lambda}{n}\right) - (\lambda - 1) \sum_{i=1}^{n} \ln y_i$$

problem

SSE_λ is the SSE when $g_\lambda(Y)$ is the response

R: boxcox(fit), see plot.

3) Now we need to choose between these $k+1$ models found in 2). Which criterion should I use?

i) $R^2_{adj} = 1 - \frac{\frac{SSE}{(n-p)}}{\frac{SST}{n-1}}$ $\quad |M| + p = k+1$

Ex: Happiness:

$\|M\| = 1$:	only intercept		
$\|M\| = 2$:	love	(3)	: 60.5
$\|M\| = 3$:	love + work	(10)	: 66.3
$\|M\| = 4$:	love + work + money	(78)	: 68.4
$\|M\| = 5$:	love + work + sex + money	(15):	67.6

ii) Mallows' $C_p = \frac{SSE}{\hat{\sigma}^2_{Full}} - n + 2|M|$

is $SPSE = SSE + 2|M| \cdot \hat{\sigma}^2_{Full}$ ← give same result.

iii) $AIC = n \cdot \ln(\hat{\sigma}^2) + 2(|M| + 1)$

$\hat{\sigma}^2 = \frac{SSE}{n - k_{full}}$

iv) $BIC = n \cdot \ln(\hat{\sigma}^2) + \ln(n)(|M| + 1)$

BIC gives more penalty than AIC to large models.

TMA4267 Linear statistical models

Part 2: Linear regression

February 20, 2017

Normal equations

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } E(\boldsymbol{\varepsilon}) = \boldsymbol{0} \text{ and } Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$$

Which of the following are *the normal equations*?

A $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$

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

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

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

The hat matrix

Design matrix $\boldsymbol{X}$ has n rows and p linearly independent columns. $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the hat-matrix.

Which of the following statements are NOT true?

A $\boldsymbol{H} = \boldsymbol{H}^T = \boldsymbol{H}^2$

B $\text{rank}(\boldsymbol{H}) = p$

C $\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}$

D $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$

Estimator for σ^2

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } E(\boldsymbol{\varepsilon}) = \boldsymbol{0} \text{ and } Cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I} \\ \boldsymbol{H} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \end{aligned}$$

An unbiased estimator for σ^2 is:

A SSE$/n$

B $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}/(n-p)$

C $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Y}/(n-p)$

D $(\boldsymbol{X}^T\boldsymbol{X})^{-1}SSE/n$

Inference about $\boldsymbol{\beta}$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

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

What are the properties of $\hat{\boldsymbol{\beta}}$?

A Chi-squared distributed with $n - p$ degrees of freedom.

B Chi-squared distributed with p degrees of freedom.

C Multivariate normal with covariance matrix $(\boldsymbol{I} - \boldsymbol{H})\sigma^2$.

D Multivariate normal with covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$.

Happiness=money+sex+love+work

	Estimate	Std. Error	t value	Pr(>|t|)
money	0.009578	0.005213	1.837	0.0749
sex	-0.149008	0.418525	-0.356	0.7240
love	1.919279	0.295451	6.496	1.97e-07
work	0.476079	0.199389	2.388	0.0227

Which of the regression coefficient estimates has the largest estimated variance?

A money **B** sex

C love **D** work

Happiness

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.072081	0.852543	-0.085	0.9331
money	0.009578	0.005213	1.837	0.0749
sex	-0.149008	0.418525	-0.356	0.7240
love	1.919279	0.295451	6.496	1.97e-07
work	0.476079	0.199389	2.388	0.0227

For which β_j would we reject the null hypothesis $\beta_j = 0$ at significance level 1%?

A money **B** sex

C love **D** work

Happiness=money+sex+love+work

The R^2 for the happiness-regression model is 71%. What does that mean?

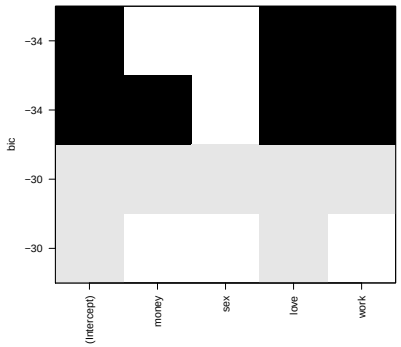
A The regression is significant for significance level 71%

B The regression explains 71% of the variability in the data

C The estimate for the variance σ^2 is 0.71

D The covariates have a correlation of 0.71

Best model

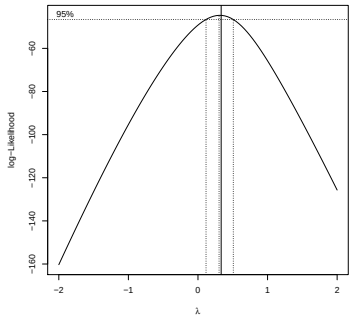


Which model does the BIC criterion report to be the best?

A love+work **B** love

C money+love+work **D** money+sex+love+work

What is this plot used for?



A Check residuals

B Assess normality of residuals

C Assess linearity

D Find transform of response

Correct?

Are you sure you want to read the correct answers? Maybe try first? The answers are explained on the next two slides.

Answers

1. C: The normal equation $(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$ is before you solve for $\hat{\boldsymbol{\beta}}$.
2. C: The hat matrix is symmetric and idempotent (so A is ok), and has rank p, but the reason for the name of the hat matrix is that is puts the hat on the $\boldsymbol{Y}$ so $\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{Y}}$. We know that for symmetric projection matrices the two matrices $\boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})$ are orthogonal so the product must be zero.

Answers

3. B: Since SSE has mean $(n-p)\sigma^2$, then SSE/(n-p) must be an unbiased estimator for σ^2. We know that $(\boldsymbol{I} - \boldsymbol{H})$ projects onto the space othogonal to the column space of the designmatrix, so that must have to do with SSE.
4. D: We know that linear combinations of multivariate normal random vectors are also multivariate normal (so the chisquare is not suitable). The residuals have $(\boldsymbol{I} - \boldsymbol{H})$ as part of their covariance matrix, but $\hat{\boldsymbol{\beta}}$ has not.

Answers

5. B: Sex has the largest estimated variance for regression estimate.
6. B: R^2 gives the percent of variability explained.
7. C: only love is significant on level 1%, since this is the only p-value below 0.01 (last column).
8. A: love+work has smallest BIC.
9. D: Box-Cox plot used to find transformation of response.