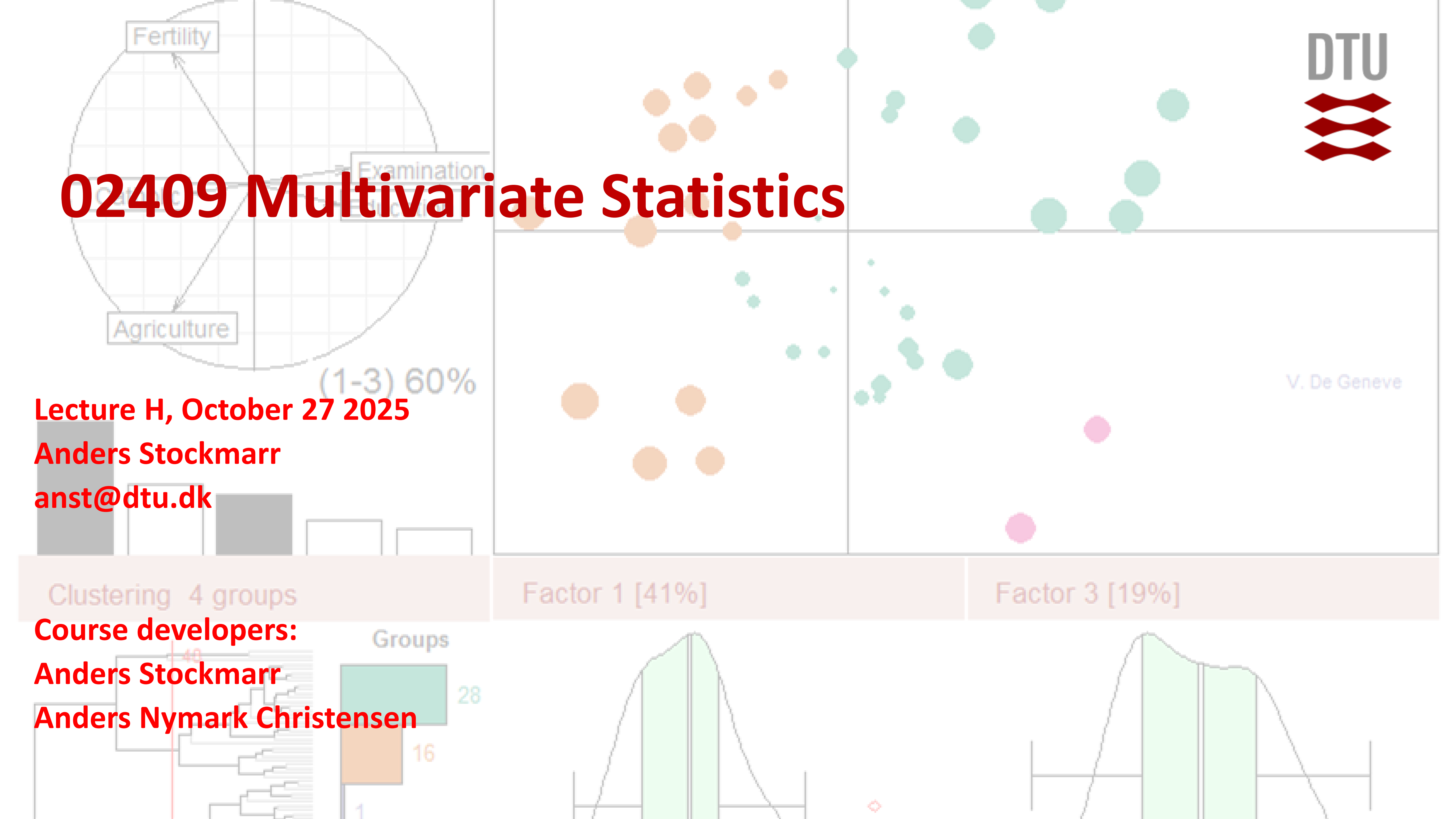
02409 Multivariate Statistics

Lecture H, October 27 2025
Anders Stockmarr
anst@dtu.dk

Course developers:
Anders Stockmarr
Anders Nymark Christensen



Agenda

- GLM
  - Estimation
  - Testing
  - Confidence and prediction intervals

The General Linear Model - GLM

Let $Y \in \mathbb{R}^n$ be a normal distributed random variable:

$$Y \sim N(\mu, \sigma^2 \Sigma)$$

where Σ and observations $\mathbf{x}$ is assumed known.

A linear model describe μ as a linear combination of the observations $\mathbf{x}$ and some unknown parameters $\theta \in \mathbb{R}^k \subset \mathbb{R}^n$.

$$\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\theta}$$

or

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

The General Linear Model - GLM

We also see that this

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

Is equivalent to this

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

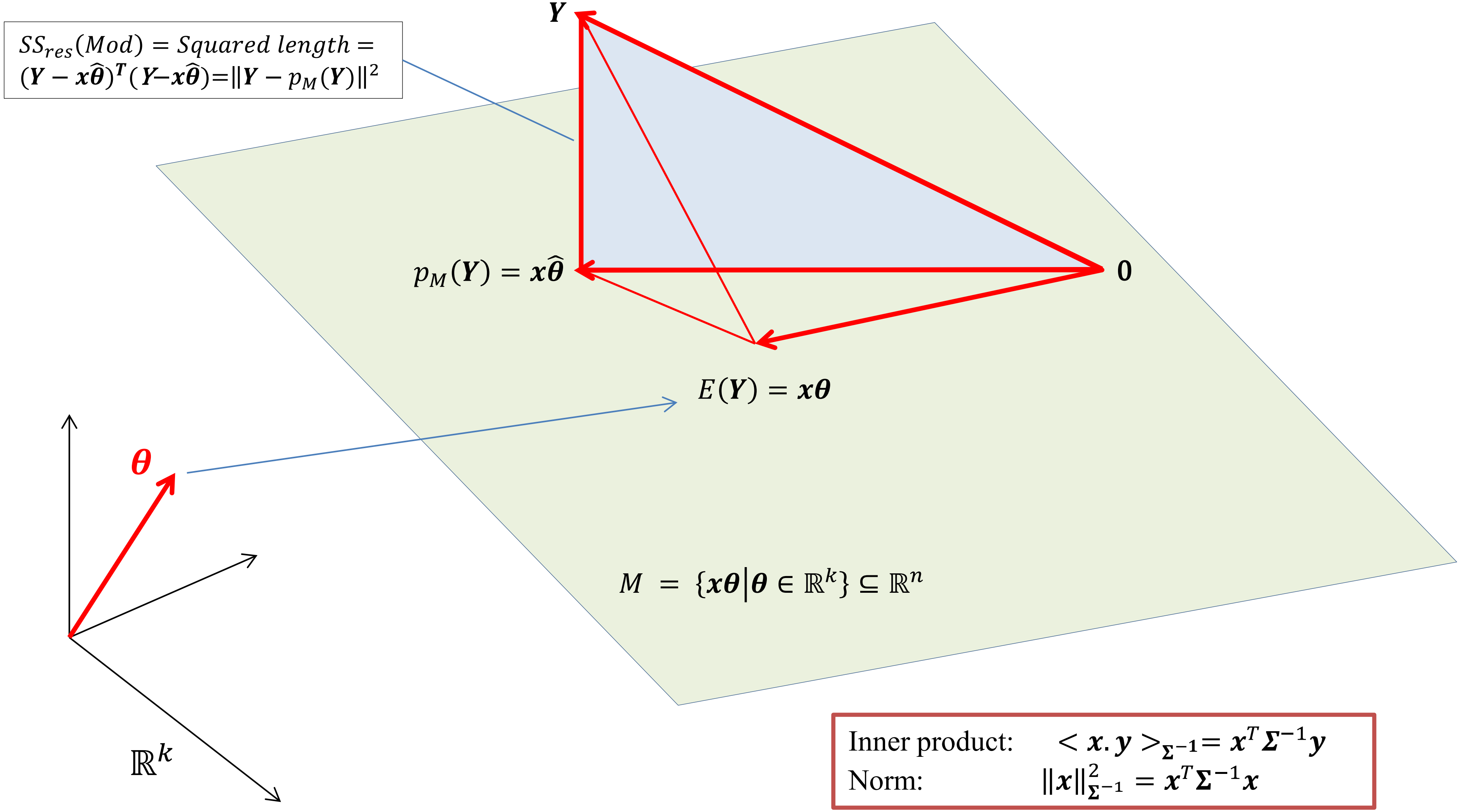
i.e.

$$\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \boldsymbol{\Sigma}), \qquad \boldsymbol{x}, \boldsymbol{\Sigma} \text{ known}$$

Geometry of (estimation in) GLM



Inner product: $< \boldsymbol{x}.\boldsymbol{y} >_{\boldsymbol{\Sigma}^{-1}} = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$

Norm: $\|\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}^{-1}} = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$

Estimation in GLM

Theorem 2.3

Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $Y \sim \mathrm{N}_n(\mathbf{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$) onto M, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called ***normal equation(s)***

$$(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

If $\mathbf{x}$ has full rank k, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}Y,$$

and being a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

Estimation in GLM

What if X doesn't have full rank?

Assume that $rk(x)=r<k$.

- The design matrix X contains unnessesary information; throw away k-r irrelevant variables. In other words, clean up the mess you have made…
- If this is not doable; use a generalized inverse to $X^T\Sigma^{-1}X$.

Generalized Inverse

- Take $G = X^T\Sigma^{-1}X$.
- A generalized inverse G^- satisfies

$$GG^-G = G$$

In practise:

Choose T such that

$$TGT^T = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

Define

$$G^- = T^T \begin{bmatrix} 1/\lambda_1 & & & & & \\ & \ddots & & & & \\ & & 1/\lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} T$$

Obviously, $GG^-G = G$.

Generalized Inverse

- Define

$$\widehat{X\theta} = X(X^T\Sigma^{-1}X)^{-}X^T\Sigma^{-1}Y$$

$$\begin{aligned} E\left(\widehat{X\theta}\right) &= X(X^T\Sigma^{-1}X)^{-}X^T\Sigma^{-1}E(Y) \\ &= X(X^T\Sigma^{-1}X)^{-}X^T\Sigma^{-1}X\theta \\ &= X\theta \end{aligned}$$

- θ **cannot be uniquely detemined**, but the mean of Y can.
- Makes the model *testable*.

Generalized Inverse

- Generalized inverse in practice:

Use an appropriate tolerance

```
G<-t(X)%*%solve(Sigma)%*%X

my.eigenvalues<-eigen(G)$values[eigen(G)$values>1e-6]
r<-length(my.eigenvalues)
T<-eigen(G)$vectors

G.geninv<-t(T)%*%diag(c(1/my.eigenvalues,rep(0,dim(X)[2]-r)))%*%T
```

Warning

- Do **NOT** construct an artificial estimator $\hat{\theta}$ through the side-subspace technique illustrated in Section 2.1.3.
- Why not? Tecnically, such an estimator works just as well as the method outlined above?
- Yes, for **testing**;

BUT - the estimator $\hat{\theta}$ is **not interpretable**
as measures of associations with the underlying explanatory variables!

- The values of $\hat{\theta}$ are **arbitrary**; a different parametrization of the side-subspace will give different values of $\hat{\theta}$.
- For testing, **we don't need any** $\widehat{\boldsymbol{\theta}}$, only $\widehat{X\theta}$.
- **Parameter estimates must be interpretable as such; otherwise they loose their meaning.**

Estimation in GLM

Theorem 2.5

Let the situation be as above. The maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\|Y - \mathbf{x}\hat{\theta}\|^2 = \frac{1}{n}(Y - \mathbf{x}\hat{\theta})^T\Sigma^{-1}(Y - \mathbf{x}\hat{\theta}).$$

The unbiased estimator of σ^2 is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}\|Y - \mathbf{x}\hat{\theta}\|^2 \\ &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}(Y - \mathbf{x}\hat{\theta})^T\Sigma^{-1}(Y - \mathbf{x}\hat{\theta})\end{aligned}$$

where $\mathbf{x}\hat{\theta}$ is the maximum likelihood estimator of $\mathrm{E}(Y)$. The following holds

$$\hat{\sigma}^2 \sim \sigma^2\chi^2(n - \mathrm{rk}(\mathbf{x}))/(n - \mathrm{rk}(\mathbf{x}))$$

and $\hat{\sigma}^2$ is independent of the maximum likelihood estimator of the expected value and is therefore independent of $\hat{\theta}$.

GLM Main Example 2.8

Example 2.8

In the production of a certain synthetic product two raw materials A and B are mainly used. The quality of the end product can be described by a stochastic variable which is normally distributed with mean value μ and variance σ^2. The mean-value is known to depend linearly on the added amount of A and B respectively i.e.

$$\mu = x_A\theta_A + x_B\theta_B,$$

Experiment	Content of A	Content of B	Outcome
1	100%	0%	90
2	0%	100%	30
3	50%	50%	75

GLM Example 2.8

The single experiments are assumed to be stochastically independent. The simultaneous distribution of the experimental results Y_1, Y_2, Y_3 is then a three dimensional normal distribution with mean value

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_{\mathrm{A}} \\ \theta_{\mathrm{B}} \end{bmatrix} = \mathbf{x}\,\boldsymbol{\theta},$$

and variance-covariance matrix $\sigma^2\mathbf{I}$.

GLM Example 2.8

Theorem 2.3 $$\hat{\theta} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}Y,$$

We have

$$\mathbf{x}^T\mathbf{x} = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{5}{4} \end{bmatrix} \quad \Rightarrow \quad (\mathbf{x}^T\mathbf{x})^{-1} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix},$$

and

$$\mathbf{x}^T\mathbf{y} = \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix},$$

giving

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{6}y_1 - \frac{1}{6}y_2 + \frac{1}{3}y_3 \\ -\frac{1}{6}y_1 + \frac{5}{6}y_2 + \frac{1}{3}y_3 \end{bmatrix}$$

Observations

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \mathbf{90} \\ \mathbf{30} \\ \mathbf{75} \end{pmatrix}$$

GLM Example 2.8

Estimated parameters

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \begin{bmatrix} y_1 + \frac{1}{2} y_3 \\ y_2 + \frac{1}{2} y_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{6} y_1 - \frac{1}{6} y_2 + \frac{1}{3} y_3 \\ -\frac{1}{6} y_1 + \frac{5}{6} y_2 + \frac{1}{3} y_3 \end{bmatrix} = \begin{bmatrix} 95 \\ 35 \end{bmatrix}$$

$$\hat{\mathrm{E}}(\boldsymbol{Y}) = \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{pmatrix} 95 \\ 35 \\ 65 \end{pmatrix}$$

Estimated value

$$\boldsymbol{Y} - \hat{\mathrm{E}}(\boldsymbol{Y}) = \boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{pmatrix} -5 \\ -5 \\ 10 \end{pmatrix}$$

Residual

Theorem 2.5

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n - \mathrm{rk}(\mathbf{x})} \|\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 \\ &= \frac{1}{n - \mathrm{rk}(\mathbf{x})} (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) \end{aligned}$$

This gives the residual sum of squares

$$(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = 25 + 25 + 100 = 150,$$

$$\frac{1}{3-2} 150 = 150$$

GLM Example 2.8

```
> synprod <- data.frame("Acontent"=c(1,0,0.5),
+                       "Bcontent"=c(0,1,0.5),
+                       "Qual"=c(90,30,75))

> glmSyn <- lm(Qual ~ Acontent + Bcontent - 1, data=synprod)

> summary(glmSyn)

Call:
lm(formula = Qual ~ Acontent + Bcontent - 1, data = synprod)

Residuals:
 1  2  3
-5 -5 10

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Acontent    95.00      11.18   8.497   0.0746 .
Bcontent    35.00      11.18   3.130   0.1968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.25 on 1 degrees of freedom
Multiple R-squared:  0.9897,    Adjusted R-squared:  0.9692
F-statistic: 48.25 on 2 and 1 DF,  p-value: 0.1013
```

Residuals: $\mathrm{R_M} = Y - \hat{E}(Y) = Y - X\hat{\theta} = \begin{bmatrix} -5 \\ -5 \\ 10 \end{bmatrix}$

Parameter estimates: $\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix}$

$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{150}$

$n - k = 1$

GLM Example 2.8

- **Simpler model**: Perhaps Bcontent doesn't play a rôle for Quality: Mean space H spanned by $x_A = \begin{bmatrix} 100 \\ 0 \\ 50 \end{bmatrix}$.

Residuals: $\mathrm{R}_H = Y - \hat{E}_A(Y) = Y - X_A\hat{\theta} = \begin{bmatrix} -12 \\ 30 \\ 24 \end{bmatrix}$

Parameter estimate: $\hat{\theta}_A$

$\hat{\hat{\sigma}} = \sqrt{\hat{\hat{\sigma}}^2} = \sqrt{810}$

$n - k = 2$

```
> glmSyn <- lm(Qual ~ Acontent - 1, data=synprod)

Residuals:
  1   2   3
-12  30  24

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Acontent   102.00      25.46   4.007    0.057 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.46 on 2 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8338
F-statistic: 16.06 on 1 and 2 DF,  p-value: 0.05701
```

GLM Example 2.8

$$H_0: \ \theta_B = 0$$

- H_0 indikates that $P_M(Y) = P_H(Y)$; ie. essentially that

$$P_M Y - P_H Y = P_{M \cap H^{\perp}} Y = \varepsilon$$

Under H_0:

$$P_{M \cap H^{\perp}} Y \sim \mathrm{N}\left(0, \sigma^2 P_{M \cap H^{\perp}}\right), \qquad \|P_M Y - P_H Y\|^2 \sim \sigma^2 \chi^2(1)$$

Because

$dim(M \cap H^{\perp}) = 1.$

GLM Example 2.8

Note that

$$< Y - P_M Y, P_M Y - P_H Y >_{I_3^{-1}} =< P_{M^\perp} Y, P_{M \cap H^\perp} Y >= 0$$

So that $Y - P_M Y, P_M Y - P_{M_A} Y$ are **stochastically independent** of each other.

Now,

$$Q_F := \frac{\frac{1}{1}\|P_M Y - P_H Y\|^2}{\frac{1}{3-2}\|Y - P_M Y\|^2} \sim \frac{\frac{1}{1}\sigma^2 \chi^2(1)}{\frac{1}{3-2}\sigma^2 \chi^2(1)} = \frac{\frac{1}{1}\chi^2(1)}{\frac{1}{3-2}\chi^2(1)} = F(1,1)$$

A Sidestep

General testing:

$$H_0 : \mu \in H \subset M \quad vs. H_1 : \mu \in M \backslash \mathrm{H}$$

taking $n = \#\, obs$, $k = dim(M), k_1 = dim(H)$:

$$Q_F := \frac{\frac{1}{k-k_1}\|P_M Y - P_H Y\|^2}{\frac{1}{n-k}\|Y - P_M Y\|^2} \sim \frac{\frac{1}{k-k_1}\sigma^2\chi^2(k-k_1)}{\frac{1}{n-k}\sigma^2\chi^2(n-k)} = \frac{\frac{1}{k-k_1}\chi^2(k-k_1)}{\frac{1}{n-k}\chi^2(n-k)} = F(k-k_1, n-k)$$

A Sidestep

- Note that with Q the Likelihood Ratio test statistic, and taking n =# observations, $k = dim(M), k_1 = dim(H)$ arbitrary,

$$Q = \frac{L(\hat{\theta}_A, \hat{\hat{\sigma}}^2)}{L(\hat{\theta}, \hat{\sigma}^2)} = \left(\frac{(2\pi)^{-n/2}\left(\hat{\hat{\sigma}}^2\right)^{-n/2}\exp(-\frac{1}{2\hat{\hat{\sigma}}^2}\|Y - P_H Y\|^2}{(2\pi)^{-n/2}(\hat{\sigma}^2)^{-n/2}\exp(-\frac{1}{2\hat{\sigma}^2}\|Y - P_M Y\|^2} \right)$$

$$= \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{-n/2} \exp\left(-\frac{n-k_1}{2} + \frac{n-k}{2}\right)$$

$$= c\left(\frac{n-k}{n-k_1}\frac{\|Y - P_H Y\|^2}{\|Y - P_M Y\|^2}\right)^{-n/2}$$

$$= c\left(\frac{n-k}{n-k_1}\left(\frac{\|Y - P_M Y\|^2 + \|P_M Y - P_H Y\|^2}{\|Y - P_M Y\|^2}\right)\right)^{-\frac{n}{2}}$$

$$= c\left(\frac{n-k}{n-k_1}\left(1 + \frac{\|P_M Y - P_H Y\|^2}{\|Y - P_M Y\|^2}\right)\right)^{-\frac{n}{2}} = c\left(\frac{n-k}{n-k_1} + \frac{k-k_1}{n-k_1}Q_F\right)^{-\frac{n}{2}}$$

- Thus, Q with small values critical, **is equivalent to $\boldsymbol{Q_F}$** with large values critical.

A Sidestep

Theorem 2.21

Let the situation be as above. Then the likelihood ratio test at level α of testing

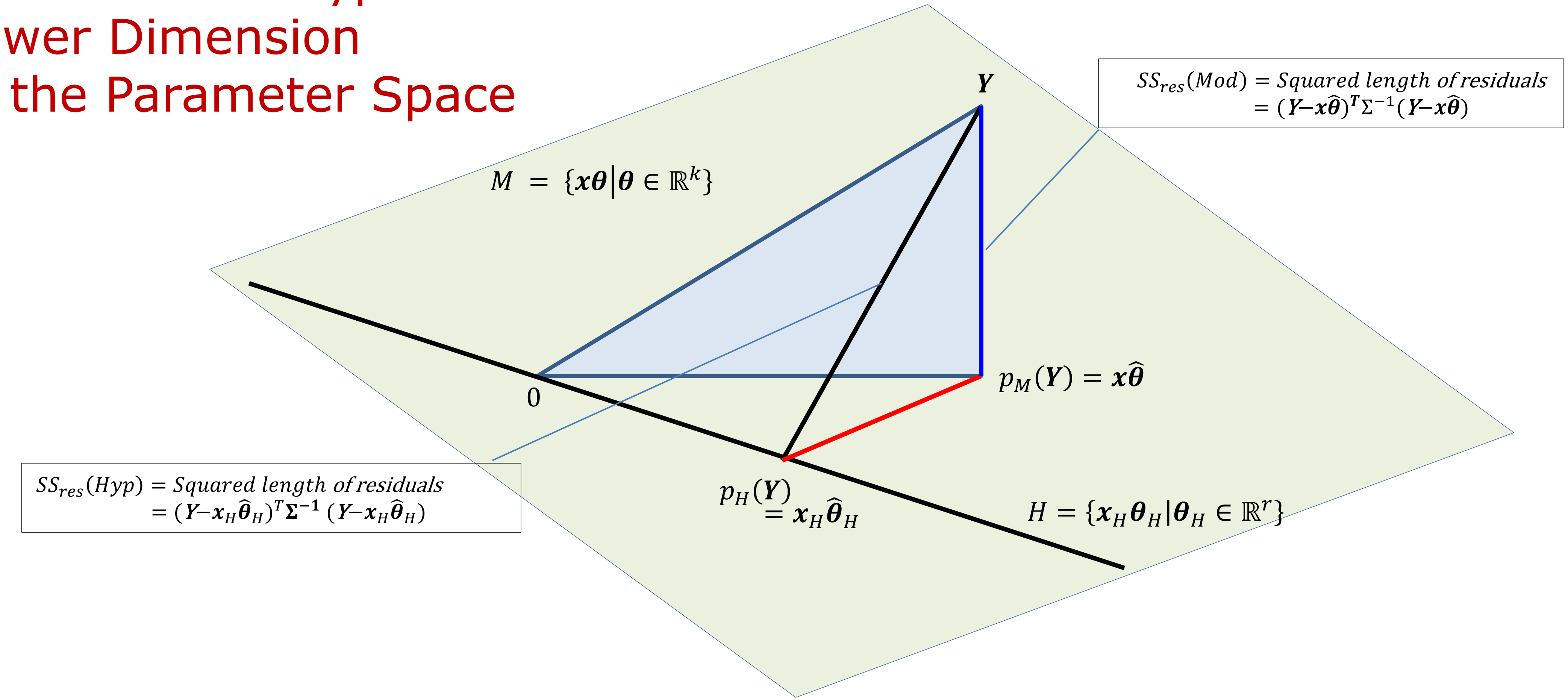
$$H_0 : \mu \in H \quad \text{versus} \quad H_1 : \mu \in M \setminus H,$$

is equivalent to the test given by the critical region

$$C_\alpha = \{(y_1, \ldots, y_n) | \frac{\|p_M(y) - p_H(y)\|^2 / (k-r)}{\|y - p_M(y)\|^2 / (n-k)} > F(k-r, n-k)_{1-\alpha}\}.$$

- Note that the proof in the book is flawed; two errors make up for each other (try to spot them).

Test of Linear Hypotheses: Lower Dimension of the Parameter Space



F-Test statistic for $H_0: E(\boldsymbol{Y}) \in H$ against $H_1: E(\boldsymbol{Y}) \in M \backslash H$:

$$Q_F = \frac{\|p_M(Y) - p_H(Y)\|^2_{\Sigma^{-1}}/(k-r)}{\|Y - p_M(Y)\|^2_{\Sigma^{-1}}/(n-k)}$$

$$= \frac{(SS_{res}(Hyp) - SS_{res}(Mod))/(DF_{res}(Hyp) - DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)}$$

GLM Example 2.8

$$Q_F := \frac{\frac{1}{1}\|P_M Y - P_H Y\|^2}{\frac{1}{3-2}\|Y - P_M Y\|^2}$$

$$\|P_M Y - P_H Y\|^2 = \|R_M - R_H\|^2 = \left\|\begin{bmatrix}-5\\-5\\10\end{bmatrix} - \begin{bmatrix}-12\\30\\24\end{bmatrix}\right\|^2$$

$$= \left\|\begin{bmatrix}-17\\-35\\-44\end{bmatrix}\right\|^2 = 17^2 + 35^2 + 44^2 = 1470$$

$$\|Y - P_M Y\|^2 = \|R_M\|^2 = \left\|\begin{bmatrix}-5\\-5\\10\end{bmatrix}\right\|^2 = 5^2 + 5^2 + 10^2 = 150$$

$$Q_F = \frac{\frac{1}{1}}{\frac{1}{3-2}}\frac{1470}{150} = 9.8, \quad p = P(Q_F > 9.8) = \texttt{1-pf(9.8,1,1)=0.20}$$

GLM Example 2.8

```
> anova(glmSyn)
Analysis of Variance Table

Response: Qual
          Df Sum Sq Mean Sq F value  Pr(>F)
Acontent   1  13005   13005    86.7 0.06811 .
Bcontent   1   1470    1470     9.8 0.19684
Residuals  1    150     150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\|P_H Y\|^2$

$\|P_M Y - P_H Y\|^2$

$\|Y - P_M Y\|^2$

$$Y = Y - P_M Y + P_M Y - P_H Y + P_H Y$$

$$\|Y\|^2 = \|Y - P_M Y\|^2 + \|P_M Y - P_H Y\|^2 + \|P_H Y\|^2$$

$$14625 = 150 + 1470 + 13005$$

- Compare with the ANOVA table in the book page 130. They are not the same.
- Test Statistics: **SEQUENTIAL TESTS!**
  2nd column/1st column=3rd column; Q_F is in the 4th column: Obtained from 3rd column/bottom value of 3rd column.
- p-value is the 5th column: `1-pf(Q_F, 1st column,1st column bottom).`

GLM Example 2.8

- Parallel tests:

```
> drop1(glmSyn,test="F")
Single term deletions

Model:
Qual ~ Acontent + Bcontent - 1
         Df Sum of Sq   RSS    AIC F value   Pr(>F)
<none>                  150 15.736
Acontent  1     10830 10980 26.616    72.2 0.07458 .
Bcontent  1      1470  1620 20.875     9.8 0.19684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p-value for Bcontent matches the `anova` output.
- Note that if we switch the sequence in the model object:

```
> temp <- lm(Qual ~ Bcontent + Acontent -1, data=synprod)
> anova(temp)
Analysis of Variance Table

Response: Qual
          Df Sum Sq Mean Sq F value  Pr(>F)
Bcontent   1   3645    3645    24.3 0.12742
Acontent   1  10830   10830    72.2 0.07458 .
Residuals  1    150     150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p-value for Acontent matches the `anova` output.

- **The top test in the ANOVA table is given the previous being accepted.**

GLM Example 2.8

Theorem 2.3

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\Sigma^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.5 & 0.5 \end{bmatrix}, X^T X = \begin{bmatrix} 1.25 & 0.25 \\ 0.25 & 1.25 \end{bmatrix},$$

$$(X^T X)^{-1} = \begin{bmatrix} 0.8333 & -0.1667 \\ -0.1667 & 0.8333 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}, \hat{\sigma}^2 = 150$$

$$\widehat{V(\hat{\theta})} = \frac{150}{6}\begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix} = \begin{bmatrix} \mathbf{125} & \mathbf{-25} \\ \mathbf{-25} & \mathbf{125} \end{bmatrix}$$

GLM Example 2.8

$$(X^T X)^{-1} = \begin{bmatrix} \mathbf{0.8333} & \mathbf{-0.1667} \\ \mathbf{-0.1667} & \mathbf{0.8333} \end{bmatrix}, V(\hat{\theta}) = \begin{bmatrix} \mathbf{125} & \mathbf{-25} \\ \mathbf{-25} & \mathbf{125} \end{bmatrix}$$

```
> summary(glmSyn)$cov.unscaled
           Acontent   Bcontent
Acontent  0.8333333 -0.1666667
Bcontent -0.1666667  0.8333333

> summary(glmSyn)$sigma^2*summary(glmSyn)$cov.unscaled
         Acontent Bcontent
Acontent      125      -25
Bcontent      -25      125
```

GLM Example 2.8

Note that

$$
\begin{aligned}
cov\left(\hat{\theta}, Y - P_M Y\right) &= cov\left((X^T X)^{-1} X^T Y, P_{M^\perp} Y\right) \\
&= cov\left((X^T X)^{-1} X^T (P_M Y + P_{M^\perp} Y), P_{M^\perp} Y\right) \\
&= cov\left((X^T X)^{-1} X^T P_M Y, P_{M^\perp} Y\right) \\
&= (X^T X)^{-1} X^T cov(P_M Y, P_{M^\perp} Y) \\
&= (X^T X)^{-1} X^T P_M V(Y) P_{M^\perp} \\
&= \sigma^2 (X^T X)^{-1} X^T < P_M , P_{M^\perp} >= 0,
\end{aligned}
$$

So that $\hat{\theta}$ and $\hat{\sigma}^2 = \frac{1}{n-k} \|Y - P_M Y\|^2$ are **independent** of each other.

Thus, $\hat{\theta}$ and $\hat{\sigma}^2$ are **independent** of each other.

GLM Example 2.8

```
glmSynA <- lm(Qual ~ Acontent +Bcontent - 1, data=synprod)

> summary(glmSynA)

Call:
lm(formula = Qual ~ Acontent + Bcontent - 1, data = synprod)

Residuals:
 1  2  3
-5 -5 10

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Acontent    95.00      11.18   8.497   0.0746 .
Bcontent    35.00      11.18   3.130   0.1968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.25 on 1 degrees of freedom
Multiple R-squared:  0.9897,    Adjusted R-squared:  0.9692
F-statistic: 48.25 on 2 and 1 DF,  p-value: 0.1013
```

Parameter estimates $\hat{\theta}$

Standard error:
$\sqrt{\hat{\sigma}^2 diag\left(\frac{1}{6}\begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}\right)}, \sqrt{125} = 11.18$

T-test statistic for testing $H_A: \theta_A = 0$ and $H_B: \theta_B = 0$: Estimate/Std. Error

p-value from the $t(n - dim(M))$ distribution

GLM Example 2.8

The t-test:

Under $H_B: \theta_B = 0$:

$$\hat{\theta}_B \sim N\left(0, \frac{5}{6}\sigma^2\right), \qquad \widehat{V(\hat{\theta}_B)} = \frac{5}{6}\hat{\sigma}^2 \sim \frac{5}{6}\sigma^2\chi^2(1),$$

So that

$$T = \frac{\hat{\theta}_B}{\sqrt{\widehat{V(\hat{\theta}_B)}}} \sim \frac{N\left(0, \frac{5}{6}\sigma^2\right)}{\sqrt{\frac{5}{6}\sigma^2\chi^2(1)}} = \frac{\sqrt{\frac{5}{6}\sigma^2} \cdot N(0,1)}{\sqrt{\frac{5}{6}\sigma^2} \cdot \sqrt{\chi^2(1)}} = \frac{N(0,1)}{\sqrt{\chi^2(1)}} = t(1)$$

Note that

$$T^2 = \frac{\hat{\theta}_B^2}{\widehat{V(\hat{\theta}_B)}} \sim \frac{N(0,1)^2}{\chi^2(1)} = \frac{\chi^2(1)}{\chi^2(1)} = F(1,1)$$

GLM Example 2.8

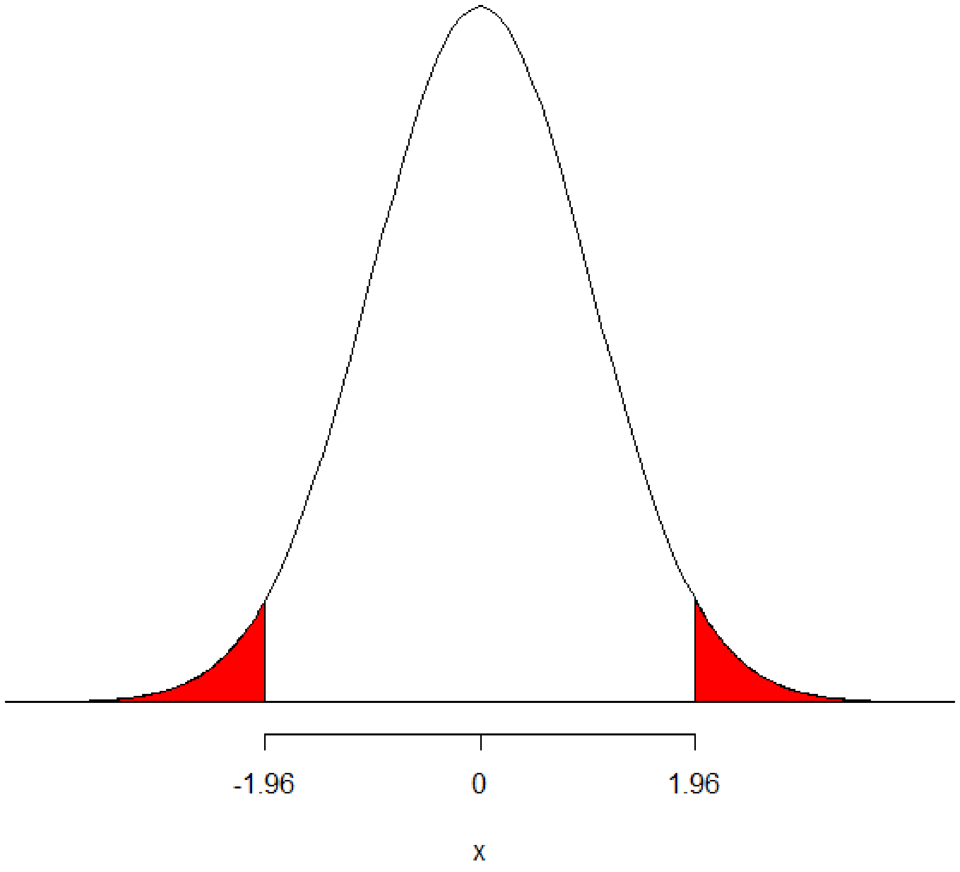
- In general, the t-test statistics can only be used as pointers, as they don’t involve estimation under H_0.
- However, in simple situations, the t-test is equivalent with the Likelihood Ratio test.
- For example, in simple linear models without interaction terms.
- Here, $T = 3.130$ (slide 30), $T^2 = 3.130^2 = 9.8 = Q_F$(slide 26), with Q_F equivalent to the Likelihood Ratio test.

Confidence Intervals

- Suppose that $Y \sim N(0,1)$. Then

$$P(Y \leq 1.96) = P(Y \leq q_{0.975}) = 0.975$$

$$P(-1.96 \leq Y \leq 1.96)$$
$$= P(q_{0.025} \leq Y \leq q_{0.975}) = 0.95$$

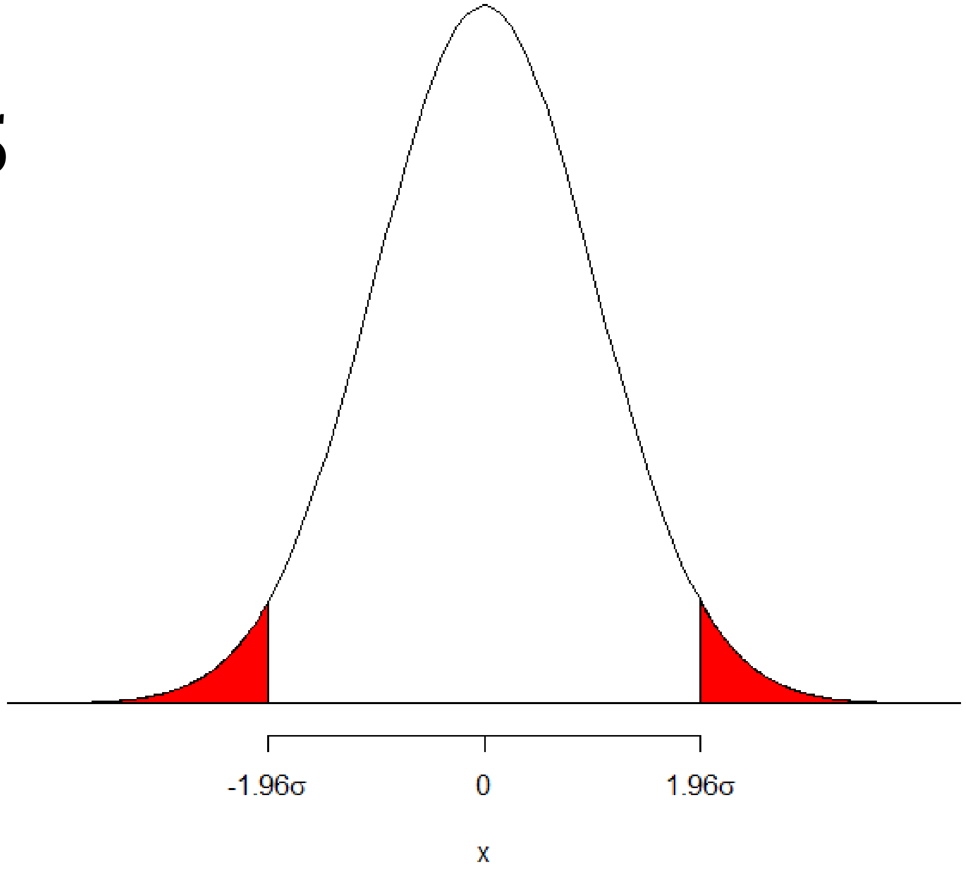


Confidence Intervals

- Suppose that $Y \sim N(0, \sigma^2)$. Then

$$P(Y \leq 1.96\sigma) = P(Y \leq q_{0.975}) = 0.975$$

$$P(-1.96\sigma \leq Y \leq 1.96\sigma)$$
$$= P(q_{0.025} \leq Y \leq q_{0.975}) = 0.95$$



Confidence Intervals

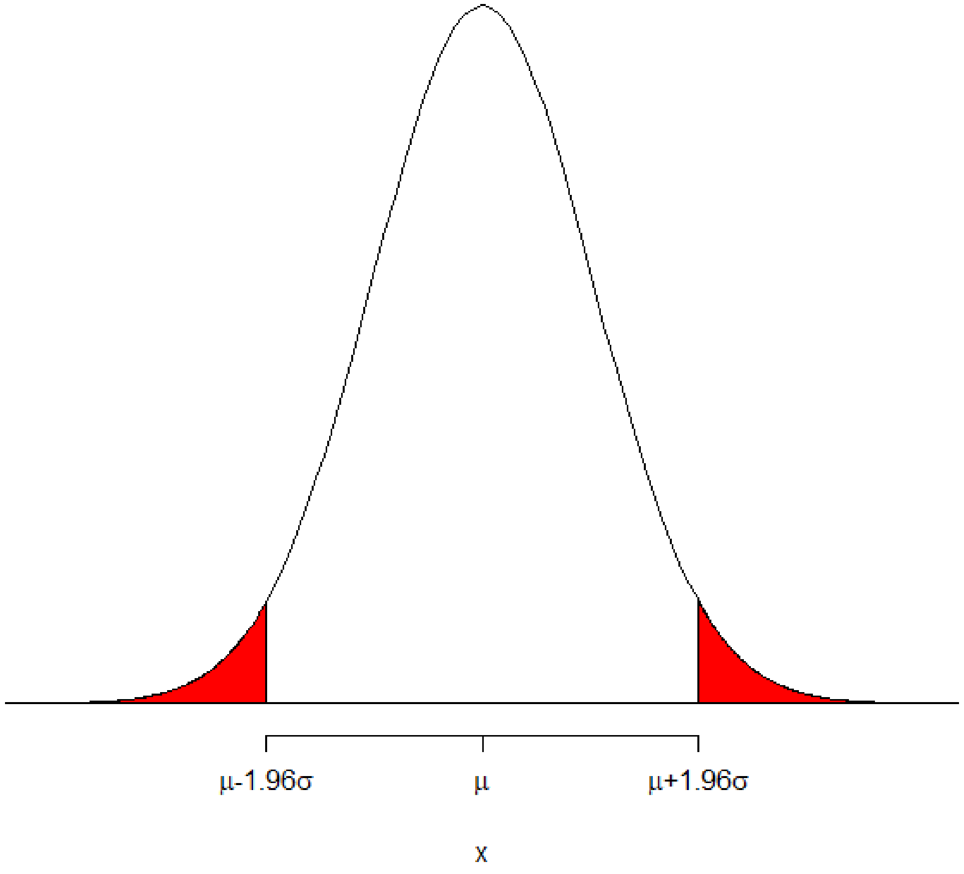
- Suppose that $Y \sim N(\mu, \sigma^2)$. Then

$$P(Y \leq \mu + 1.96\sigma) = P(Y \leq q_{0.975}) = 0.975$$

$$P(\mu - 1.96\sigma \leq Y \leq \mu + 1.96\sigma)$$
$$= P(q_{0.025} \leq Y \leq q_{0.975}) = 0.95$$

Follows from that

$$\frac{Y - \mu}{\sigma} \sim N(0,1)$$



Confidence Intervals

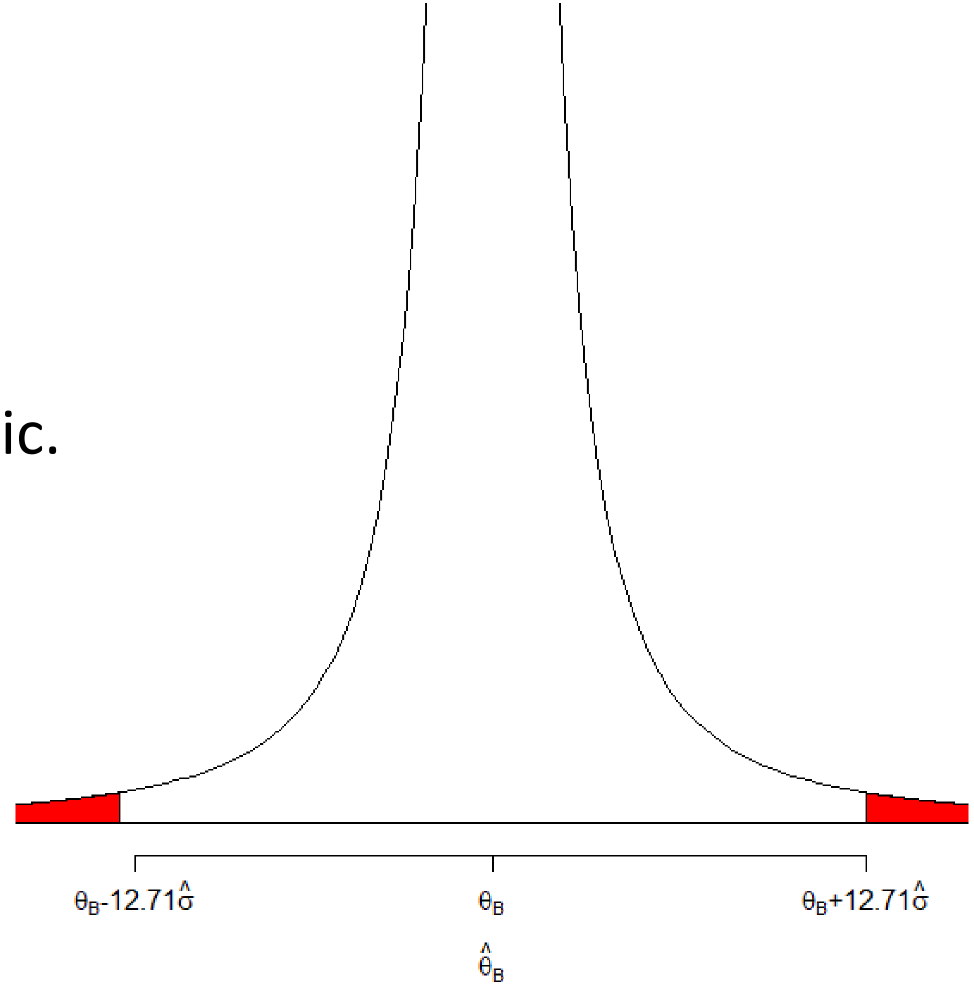
- In practice, we will not know μ and σ.
- Introduces additional variation:

$$\frac{Y-\hat{\mu}}{\hat{\sigma}} \sim t(n-k)$$

Example 2.8, $Y=\hat{\theta}_B$: $\frac{Y-\theta_B}{\hat{\sigma}} \sim t(1)$ from the same logic.

$$P\left(\frac{Y-\theta_B}{\hat{\sigma}} \leq 12.71\right)$$
$$= P\left(\frac{Y-\theta_B}{\hat{\sigma}} \leq q_{0.975}\right) = 0.975$$

Uncertainty increased with a factor 6!



Confidence Intervals

- $\widehat{V(\hat{\theta}_B)} = 11.18, \hat{\theta}_B = 35$ (slide 31).

$$
\begin{aligned}
0.95 &= P\left(-12.71 \le \frac{\hat{\theta}_B - \theta_B}{\hat{\sigma}} \le 12.71\right) \\
&= P\left(\theta_B - 12.71\hat{\sigma} \le \hat{\theta}_B \le \theta_B + 12.71\hat{\sigma}\right) \\
&= P\left(\hat{\theta}_B - 12.71\hat{\sigma} \le \theta_B \le \hat{\theta}_B + 12.71\hat{\sigma}\right)
\end{aligned}
$$

- Thus,

$$\theta_B \in (35 - 12.71 \cdot 11.18; 35 + 12.71 \cdot 11.18) = (-107.10; 177.10)$$

with 95% probability.

- $(-107.10; 177.10)$ is a $stochastic\ interval$ that with 95% probability contains the true parameter θ_B. This is where we have **95% confidence** that θ_B is.

- Very little information from 3 observations.

Confidence Intervals

Theorem 2.23

Let the situation be as above. Then the critical region for testing H_0 against H_1 at significance level α is

$$C_\alpha = \left\{ (y_1, \cdots, y_n) \,\middle|\, \hat{\theta}_{i_0} < c - t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}\left(\hat{\theta}_{i_0}\right)} \text{ or } \hat{\theta}_{i_0} > c + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}\left(\hat{\theta}_{i_0}\right)} \right\}$$

Remark 2.24

The estimated standard deviation $(\hat{V}(\hat{\theta}_{i_0}))^{1/2}$ of $\hat{\theta}_{i_0}$is often provided by standard software as '*standard error of estimate*' or similar. It is thus straight forward to compute the critical limits. This result may also be used in setting up confidence limits for θ_{i_0}. More specifically, a $(1-\alpha)$ confidence interval becomes

$$\left[\hat{\theta}_{i_0} - t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})},\ \ \hat{\theta}_{i_0} + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \right]$$

Warning

- Do not trust remark 2.25 in the book.
- While we have shown that it always is so that $T^2 \sim F(1, n-k)$, it is **NOT** the same thing as T^2 being equivalent to Q_F.

- In more complicated models that includes interactions, Q_F and T^2 may lead to different conclusions (use Q_F).

Prediction Intervals
Example 2.20

Example 2.20

We consider the following corresponding observations of an independent variable x and a dependent variable y:

x	1	2	3	4	5	6
y	0.3	1.5	1.3	1.9	4.2	8

We assume that the y's originate from independent stochastic variables $Y_1, \ldots, Y_6$ which are normally distributed with mean values

$$\mathrm{E}(Y|x) = \beta x^2$$

and variances

$$\mathrm{V}(Y|x) = x^2\sigma^2$$

We would now like to find a confidence interval for a new (or future) observation corresponding to $x = 10$. This observation is called Y, and we have

$$\begin{aligned} \mathrm{E}(Y) &= 100\beta \\ \mathrm{V}(Y) &= 100\sigma^2 . \end{aligned}$$

Example 2.20

- Matrix representation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}, \qquad Y = X\theta + \varepsilon,$$

$$V(\varepsilon) = \sigma^2 \Sigma = \sigma^2 \begin{bmatrix} 1 & & & & & \\ & 4 & & & & \\ & & 9 & & & \\ & & & 16 & & \\ & & & & 25 & \\ & & & & & 36 \end{bmatrix}$$

Example 2.20

- Parameter estimation:

$$\hat{\theta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y = 0.1890,$$
$$\hat{\sigma}^2 = \frac{1}{6-1}\left(Y - X\hat{\theta}\right)^T\Sigma^{-1}\left(Y - X\hat{\theta}\right) = 0.0597$$

- R code:

```
> Y<-c(0.3,1.5,1.3,1.9,4.2,8)
> X<-cbind((1:6)^2)
> Sigma<-diag((1:6)^2)
> (thetahat<-solve(t(X)%*%solve(Sigma)%*%X)%*%t(X)%*%solve(Sigma)%*%Y)
         [,1]
[1,] 0.189011
> (sigma2hat<-(1/(6-1))*t(Y-X%*%thetahat)%*%solve(Sigma)%*%(Y-X%*%thetahat))
           [,1]
[1,] 0.05965831
```

Example 2.20

- Parameter estimator distribution:

$$E(\hat{\theta}) = \theta, \qquad V(\hat{\theta}) = \sigma^2(X^T\Sigma^{-1}X)^{-1} = \sigma^2\left(\sum_{i=1}^{6} X_i^2/i^2\right)^{-1} = \sigma^2\left(\sum_{i=1}^{6} i^4/i^2\right)^{-1} = \frac{\sigma^2}{91};$$

$$\hat{\theta} \sim N\left(\theta, \frac{\sigma^2}{91}\right).$$

Example 2.20

- Now let $Y_{10} \sim N(10^2\theta, 10^2\sigma^2), = N(100\theta, 100\,\sigma^2)$,
  Y_{10} independent of $Y_1, \dots, Y_6$.

We expect Y_{10} to be around 100θ;
Confidence interval for $E(Y_{10})$:

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\sigma}^2}{91}}} \sim t(5).$$

Since $q_{0.975} = \min_q P(t(5) \leq q) \geq 0.975 =$ `qt(0.975,df=5)` $= 2.57$,it holds that

$$\theta = \hat{\theta} \pm 2.57\sqrt{\frac{\hat{\sigma}^2}{91}} = 0.1890 \pm 0.066, \qquad ie.\ \ \theta \in (0.1232; 0.2548)$$

With probability 0.95, so that

$$\boldsymbol{E(Y_{10}) = 100\theta \in (12.32; 25.48)}$$

With probability 0.95.

Example 2.20

- $100\hat{\theta} \sim N\left(100\theta, \frac{100^2}{91}\sigma^2\right) = N(100\theta, c\sigma^2)$. Take $u = 100\hat{\theta}$.

Theorem 2.15

Let the situation be as above. Then the $(1-\alpha)$-*confidence interval* for the expected value of a new observation Y will be

$$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}, \quad u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}].$$

Example 2.20

- $Y_{10} \sim N(10^2\theta, 10^2\sigma^2), = N(100\theta, 100\,\sigma^2)$,
  Y_{10} independent of $Y_1, \dots, Y_6$.

  How about Y_{10} itself?

- since $\hat{\theta}$ is anunbiased estimator for θ, we expect $Y_{10} - \hat{\theta}$ to be around 0:

$$E\left(Y_{10} - 100\hat{\theta}\right) = 0,$$

$$V\left(Y_{10} - 100\hat{\theta}\right) = V(Y_{10}) + 100^2 V\left(\hat{\theta}\right) = 100\sigma^2 + \frac{100^2}{91}\sigma^2 = \frac{19100}{91}\sigma^2$$

Example 2.20

$$Y_{10} - 100\hat{\theta} \sim N\left(0, \frac{19100}{91}\sigma^2\right):$$

$$\frac{Y_{10} - 100\hat{\theta}}{\sqrt{\frac{19100}{91}\hat{\sigma}^2}} = \frac{\frac{1}{\sigma}\sqrt{\frac{91}{19100}}\left(Y_{10} - 100\hat{\theta}\right)}{\frac{1}{\sigma}\hat{\sigma}} \sim \frac{N(0,1)}{\sqrt{\chi^2(5)}} = t(5)$$

Thus,

$$Y_{10} = 100\hat{\theta} \pm 2.57\sqrt{\frac{19100}{91}\hat{\sigma}^2} = 18.90 \pm 9.10,$$

ie.

$$Y_{10} \in (18.90 - 9.10; 18.90 + 9.10) = (\mathbf{9.80}; \mathbf{28.00})$$

With probability 0.95.

Example 2.20

Theorem 2.17

Let us assume that a new observation taken at $(z_1, \ldots, z_k)$ has a variance $c_1\sigma^2$. Furthermore, it is independent of the earlier observations. In that case a $(1-\alpha)$-*prediction interval* for the new observation equals the interval

$$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}, u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}].$$

Example 2.20

95% Confidence interval for $E(Y_{10})$:

$$\mathbf{(12.32; 25.48)}$$

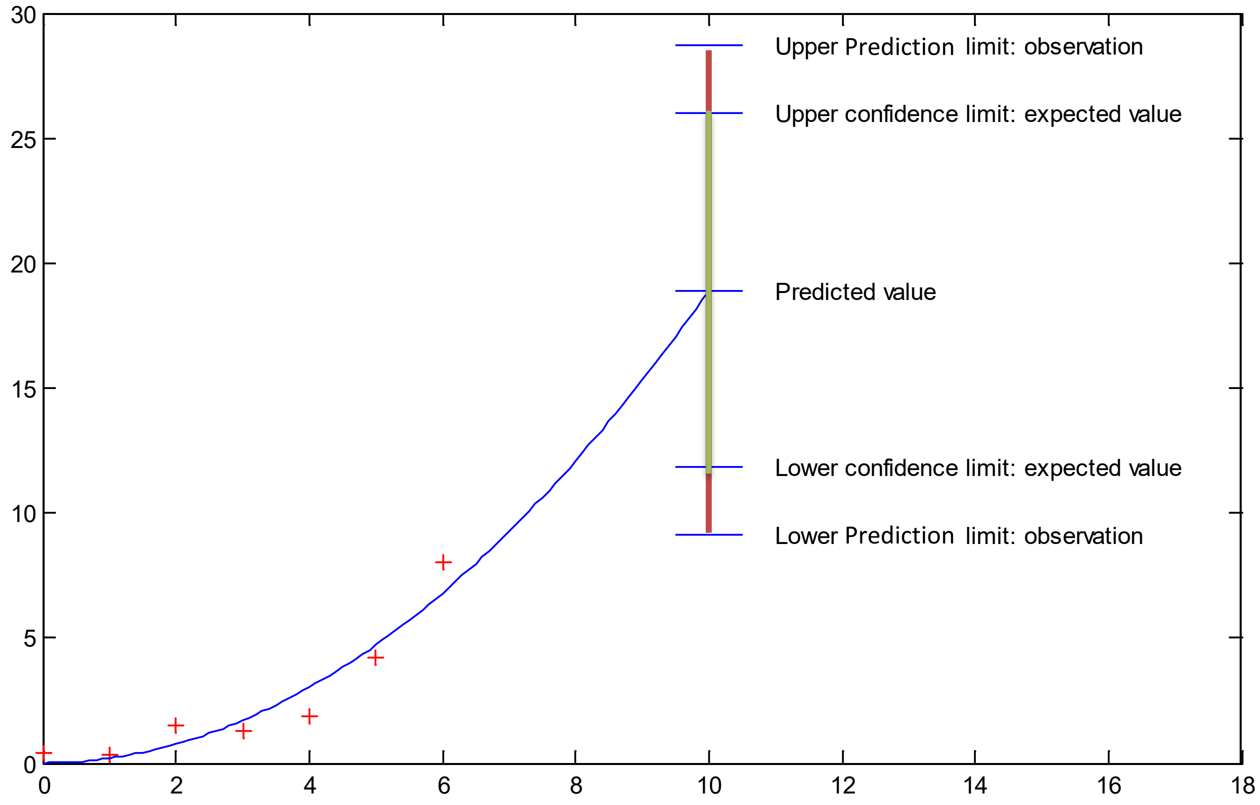
95% Prediction interval for Y_{10}:

$$\mathbf{(9.80; 28.00)}$$

Prediction interval is **wider**.

Confidence intervals are for parameters;
Prediction intervals are for observations

Example 2.20



Upper Prediction limit: observation
Upper confidence limit: expected value
Predicted value
Lower confidence limit: expected value
Lower Prediction limit: observation
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18

Exercises

- 3.3 - GLM

- 3.4 - GLM

- 3.6 – GLM. Compare what the `Anova()` function from the `car` package does, to the `drop1()` function. Data can alternatively be accessed from the dataset `Cars93` from the `MASS` package: `library(MASS); data(Cars93)`.