## Dagens program:

Mandag den 14. februar

**Den generelle lineære model**

- Kort om likelihoodfunktion
- Flerdimensional normalfordeling
- Generel lineær model
- Estimation, fordeling af residualer mv
- Test for modelreduktion
- Successiv testning

Eksempel, hjerneblodgennemstrømning hos malere



## Likelihoodfunktion :

Kommer fra udtrykket for sandsynlighedstætheden

Sandsynlighedstætheden beskriver en **hyppighedsfordeling** af observationer - for en forelagt værdi af parameteren

likelihoodfunktionen er en funktion af parameteren - for forelagte værdier af observationerne

likelihoodfunktionen beskriver **parameterusikkerhed**, nemlig grad af overensstemmelse mellem data og parameter.



## Likelihoodfunktion, begreber :

- **score-funktion**: første afledede af log-likelihood
  har middelværdien nul
- **observeret information**: anden afledede af log-likelihood (krumning)
- **forventet information**: middelværdien af den observerede information
  Og det er netop dispersionsmatricen for scorefunktionen



## Likelihoodfunktion, transformation af parametre :

Lad $\theta = \theta(\beta) \in \mathbb{R}^k$ for $\boldsymbol{\beta} \in \mathsf{B} \subset \mathbb{R}^m$ med $m \leq k$

Da er scorefunktionen (mht $\boldsymbol{\beta}$)

$$l'_\beta(\beta;\ \mathbf{y}) = \mathbf{J}^T\ l'_\theta(\theta(\beta);\ \mathbf{y}) \tag{2.11}$$

hvor

$$\mathbf{J} = \frac{\partial \theta}{\partial \beta}$$

med elementer

$$\mathbf{J}_{ij} = \frac{\partial \theta_i}{\partial \beta_j} \quad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m$$

og den forventede information er

$$\mathbf{i}_\beta(\beta) = -\mathsf{E}\left[\frac{\partial^2}{\partial \beta \partial \beta^\mathsf{T}}\ \ell(\theta(\beta);\ \mathbf{Y})\right] = \mathbf{J}^\mathsf{T}\ \mathbf{i}_\theta(\theta(\beta))\ \mathbf{J} \tag{2.12}$$

## Normalfordeling :

$$f(\mathbf{y};\boldsymbol{\mu},\sigma^2) = \frac{1}{(\sqrt{2\pi})^k\sigma^k\mathsf{det}(\boldsymbol{\Sigma})}\ \exp\left[\frac{-1}{2\sigma^2}\,(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \quad \text{for } \mathbf{y}\in\mathbb{R}^k \tag{3.4}$$

$$\delta_\Sigma(\mathbf{y}_1,\mathbf{y}_2) = \mathbf{y}_1'\boldsymbol{\Sigma}^{-1}\mathbf{y}_2 \tag{3.5}$$

et **indre produkt** på $\mathbb{R}^k$.
Tilsvarende **ortogonalitetsbegreb** (svarende til at det indre produkt er nul), og en **norm** (defineret ved $\|\mathbf{y}\|_\Sigma = \sqrt{\delta(\mathbf{y},\mathbf{y})}$.

Med notationen

$$D(\mathbf{y};\boldsymbol{\mu}) = \|\mathbf{y}-\boldsymbol{\mu}\|^2_\Sigma = (\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.6}$$

bliver tæthedsfunktionen

$$f(\mathbf{y};\boldsymbol{\mu},\sigma^2) = \frac{1}{(\sqrt{2\pi})^k\sigma^k\mathsf{det}(\boldsymbol{\Sigma})}\ \exp\left[\frac{-1}{2\sigma^2}\,D(\mathbf{y};\boldsymbol{\mu})\right] \tag{3.7}$$



## Normalfordeling, scorefunktion og information :

**log-likelihood**

$$\ell_\mu(\boldsymbol{\mu},\sigma^2;\mathbf{y}) = -(k/2)\log(\sigma^2) - \frac{1}{2\sigma^2}\,(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.8}$$

**Scorefunktion** (m.h.t. $\boldsymbol{\mu}$)

$$\frac{\partial}{\partial\boldsymbol{\mu}}\,\ell_\mu(\boldsymbol{\mu},\sigma^2;\mathbf{y}) = \frac{1}{\sigma^2}\left[\boldsymbol{\Sigma}^{-1}\mathbf{y}-\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.9}$$

**Observeret information** (m.h.t. $\boldsymbol{\mu}$)

$$\mathbf{j}(\boldsymbol{\mu};\mathbf{y}) = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}\,. \tag{3.10}$$

afhænger ikke af observationerne $y$, hvorfor
den **forventede information** er

$$\mathbf{i}(\boldsymbol{\mu}) = \frac{1}{\sigma^2}\boldsymbol{\Sigma}^{-1}\,.$$



## Generel lineær model:

**Model** (koordinatfri):

$$H_0\ :\ \boldsymbol{\mu}\in L\,,$$

et lineært (affint) $m$-dimensionalt underrum af $\mathbb{R}^k$
(egentlig $\boldsymbol{\mu}-\boldsymbol{\mu}_0\in L$

**Parametriseret** (i koordinater):

$$H_0\ :\ \boldsymbol{\mu}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$$

Søjlerne i $\mathbf{X}$ udspænder underrummet $L$ (et lokalt koordinatsystem).
$\boldsymbol{\beta}$ angiver punkternes lokale koordinater.

Utallige ækvivalente parametriske fremstillinger af en given model.



## Eksempel, sædvanlig endimensional regression:

Ramushøjder, afsnit 3.5

| Alder (år) $x_i$ | Højde (mm) $y_i$ |
|---|---|
| 8.0 | 51.2 |
| 8.5 | 53.0 |
| 9.0 | 54.3 |
| 9.5 | 54.5 |

## Eksempel, sædvanlig endimensional regression:

Model:

$$\mathsf{E[Y_]} = \beta_0 + \beta_1 \mathsf{x_i}$$

$$\begin{aligned}
\mu_1 &= \beta_0 + \beta_1 x_1 \\
\mu_2 &= \beta_0 + \beta_1 x_2 \\
\mu_3 &= \beta_0 + \beta_1 x_3 \\
\mu_4 &= \beta_0 + \beta_1 x_4
\end{aligned}$$

dvs

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$



## Eksempel, sædvanlig endimensional regression:

med

$$\mathbf{X} = \begin{pmatrix} 1 & 8.0 \\ 1 & 8.5 \\ 1 & 9.0 \\ 1 & 9.5 \end{pmatrix}$$



## Generel lineær model, estimation under model:

**Koordinatfrit**:
maksimum-likelihood estimatet $\widehat{\boldsymbol{\mu}}$ for sættet af middelværdier fås ved at minimere $D(\mathbf{y};\boldsymbol{\mu})$, dvs som den ortogonale (m.h.t $\delta_\Sigma$) projektion, $p_L(\mathbf{y})$ af punktet $\mathbf{y}$ ned på det lineære underrum, $L$.

**Parametrisk**:
Vi bruger (2.11) og (3.9) og får scorefunktionen

$$\frac{\partial}{\partial\boldsymbol{\beta}}\,\ell_\beta(\boldsymbol{\beta},\sigma^2;\mathbf{y}) = \left[\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right]' \frac{\partial}{\partial\boldsymbol{\mu}}\ell_\mu(\boldsymbol{\mu}(\boldsymbol{\beta}),\sigma^2;\mathbf{y}) = \frac{1}{\sigma^2}\mathbf{X}'\left[\boldsymbol{\Sigma}^{-1}\mathbf{y} - \boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta}\right] \tag{3.11}$$

hvorfor maksimum-likelihood estimatet for $\boldsymbol{\beta}$ fås som løsning til **normalligningen**

$$\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}} \tag{3.12}$$

Altså $m$ ligninger med $m$ ubekendte.



## Generel lineær model, estimation under model:

Hvis $\mathbf{X}$ har fuld rang ($m$), er der en entydig løsning, som bestemmes ved

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{3.16}$$

**Hatmatricen**

$$\mathbf{H} = \mathbf{X}[\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}]^{-1}\mathbf{X}'$$

er projektionsmatricen, der overfører $\mathbf{y}$ til de fittede værdier $\widehat{\boldsymbol{\mu}}$ (nemlig sætter hat på $\mu$'erne).

idet nemlig

$$\widehat{\boldsymbol{\mu}} = p_L(\mathbf{y}) = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

**Residualerne** er

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{3.17}$$

## Generel lineær model, residualer og fittede værdier:

Opspaltning:

$$\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{r} = \mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Normalligningen

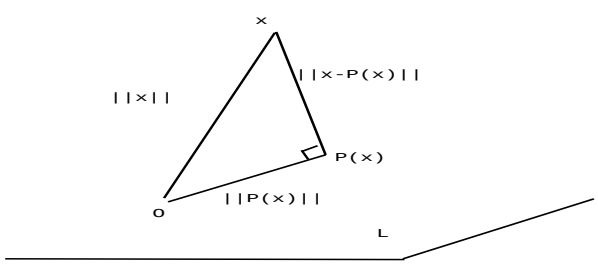
$$\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0} \tag{3.19}$$

viser, at residualerne $\mathbf{r} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ er **ortogonale** (mht $\boldsymbol{\Sigma}^{-1}$) på underrummet, $L$, udspændt af $\mathbf{X}$



## Generel lineær model, opspaltning :

$$\begin{aligned} D(\mathbf{y};\mathbf{0}) &= D(\mathbf{y};\mathbf{X}\widehat{\boldsymbol{\beta}}) + D(\mathbf{X}\widehat{\boldsymbol{\beta}};\mathbf{0}) \\ &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{3.20}$$






## Generel lineær model, fordeling af estimatorer::

$$\begin{aligned} \mathsf{E}[\widehat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} && (3.21) \\ \mathbf{D}\,[\widehat{\boldsymbol{\beta}}] &= \sigma^2\,(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} && (3.22) \end{aligned}$$

Bruges ved test af enkelte parameterværdier

$$\mathbf{D}\,[\mathbf{X}\widehat{\boldsymbol{\beta}}] = \sigma^2\mathbf{X}[\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}]^{-1}\mathbf{X}' = \sigma^2\mathbf{H}$$

$$\mathbf{D}\,[\mathbf{r}] = \sigma^2(\mathbf{I} - \mathbf{H})$$

rang (frihedsgrader) $k - m$.

Bruges til at **standardisere** residualer



## Generel lineær model, estimation af $\sigma^2$:

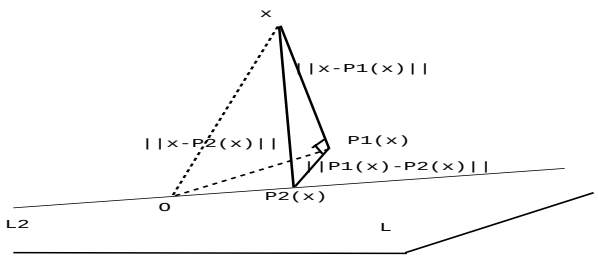
$\sigma^2$ estimeres ved

$$s^2 = \mathbf{r}'\boldsymbol{\Sigma}^{-1}\mathbf{r}/(k - m) \tag{3.24}$$

som er uafhængig af $\widehat{\boldsymbol{\mu}}$

## Generel lineær model, test af delhypotese:

Delhypotese: $\boldsymbol{\mu} \in L_2 \subset L$






## Generel lineær model, kvotienttest af delhypotese::

$$
\begin{aligned}
q_{2|1}(\mathbf{y}) &= \left[\frac{D(\mathbf{y};p_2(\mathbf{y}))}{D(\mathbf{y};p_1(\mathbf{y}))+D(p_1(\mathbf{y});p_2(\mathbf{y}))}\right]^{k/2} \\
&= \left[\frac{1}{1+\dfrac{D(p_1(\mathbf{y});p_2(\mathbf{y}))}{D(\mathbf{y};p_1(\mathbf{y}))}}\right]^{k/2} = \left[\frac{1}{1+\dfrac{m_1-m_2}{k-m_1}F}\right]^{k/2}
\end{aligned}
$$

Kvotienttestet fører således til F-teststørrelsen

$$F = \frac{D(p_1(\mathbf{y});p_2(\mathbf{y}))/(m_1-m_2)}{D(\mathbf{y};p_1(\mathbf{y}))/(k-m_1)}$$

der følger en F$(m_1-m_2, k-m_1)$-fordeling (eksakt resultat)



## Partielt test, type III opspaltning:

Antag at H$_i$ gælder.

Vi vil nu undersøge H$_{i+1} \subset$ H$_i$

Kvotienttestet for H$_{i+1}$ under antagelse af at H$_i$ kan opretholdes har teststørrelsen

$$F(\mathbf{y}) = \frac{D(\widehat{\boldsymbol{\mu}};\widehat{\widehat{\boldsymbol{\mu}}})/(m_1-m_2)}{s_1^2}\,, \tag{3.33}$$

hvor $\widehat{\boldsymbol{\mu}}$ angiver de fittede værdier under H$_i$ og $\widehat{\widehat{\boldsymbol{\mu}}}$ angiver de fittede værdier under H$_{i+1}$ og $s_1^2$ angiver residualvariansen under H$_i$.

Testet kaldes det partielle kvotienttest - eller et Type III test.

Testet "korrigerer" for alle de effekter, der er i modellen på det pågældende trin.

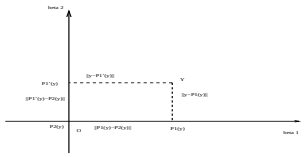


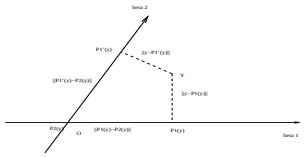
## Successiv testning i hypotesekæde:

$$\mathbb{R} \subseteq L_M \ldots \subset L_2 \subset L_1 \subset L_0 \subset \mathbb{R}^k\,, \tag{3.32}$$

Type I opspaltning udtrykker successice projektioner ned gennem kæden af underrum.

Størrelsen af leddene i Type I opspaltningen kan afhænge af rækkefølgen i modelformlen



## Successiv testning i hypotesekæde:

Type III opspaltning svarer til at man fjerner et led - og vurderer reduktionen i model SSQ ved at fjerne det

så smider man det ind igen, og fjerner et af de andre.
Type III opspaltningen afhænger ikke af rækkefølgen i modelformlen.

I praksis tager man udgangspunkt i en type III opspaltning og fjerner mindst betydende - og mest komplicerede led.
derefter reestimeres - og reduceres



## Diagnostikker:

Oversigt over diagnostiske størrelser samt parametriseringer i notens afsnit 3.9.1 til 3.9.4



## Indskud om multikollinearitet:

Det kan ske, at to eller flere af de forklarende variable følges ad (er næsten lineært afhængige). Konsekvenser af dette er, at modellen er signifikant, men estimaterne for de enkelte koefficienter har stor varians ($X'X$ er næsten singulær). Desuden er fortolkningen af de enkelte koefficienter højst tvivlsom.

Diagnostikker for multikollinearitet udskrives sammen med estimaterne:

$$TOL_i = 1 - R_i^2$$

hvor $R_i^2$ er den multiple determinationskoefficient fra regressionen af den $i$'te forklarende variable på de øvrige. Lille TOL er tegn på multikollinearitet.
V[$\widehat{\beta}_i \propto 1/\text{TOL}_i$

$$VIF_i = 1/TOL_i$$

Hvis VIF$_i > 10$ grund til bekymring



## Eksempel, Blodgennemstrømningsindeks hos malere:

Eksempel på

- Brug af forklarende regressionsvariabel til korrektion for forskelle i populationer (Kovariansanalyse)
- Brug af modelformler
- Forskel på typel og type III opspaltning

## Eksempel, Blodgennemstrømningsindeks hos malere:

Først tegnes boxplot for ISI for de to grupper
Ikke stærke tegn på forskel

Så proves t-test for to uafhængige stikprøver
ved brug af FIT-menuen

Tegn ISI mod alder, marker de to grupper hver for sig
Indikation af sammenhæng mellem ISI og alder

Korriger for alderseffekten ved stor model:
ISI = ORG ALDER ORG $*$ ALDER
To rette linier
Test for reduktion af model til to parallelle linier:
ISI = ORG ALDER
Nu er effekten af brugen af opløsningsmidler signifikant.