Forelesning 3 STK3100

8. september 2008

S. O. Samuelsen

Plan for forelesning:

1. Generelt om lineære modeller
2. Variansanalyse - Kategoriske kovariater
3. Koding av kategoriske kovariater
4. Hat-matrise
5. Residualer

Lineær modell:

Uavhengige responser, $i = 1, \ldots, n$,

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

der

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \beta' \mathbf{x}_i$$

der x_{ij}-ene er forklaringsvariable eller kovariater for respons nr. i og og β_j-ene regresjonsparametre.

Alternativt kan vi skrive dette som

- Y_i-ene er uavhengige
- med forventning $\mu_i = \beta' \mathbf{x}_i$
- konstant varians $\mathrm{Var}(Y_i) = \sigma^2$
- Y_i-ene er normalfordelt

Eks. Fødselsvekt mot svangerskapslengde og kjønn

Fødselsvekt Y_i, indikator for gutt x_{i1}, indikator for jente x_{i2}, svangerskapslengde x_{i3}.

Generell (apriori) modell tillater ulike vekstrater for gutter og jenter:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} * x_{i1} + \beta_4 x_{i3} * x_{i2} + \varepsilon_i$$

Vil teste nullhypotesen $\mathrm{H}_0 : \beta_3 = \beta_4$, dvs. samme veksthastighet, som er gitt ved modell

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Notasjon - Data

Respons for "individ" nr. i: $Y_i, i = 1, \ldots n$

$$\text{Vektor av responser } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

x_{ij} = Forklaringsvariabel nr $j, j = 1, \ldots, p$ for individ nr. i
Kovariatmatrise eller **Designmatrise** for forklaringsvariablene:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Designmatrise for apriorimodell. fødselvekter:

$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$ der

x_{i1} = indikatorvariabel for gutt,

x_{i2} = indikatorvariabel for jente,

$x_{i4} = x_{i1}x_{i3}$ = produkt av varighet og indikator gutt og

$x_{i5} = x_{i2}x_{i3}$ = produkt av varighet og indikator jente

Designmatrisen blir da, når det er nummerert slik at de første 12 individene er gutter,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & x_{1,3} & 0 \\ \vdots & \vdots & & \\ 1 & 0 & x_{12,3} & 0 \\ 0 & 1 & 0 & x_{13,3} \\ \vdots & \vdots & & \\ 0 & 1 & 0 & x_{24,3} \end{bmatrix}$$

Designmatrise for nullhypotesemodell

$Y_i =$ fødselvekt individ nr. i

$x_{i1} =$ indikatorvariabel for gutt

$x_{i2} =$ indikatorvariabel for jente

$x_{i3} =$ svangerskapslengde

Nullhypotesemodellen blir da $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

Nå blir designmatrisen

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & x_{1,3} \\ \vdots & \vdots & \\ 1 & 0 & x_{12,3} \\ 0 & 1 & x_{13,3} \\ \vdots & \vdots & \\ 0 & 1 & x_{24,3} \end{bmatrix}$$

Kvadratsum / Likelihood

På matriseform kan vi skrive $\mu = (\mu_1, \ldots, \mu_n)' = \mathbf{X}\beta$. Dermed kan kvadratsummen kan skrives

$$S(\beta) = \sum_{i=1}^{n} (Y_i - \mu_i)^2 = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

Likelihood for $Y_1, \ldots, Y_n$ blir da

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \left[\tfrac{1}{\sqrt{2\pi\sigma^2}} \exp(-\tfrac{1}{2}(Y_i - \mu_i)^2) \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp(-\tfrac{1}{2} S(\beta)) \end{aligned}$$

Estimering: MK = ML

Med $\mathbf{Y}$ vektor av responser og $\mathbf{X}$ designmatrisen blir log-likelihood

$$l(\beta) = -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) + K$$

(der K ikke inneholder β) og score-ligninger

$$\frac{1}{\sigma^2}(-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\hat{\beta})) = 0,$$

så MLE = MKE gis ved

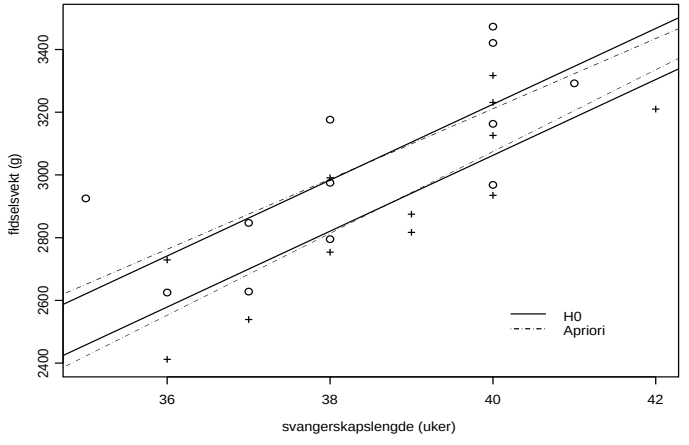
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

såsant $\mathbf{X}'\mathbf{X}$ er inverterbar.

R-tilpasning fødselsvekt, begge modeller

```
> lm(vekt~factor(kjonn)+uker-1,data=fvekt)

Call:
lm(formula = vekt ~ factor(kjonn) + uker - 1, data = fvekt)

Coefficients:
factor(kjonn)1  factor(kjonn)2            uker
       -1610.3         -1773.3           120.9

> x4<-uker*(kjonn==1)
> x5<-uker*(kjonn==2)
> lm(vekt~factor(kjonn)+x4+x5-1,data=fvekt)

Call:
lm(formula = vekt ~ factor(kjonn) + x4 + x5 - 1, data = fvekt)

Coefficients:
factor(kjonn)1  factor(kjonn)2              x4              x5
       -1268.7         -2141.7           112.0           130.4
```


Eks: Fødselsvekt



Felles (H_0) stigningskoeffisient: 120.894
Apriori 111.983 for gutter og 130.400 for jenter.

Fordelingsegenskaper

Siden $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\beta$ fås

$$\mathrm{E}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta,$$

dvs. forventningsrett.
Dessuten er kovarians-matrisen til $\mathbf{Y}$ gitt som $\sigma^2 I$ der I er en nxn identitetsmatrise, dermed blir kovariansmatrisen til $\hat{\beta}$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

og vi har altså

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

eksakt siden $\hat{\beta}$ er en lineærkombinasjon av normale Y_i.

Variansen σ^2

estimeres forventningsrett ved

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}'\mathbf{x}_i)^2}{n-p} = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}'\hat{\beta})'(\mathbf{Y} - \mathbf{X}'\hat{\beta})$$

Dessuten har vi at

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

igjen et eksakt resultat når Y_i-ene er normalfordelt.

Likelihood Ratio Test (LRT)

Anta at Mod_0 er et spesialtilfelle av Mod_1 og at vi vil teste nullhypotesen

$$\text{H}_0 : \text{Mod}_0 \text{ er sann}$$

under antagelse på forhånd (apriori) av at Mod_1. LRT består da i å forkaste hvis $\Delta = 2(\hat{l} - l^*)$ er stor der $\hat{l}$ og l^* er maksimal loglikelihood under hhv. Mod_1 og Mod_0. Vi har da tilnærmet under H_0 at

$$\Delta = 2(\hat{l} - l^*) \sim \chi^2_q$$

når q antall færre parametre i Mod_0.

LRT og lineærnormale modeller

Anta at Mod_0 gis ved $\text{H}_0 : \beta_{p-q+1} = \cdots = \beta_p = 0$ samt at σ^2 er kjent. La dessuten $\hat{\beta}$ og β^* være MLE/MKE under hhv. apriorispesifikasjon og H_0 slik at

$$\beta^*_{p-q+1} = \cdots = \beta^*_p = 0$$

Da blir

$$\Delta = 2(\hat{l} - l^*) = \frac{S(\beta^*) - S(\hat{\beta})}{\sigma^2}.$$

For lineærnormale modeller holder dessuten $\Delta \sim \chi^2_q$ *eksakt*.

F-test og LRT

Dessuten er $\Delta = \frac{S(\beta^*) - S(\hat{\beta})}{\sigma^2}$ og $\hat{\sigma}^2$ uavhengige, dermed får vi at eksakt

$$F = \frac{[S(\beta^*) - S(\hat{\beta})]/q}{\hat{\sigma}^2} \sim F_{q,n-p}$$

dvs. Fisher-fordelt med q og $n - p$ frihetsgrader under $\text{H}_0 : \beta_{p-q+1} = \cdots = \beta_p = 0$ når Y_i-ene er normale.

Dette resultatet kan brukes til å teste f.eks. ingen effekt av en kovariat x som inngår i modellen både med leddet x og x^2.

Men mer typisk brukes det til å teste om det er en effekt av en kategorisk kovariat.

En noe mer generell nullhypotese, de J & H

La C være en $q \times p$ matrise. Vi kan mer generelt være interessert i å teste

$$\text{H}_0 : \mathbf{C}\beta = 0$$

Spesielt får vi testen presentert foran ved å sette

$$\mathbf{C} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Vi antar at rangen til $\mathbf{C}$ er lik $q(< p)$.

En noe mer generell F-test, de J & H

La som før $\mathbf{X}$ være designmatrisen og $\hat{\beta}$ paramterestimat for apriorimodellen, $\mathbf{X}^\star$ og $\beta^\star$ er designmatrise og parameterestimater under nullhypotesen. Vi kan skrive kvadratsummen

$$\begin{aligned} S(\hat{\beta}) &= (\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta}) = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \end{aligned}$$

siden $\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} = \hat{\beta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\beta}'\mathbf{X}'\mathbf{Y}$.
Tilsvarende fås for nullhypotesemodellen
$S(\beta^\star) = \mathbf{Y}'\mathbf{Y} - \beta^{\star\prime}\mathbf{X}^{\star\prime}\mathbf{Y}$ der $\mathbf{X}^\star$ er designmatrisen generert av $\mathbf{C}\beta = 0$. Dermed blir F-observatoren

$$F = \frac{(S(\beta^\star) - S(\hat{\beta}))/q}{\hat{\sigma}^2} = \frac{(\hat{\beta}'\mathbf{X}'\mathbf{Y} - \beta^{\star\prime}\mathbf{X}^{\star\prime}\mathbf{Y})/q}{\hat{\sigma}^2} \sim F_{q,n-p}$$

F-test for ulik veksthastighet mellom kjønn

```
> mod0<-lm(vekt~factor(kjonn)+uker-1,data=fvekt)
> mod1<-lm(vekt~factor(kjonn)+x4+x5-1,data=fvekt)
> anova(mod0,mod1)
Analysis of Variance Table

Model 1: vekt ~ factor(kjonn) + uker - 1
Model 2: vekt ~ factor(kjonn) + x4 + x5 - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     21 658771
2     20 652425  1      6346 0.1945 0.6639
```

Siden $n = 24, p = 4$ og $q = 1$ blir

$$F = \frac{(658770.8 - 652424.5)/1}{652424.5/20} = 0.19$$

Ikke-signifikante relativt til F-fordeling, p=0.66.

Noen ganger inndeles lineære modeller i

1. Multippel lineær regresjon
   - Kun "skala"-kovariater
2. Variansanalyse - ANOVA
   - Kun kategoriske kovariater
3. Kovariansanalyse
   - Både kategoriske kovariater og skala-kovariater

Denne inndelingen er imidlertid ikke så vanlig innen GLM-rammen, vi har typisk både skala- og kategoriske kovariater, og snakker uansett om multippel regresjon.

Eks: Kategoriske ¨kovariater

Y = Inntekt etter kjønn og sosioøkonomisk gruppe:

	Sted 1	Sted 2	Sted 3
Mann	300 350 370 360	400 370 420 390	400 430 420 410
Kvinne	300 320 310 305	350 370 340 355	370 380 360 365

Altså to kategoriske kovariater = faktorer i R-terminologi:
1. Kjønn med 2 nivåer
2. Sted med 3 nivåer

Parametrisering med enveis variansanalyse

Sammenligning av forventning mellom J grupper:

Modell: Anta at individ i er i gruppe j. Da er $Y_i \sim \mathrm{N}(\mu_j, \sigma^2)$ La for $j = 1, \ldots, J$

$$x_{ij} = \begin{cases} 1 & \text{hvis } i \text{ er i gruppe } j \\ 0 & \text{ellers} \end{cases}$$

Da kan vi skrive dette som en lineær modell uten konstantledd

$$\mu_j = \mathrm{E}[Y_i] = \mu_1 x_{i1} + \mu_2 x_{i2} + \cdots + \mu_J x_{iJ}$$

Denne parametriseringen er imidlertid kan imidlertid ikke benyttes med flere kategoriske kovariater
(for flere kovariater samtidig).

Eks: Inntekt over sted

Ser bort fra kjønn. Kun en faktor og altså enveis ANOVA.

```
> inntekt<-c(300,350,370,360,400,370,420,390,400,430,420,410,300,320,310
> kjonn<-c(rep(1,12),rep(2,12))
> sted<-rep(c(1,1,1,1,2,2,2,2,3,3,3,3),2)

> lm(inntekt~factor(sted)-1)

Coefficients:
factor(sted)1  factor(sted)2  factor(sted)3
        326.9          374.4          391.9

> summary(lm(inntekt~factor(sted)-1))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
factor(sted)1  326.875      9.733   33.58   <2e-16 ***
factor(sted)2  374.375      9.733   38.46   <2e-16 ***
factor(sted)3  391.875      9.733   40.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Eks: Designmatrise for Sted uten konstantledd

```
> enveisfit<-lm(inntekt~factor(sted)-1,x=T)
> enveisfit$x
   factor(sted)1 factor(sted)2 factor(sted)3
1              1             0             0
2              1             0             0
3              1             0             0
4              1             0             0
5              0             1             0
6              0             1             0
7              0             1             0
8              0             1             0
9              0             0             1
10             0             0             1
11             0             0             1
12             0             0             1
13             1             0             0
14             1             0             0
15             1             0             0
16             1             0             0
17             0             1             0
18             0             1             0
19             0             1             0
```


Hjørnepunkt-parametrisering = "treatment-kontrast"

Vi kan velge en av gruppene som referansegruppe, f.eks. gruppe 1 og skrive om enveis ANOVA til

$$\begin{aligned} \mu_j = \mathrm{E}[Y_i] &= \mu_1 + (\mu_2 - \mu_1)x_{i2} + \cdots + (\mu_J - \mu_1)x_{iJ} \\ &= \beta_1 + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} \end{aligned}$$

der altså $\beta_1 = \mu_1$ og $\beta_j = \mu_j - \mu_1$ for $j > 1$.

Denne parametriseringen er naturlig hvis man vil sammenligne $J - 1$ nye behandlinger med en tradisjonell behandling.

Hjørnepunkt-parametrisering / treatment-contrast er default i R.

Eksempel inntekt med hjørnepunkt-parametrisering

```
> summary(lm(inntekt~factor(sted)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    326.875      9.733  33.583  < 2e-16 ***
factor(sted)2   47.500     13.765   3.451 0.002394 **
factor(sted)3   65.000     13.765   4.722 0.000116 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.53 on 21 degrees of freedom
Multiple R-Squared: 0.5321,     Adjusted R-squared: 0.4875
F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344

> anova(lm(inntekt~factor(sted)))
Analysis of Variance Table

Response: inntekt
             Df  Sum Sq Mean Sq F value   Pr(>F)
factor(sted)  2 18100.0  9050.0  11.941 0.000344 ***
Residuals    21 15915.6   757.9
---
```


Eks: Designmatrise for Sted med treatment-kontrast

```
> enveisfit<-lm(inntekt~factor(sted),x=T)
> enveisfit$x
   (Intercept) factor(sted)2 factor(sted)3
1            1             0             0
2            1             0             0
3            1             0             0
4            1             0             0
5            1             1             0
6            1             1             0
7            1             1             0
8            1             1             0
9            1             0             1
10           1             0             1
11           1             0             1
12           1             0             1
13           1             0             0
14           1             0             0
15           1             0             0
16           1             0             0
17           1             1             0
18           1             1             0
19           1             1             0
```


Sum-parametrisering (kontrast)

Tradisjonelt i ANOVA benyttes imidlertid ofte "sum-parametrisering" med

$$\mu_j = \mathrm{E}[Y_i] = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_J x_{iJ}$$

der

$$\alpha_1 + \alpha_2 + \cdots + \alpha_J = 0$$

Merk at $\sum_{j=1}^{J} x_{ij} = 1$ og med konstantledd α_0 i modellen er det overparametrisert uten en restriksjonen som $\sum_{j=1}^{J} \alpha_j = 0$

Med sum-parametrisering blir $\alpha_J = -(\alpha_1 + \cdots + \alpha_{J-1})$ og

$$\begin{aligned}
\mu_j &= \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{J-1} x_{i,J-1} - (\alpha_1 + \cdots + \alpha_{J-1}) x_{iJ} \\
&= \alpha_0 + \alpha_1 (x_{i1} - x_{iJ}) + \cdots + \alpha_{J-1} (x_{i,J-1} - x_{iJ}) \\
&= \alpha_0 + \alpha_1 x'_{i1} + \cdots + \alpha_{J-1} x'_{i,J-1}
\end{aligned}$$

Sum-parametrisering (kontrast), forts.

Sum-parametriseringen gir altså J parametre i - på samme måte som hjørnepunkt-paramterisering - men med kovariater

$$x'_{ij} = x_{ij} - x_{iJ}$$

Sum-kontrast spesifiseres i R ved

```
options(contrasts=c("contr.sum","contr.poly"))
```

Se bare bort fra `"contr.poly"` som benyttes for en spesiell type kategorisk kovariat.

For å komme tilbake til hjørnepunk/treatment-parametrisering:

```
options(contrasts=c("contr.treatment","contr.poly"))
```

Eks: Inntekt over sted med sum-kontrast

```
> options(contrasts=c("contr.sum","contr.poly"))
> summary(lm(inntekt~factor(sted)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    364.375      5.619  64.841  < 2e-16 ***
factor(sted)1  -37.500      7.947  -4.719 0.000117 ***
factor(sted)2   10.000      7.947   1.258 0.222090
---
Residual standard error: 27.53 on 21 degrees of freedom
Multiple R-Squared: 0.5321,     Adjusted R-squared: 0.4875
F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344

> anova(lm(inntekt~factor(sted)))
Analysis of Variance Table

Response: inntekt
             Df  Sum Sq Mean Sq F value   Pr(>F)
factor(sted)  2 18100.0  9050.0  11.941 0.000344 ***
Residuals    21 15915.6   757.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Eks: Designmatrise for Sted uten sum-kontrast

```
> options(contrasts=c("contr.sum","contr.poly"))
> enveisfit<-lm(inntekt~factor(sted),x=T)
> enveisfit$x
   (Intercept) factor(sted)1 factor(sted)2
1            1             1             0
2            1             1             0
3            1             1             0
4            1             1             0
5            1             0             1
6            1             0             1
7            1             0             1
8            1             0             1
9            1            -1            -1
10           1            -1            -1
11           1            -1            -1
12           1            -1            -1
13           1             1             0
14           1             1             0
15           1             1             0
16           1             1             0
17           1             0             1
18           1             0             1
```


Toveis variansanalyse uten interaksjon

$Y_i \sim \mathrm{N}(\mathrm{E}[Y_i], \sigma^2)$ uavhengige med

- nivå j på faktor 1 med ialt J nivåer
- nivå k på faktor 2 med ialt K nivåer

Med hjørnepunkt-parametrisering kodes 1. faktor ved $x_{ij} = 1$ mens $x_{ij'} = 0$ og 2. faktor ved $z_{ik} = 1$ mens $z_{ik'} = 0$ slik at forventningen blir (med $\beta_1 = \alpha_1 = 0$)

$$\mathrm{E}[Y_i] = \beta_0 + \sum_{j=2}^{J} \beta_j x_{ij} + \sum_{k=2}^{K} \alpha_k z_{ik} = \beta_0 + \beta_j + \alpha_k$$

Eks: Toveis-Anova med hjørnepunkt

```
> options(contrasts=c("contr.treatment","contr.poly"))
> toveisfit<-lm(inntekt~factor(sted)+factor(kjonn),x=T)
> summary(toveisfit)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     347.500      6.896  50.393  < 2e-16 ***
factor(sted)2    47.500      8.446   5.624 1.67e-05 ***
factor(sted)3    65.000      8.446   7.696 2.11e-07 ***
factor(kjonn)2  -41.250      6.896  -5.982 7.54e-06 ***
---
Residual standard error: 16.89 on 20 degrees of freedom
Multiple R-Squared: 0.8322,     Adjusted R-squared: 0.8071
F-statistic: 33.07 on 3 and 20 DF,  p-value: 6.012e-08

> anova(toveisfit)
Analysis of Variance Table
Response: inntekt
              Df  Sum Sq Mean Sq F value    Pr(>F)
factor(sted)   2 18100.0  9050.0  31.720 6.260e-07 ***
factor(kjonn)  1 10209.4 10209.4  35.783 7.537e-06 ***
Residuals     20  5706.2   285.3
```

Toveis ANOVA uten interaksjon med sum-kontrast

$Y_i \sim \mathrm{N}(\mathrm{E}[Y_i], \sigma^2)$ uavhengige med

- nivå j på faktor 1 med ialt J nivåer
- nivå k på faktor 2 med ialt K nivåer

Med x_{ij} ogz_{ik} som ved hjørnepunkt-parametrisering kodes nå 1. faktor ved $x'_{ij} = x_{ij} - x_{iJ}$ og 2. faktor ved $z'_{ik} = z_{ik} - z_{iK}$

Eks: Toveis-anova med sum-parametrisering

```
> options(contrasts=c("contr.sum","contr.poly"))
> toveisfit<-lm(inntekt~factor(sted)+factor(kjonn),x=T)
> summary(toveisfit)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     364.375      3.448 105.680  < 2e-16 ***
factor(sted)1   -37.500      4.876  -7.691 2.13e-07 ***
factor(sted)2    10.000      4.876   2.051   0.0536 .
factor(kjonn)1   20.625      3.448   5.982 7.54e-06 ***
---
Residual standard error: 16.89 on 20 degrees of freedom
Multiple R-Squared: 0.8322,     Adjusted R-squared: 0.8071
F-statistic: 33.07 on 3 and 20 DF,  p-value: 6.012e-08

> anova(toveisfit)
Analysis of Variance Table
Response: inntekt
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(sted)   2 18100.0  9050.0  31.720 6.260e-07 ***
factor(kjonn)  1 10209.4 10209.4  35.783 7.537e-06 ***
Residuals     20  5706.2   285.3
```


Toveis anova med interaksjon

$Y_i \sim \mathrm{N}(\mathrm{E}[Y_i], \sigma^2)$ uavhengige med

- nivå j på faktor 1 med ialt J nivåer
- nivå k på faktor 2 med ialt K nivåer

og

$$\mathrm{E}[Y_i] = \alpha_0 + \beta_j + \gamma_k + (\beta\gamma)_{jk},$$

dvs. et nivå for hver kombinasjon nivå j på faktor 1 og nivå k på faktor 2.

Eks: Toveis-anova med sum-parametrisering

```
> toveisfit<-lm(inntekt~factor(sted)+factor(kjonn)
                        +factor(sted)*factor(kjonn))

> anova(toveisfit)
Analysis of Variance Table

Response: inntekt
                           Df  Sum Sq Mean Sq F value    Pr(>F)
factor(sted)                2 18100.0  9050.0 29.0569 2.314e-06 ***
factor(kjonn)               1 10209.4 10209.4 32.7793 1.988e-05 ***
factor(sted):factor(kjonn)  2   100.0    50.0  0.1605    0.8529
Residuals                  18  5606.2   311.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predikerte verdier og Hat-matrisen

Med estimater MKE $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ får vi predikerte verdier for μ_i og for Y_i ved

$$\hat{\mu}_i = \hat{Y}_i = \hat{\beta}'\mathbf{x_i}$$

For vektoren $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ blir disse gitt ved

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

der

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

kalles Hat-matrisen fordi den "setter hatt på" $\mathbf{Y}$.

Residualer på matriseform

La $\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{\mu}_i$ være de vanlige residualene og $\hat{\mathbf{e}} = (\hat{e}_1, \ldots, \hat{e}_n)'$ vektoren av residualer. Da finner vi

$$\hat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{I}\mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y}$$

der $\mathbf{I}$ er nxn identitetsmatrisen og $\mathbf{M} = \mathbf{I} - \mathbf{H}$.

Siden $\hat{\mathbf{e}}$ er lineært avhengig av $\mathbf{Y}$ er de normalfordelt med forventning

$$\mathrm{E}[\hat{\mathbf{e}}] = \mathbf{M}\mathrm{E}[\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{Y}] = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\beta = \mathbf{0}$$

og kovariansmatrise

$$\mathbf{M}\sigma^2\mathbf{I}\mathbf{M}' = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{H})$$

hvor første likhet vises på neste side.

H og M idempotent og symmetriske

For en symmetrisk matrise A er $A = A'$

En *idempotent* matrise A tilfredstiller $A^2 = AA = A$.

Resultat: $\mathbf{H}$ og $\mathbf{M}$ er symmetriske og idempotente.

Viser at $\mathbf{H}$ er idempotent:

$$\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$$

Siden dette også gjelder for $\mathbf{M}$ (oppgave!) får vi

$$\mathbf{M}\sigma^2\mathbf{I}\mathbf{M}' = \sigma^2\mathbf{M}\mathbf{M} = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{H})$$

Studentiserte residualer og "leverage"

Med $\mathbf{H} = [h_{ij}]_{i,j=1}^n$ får vi at standardavviket til residualen $\hat{e}_i$ er lik $\sigma\sqrt{(1-h_{ii})}$, dvs. avhenger av h_{ii}.
Dette foreslår at vi bør se på Studentifiserte residualer

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

heller enn $\hat{e}_i = Y_i - \hat{Y}_i$.

Størrelsene h_{ii} kalles "leverage" = "moment på vektstang" og angir evnen til å påvirke $\hat{\beta}$: Influens.

Oppgave: Vis at for enkel lineær regresjon, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, blir leverage

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Delta-Betas og Cook's distanse

Observasjoner $Y_i, \mathbf{x}_i$ som har stor innflytelse på $\hat{\beta}$ kan også observeres ved såkalte Delta-Betas (eller noen ganger df-betas)

$$\Delta_i \hat{\beta}_j = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

der $\hat{\beta}_{j(i)}$ er MKE for β_j observasjon utelates fra estimeringen.

Utfra uttrykk som leverage og influens kan man vel vente at det er en sammenheng mellom $\Delta_i \hat{\beta}_j$ og leverages h_{ii}. Denne kan uttrykkes som gjennom *Cook's avstand*

$$D_i = \frac{1}{p} \frac{h_{ii}}{1 - h_{ii}} r_i^2 = \frac{1}{p\hat{\sigma}^2} (\hat{\beta} - \hat{\beta}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})$$

der $\hat{\beta}_{(i)}) = (\hat{\beta}_{1(i)}, \ldots, \hat{\beta}_{p(i)})'$.

Delta-Betas og Cook's distanse, forts.

Man ser av denne ligningnen at estimatene påvirkes betydelig dersom

- Studentisert residual r_i er stor
- Leverage h_{ii} er stor

Som en tommelfingerregel bør verdier sjekkes

- hvis $D_i > 0.5$
- alltid hvis $D_i > 1$

(Cook & Weisberg, 1999, Applied Regression Including Computing and Graphics).

Delta-Betas og Cook's distanse, forts. II

En intuitiv begrunnelse for disse kriteriene fås ved å legge merke til at

$$\frac{1}{p\hat{\sigma}^2} (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \sim F_{p,n-p}$$

og siden senter i F-fordelinger ≈ 1 vil en endring i $D_i \approx 1$ være verdt å merke seg.

Eksempel: Fødselsvekt mot kjønn og sv.lengde

Residualplottet nederst til høyre gir Cook's distanse D_i, max-verdi ≈ 0.5.