Tue Tjur
April 2007

Stat. MØK 2. år

# Statistik

**Oversigt over begreber og modeller**

## Kapitel 1: Likelihood metoden

En **statistisk model** til beskrivelse af en **observation** $x$ fra **observationsrummet** $E$ er en parametriseret familie $\{P_\vartheta \mid \vartheta \in \Theta\}$ af sandsynlighedsfordelinger på $E$. Her vil **parameterrummet** $\Theta$ sædvanligvis være en åben delmængde af et reelt talrum $\mathbf{R}^d$. I så fald kaldes **parameteren** $\vartheta$ også for **parametervektoren** og skrives $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$, og dens koordinater $\vartheta_1, \ldots, \vartheta_d$ omtales også som parametre. Hvis parametriseringen er injektiv kaldes $d$ modellens **dimension**. Fortolkningen af den statistiske model går ud på, at $x$ er det observerede udfald af en stokastisk variabel $X$ med fordeling $P_\vartheta$ for en eller anden (ukendt) værdi af $\vartheta$.

**Likelihoodfunktionen** eller **likelihooden** $L : \Theta \to \mathbf{R}$ er givet ved $L(\vartheta) = p_\vartheta(x)$, hvor $p_\vartheta$ er sandsynlighedsfunktionen eller tætheden hørende til $P_\vartheta$. **Log–likelihooden** $l(\vartheta) = \log L(\vartheta)$ er den naturlige logaritme til likelihooden.

**Estimation** (dvs. skønsmæssig bestemmelse) af den ukendte parameter foretages oftest ved **maksimum likelihood metoden**, som går ud på at gætte på den værdi $\hat{\vartheta}$ af $\vartheta$ der giver likelihooden (eller log–likelihooden) sin maksimale værdi. $\hat{\vartheta}$ kaldes **maksimaliseringsestimatoren** eller **ML–estimatoren**. I praksis udregnes $\hat{\vartheta}$ som løsning til **likelihoodligningerne**, som er de ligninger der sætter log–likelihoodens partielle differentialkvotienter lig med nul.

**Usikkerheden på et estimat** angives normalt ved hjælp af (eksakte eller approksimative) **konfidens**– eller **sikkerhedsintervaller**. For eksempel defineres et 95% sikkerhedsinterval for en reel parameter (eller parameterfunktion) $\eta = \eta(\theta)$ som et interval $I(x) = [\eta_{\text{nedre}}(x), \eta_{\text{øvre}}(x)]$ med den egenskab, at hændelsen $\{\eta \in I(X)\}$ har sandsynlighed (eksakt eller approksimativt) 0.95. Intervallets endepunkter kaldes **konfidens**– eller **sikkerhedsgrænser**.

**Test** af en **hypotese** af formen $\vartheta \in \Theta_0$, hvor $\Theta_0 \subset \Theta$ typisk er en delmængde af lavere dimension $d_0$, foretages sædvanligvis ved hjælp af et **kvotienttest**, som er baseret på **kvotientteststørrelsen**

$$-2\log q = -2\log\frac{L(\hat{\vartheta}_0)}{L(\hat{\vartheta})} = -2\log\left(\frac{\max_{\vartheta\in\Theta_0} L(\vartheta)}{\max_{\vartheta\in\Theta} L(\vartheta)}\right) = 2\left(l(\hat{\vartheta}) - l(\hat{\vartheta}_0)\right)$$

hvor $\hat{\vartheta}_0$ betegner maksimaliseringsestimatoren i den **reducerede model** $\{P_\vartheta \mid \vartheta \in \Theta_0\}$. I mange tilfælde gælder det, at fordelingen af den tilsvarende stokastiske variabel $-2\log Q$ under hypotesen er approksimativt uafhængig af $\vartheta \in \Theta_0$, og under passende omstændigheder kan den approksimeres med en $\chi^2$–fordeling med $d-d_0$ frihedsgrader. Hypotesen **forkastes** hvis den observerede værdi $-2\log q$ af $-2\log Q$ er ekstremt stor i sin fordeling under hypotesen. Som mål for graden af **signifikans** — dvs. den sikkerhed, hvormed hypotesen kan afvises — benyttes **P–værdien**, som er halesandsynligheden i fordelingen af $-2\log Q$ under hypotesen, regnet fra $-2\log q$ og udefter; altså sandsynligheden for at få en lige så stor eller endnu større teststørrelse under antagelse af at hypotesen er sand. En *lille* P–værdi svarer til *høj* grad af signifikans, dvs. høj grad af uoverensstemmelse mellem observation og hypotese. Omvendt betyder en stor P–værdi at hypotesen kan **godkendes**. I almindelighed regnes P–værdier over 0.05 for **insignifikante**, medens P–værdier under 0.01 (undertiden helt op til 0.05) anses for signifikante. Udsagn om signifikans bør normalt gradueres, bedst ved angivelse af selve P–værdien, eller (hvis man mangler værktøj til udregning af de relevante halesandsynligheder) ved brug af følgende konvention: Hvis en P–værdi ved tabelopslag findes at være (for eksempel) under 0.005 (dvs. hvis teststørrelsen er større end 99.5%–fraktilen i den relevante fordeling) siger man at testet **udviser signifikans på 0.5%–niveauet**.

## Kapitel 2: Estimation og testning i en binomialfordeling

**Model:** $x$ er (en observation af en stokastisk variabel $X$ som er) binomialfordelt med antalsparameter $n$ og sandsynlighedsparameter $p \in ]0,1[$.

**ML–estimatoren** for $p$ er

$$\hat{p} = \frac{x}{n}\,.$$

Approksimative **95% sikkerhedsgrænser** udregnes således:

$$p = \hat{p} \pm 1.96\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})} = \frac{x}{n} \pm 1.96\sqrt{\frac{x(n-x)}{n^3}}\,.$$

Test for en **simpel hypotese** af formen $p = p_0$ foretages enten ved hjælp af **kvotientteststørrelsen**

$$2\left(x\log\frac{x}{n} + (n-x)\log\frac{n-x}{n} - x\log p_0 - (n-x)\log(1-p_0)\right)$$

eller **Pearsons teststørrelse**

$$\frac{(x-np_0)^2}{np_0(1-p_0)}\,.$$

Disse størrelser skal vurderes i en $\chi^2$–fordeling med 1 frihedsgrad. Testet baseret på Pearsons teststørrelse kan også foretages ved tosidet vurdering af teststørrelsen

$$u = \frac{x-np_0}{\sqrt{np_0(1-p_0)}}$$

i en normeret normalfordeling. Hvis antallene $x$ og $n-x$ er små kan det være en fordel at anvende **kontinuitetskorrektion**, som går ud på at ændre Pearsons teststørrelse til

$$\left(\frac{|x-np_0|-\frac{1}{2}}{\sqrt{np_0(1-p_0)}}\right)^2 .$$

Endnu bedre er et **eksakt binomialtest**, hvor P–værdien direkte udregnes som sandsynligheden for, at en binomialfordelt stokastisk variabel med parametre $(n, p_0)$ ligger mindst lige så langt fra sin forventede værdi $np_0$ som selve observationen $x$:

$$P(|X-np_0| \geq |x-np_0|) \approx \begin{cases} 2\sum_{k=x}^{n}\binom{n}{k}p_0^k(1-p_0)^{n-k} & \text{for } x \geq np_0, \\ 2\sum_{k=0}^{x}\binom{n}{k}p_0^k(1-p_0)^{n-k} & \text{for } x \leq np_0. \end{cases}$$

Et **ensidet test** benyttes i helt specielle situationer, hvor hypotesen $p = p_0$ skal testes mod et alternativ af formen $p > p_0$. I så fald regnes en observation $x$ som er mindre end $np_0$ under alle omstændigheder for insignifikant, medens en observation som er større end $np_0$ giver anledning til P–værdien

$$P(X \geq x) = \sum_{k=x}^{n}\binom{n}{k}p_0^k(1-p_0)^{n-k} .$$

## Kapitel 3: Sammenligning af to binomialfordelinger

**Model:** $x_1$ og $x_2$ er stokastisk uafhængige, binomialfordelte med antalsparamtre $n_1$ og $n_2$ og sandsynlighedsparametre $p_1$ og $p_2$.

$p_1$ og $p_2$ estimeres hver for sig i "deres egne" modeller som i kapitel 2, og sikkerhedsgrænser udregnes også på samme måde.

**Test for hypotesen** $p_1 = p_2$. **ML–estimatoren** for den fælles sandsynlighedsparameter i den reducerede model givet ved $p_1 = p_2 = p$ er $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$. Kvotientteststørrelsen udregnes ved at man på hver plads i antalstabellen

$$\begin{array}{ccc} x_1 & n_1 - x_1 & n_1 \\ x_2 & n_2 - x_2 & n_2 \\ & & \\ x. & n. - x. & n. \end{array}$$

(inklusive række– og søjlesummer og totalsummen) anvender funktionen $2x \log x$. Kvotientteststørrelsen $-2\log q$ er en sum af disse størrelser, regnet med fortegn, således at der for selve indgangene i tabellen og totalsummen benyttes positivt fortegn, for række– og søjlesummer negativt. Man kan også benytte **Pearsons teststørrelse**

$$\frac{(\hat{p}_1-\hat{p}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})} = \left(\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}\right)^2 .$$

Begge disse teststørrelser skal vurderes i deres approksimative fordeling under hypotesen, som er en $\chi^2$–fordeling med 1 frihedsgrad. Hvis nogle af tallene i tabellen er små (under 5) er det sikrere at bruge **Fisher's eksakte test**. Dette forudsætter at man har adgang til et program som kan udregne halesandsynligheder i den hypergeometriske fordeling. Dette test går ud på at undersøge om $x_1$ ligger ekstremt langt ude i en af halerne af den hypergeometriske fordeling, der kan fortolkes som fordelingen af antal røde kugler i en stikprøve på $x.$ kugler udtaget uden tilbagelægning fra en kasse med $n_1$ røde og $n_2$ hvide kugler.

## Kapitel 4: Estimation og testning i en polynomialfordeling

**Model:** $x = (x_1, \ldots, x_k)$ er polynomialfordelt af orden $k$ med antalsparameter $n$ og ukendte sandsynlighedsparametre $(p_1, \ldots, p_k)$.

Observationsrummet er

$$D(n, k) = \{(x_1, \ldots, x_k) \in \mathbf{N}_0^k \mid x_1 + \cdots + x_k = n\},$$

og parameterrummet svarende til **den fulde model** er

$$\Delta_k = \{(p_1, \ldots, p_k) \in \mathbf{R}_+^k \mid p_1 + \cdots + p_k = 1\}.$$

Dimensionen af den fulde model er $k - 1$.

**ML–estimatorerne** for sandsynlighedsparametrene er de tilsvarende relative hyppigheder $\hat{p}_j = \frac{x_j}{n}$, og **sikkerhedsgrænser** kan udregnes ligesom i kapitel 2 (idet den marginale fordeling af $x_j$ er binomialfordelingen med antalsparameter $n$ sandsynlighedsparameter $p_j$).

**Test for dimensionsreducerende hypotese** af formen $(p_1, \ldots, p_k) \in \Theta_0$, hvor $\Theta_0 \subset \Delta_k$ er en glat delmængde af dimension $d_0 < k - 1$ kan foretages ved hjælp af **kvotientteststørrelsen**

$$2\left(x_1 \log \frac{\hat{p}_1}{\hat{p}_{01}} + \cdots + x_k \log \frac{\hat{p}_k}{\hat{p}_{0k}}\right)$$

hvor $\hat{p}_{01}, \ldots, \hat{p}_{0k}$ betegner ML–estimatorerne i den reducerede model. Alternativt kan man bruge **Pearsons teststørrelse**

$$\sum_{j=1}^{k} \frac{(x_j - n\hat{p}_{0j})^2}{n\hat{p}_{0j}}.$$

Begge teststørrelser skal vurderes i en $\chi^2$–fordeling med $k - 1 - d_0$ frihedsgrader. Som tommelfingerregel er fordelingsapproksimation brugbar når alle de **fittede værdier** $n\hat{p}_{0j}$ er $\geq 5$. I modsat fald må man tage passende forbehold vedrørende P–værdiens præcision.

**Test for homogenitet** betyder i denne sammenhæng test for den simple hypotese $p_1 = \cdots = p_k = \frac{1}{k}$ imod den fulde model. Kvotientteststørrelsen, som i dette tilfælde kan skrives

$$-2\log q = 2\left(x_1 \log x_1 + \cdots + x_k \log x_k - n\log n + n\log k\right)$$

skal her vurderes i en $\chi^2$–fordeling med $k-1$ frihedsgrader.

Når der er flere hypoteser i spil bør man normalt bruge **successiv testning.** Hvis hypotesen $\vartheta \in \Theta_0$ er godkendt, og man derefter ønsker at teste for yderligere reduktion til en model bestemt ved en glat delmængde $\Theta_{00} \subset \Theta_0$ af dimension $d_{00} < d_0$, tager man den sidst accepterede hypotese som udgangspunkt og benytter **kvotientteststørrelsen**

$$2\left(x_1 \log \frac{\hat{p}_{01}}{\hat{p}_{001}} + \cdots + x_k \log \frac{\hat{p}_{0k}}{\hat{p}_{00k}}\right)$$

eller **Pearsons teststørrelse**

$$\sum_{j=1}^{k} \frac{\left(n\hat{p}_{0j} - n\hat{p}_{00j}\right)^2}{n\hat{p}_{00j}}\,.$$

Begge størrelser skal vurderes i en $\chi^2$–fordeling med $d_0 - d_{00}$ frihedsgrader. Igen forudsættes, som en forsigtig tommelfingerregel, at alle fittede værdier $n\hat{p}_{00j}$ i den yderligere reducerede model er $\geq 5$.

## Kapitel 5: Analyse af tosidede antalstabeller

**Model: Data** (dvs. observationen) er givet i form af en tosidet antalstabel

$$(x_{rs} \mid r = 1, \ldots, R,\ s = 1, \ldots, S)$$

som antages at være en observation fra en polynomialfordeling af orden $RS$, med antalsparameter $n = x_{\cdot\cdot}$ og ukendte sandsynlighedsparametre

$$(p_{rs} \mid r = 1, \ldots, R,\ s = 1, \ldots, S) \in \Delta_{RS}\,.$$

Dimensionen af den fulde model er $RS - 1$. Den vigtigste reducerede model er givet ved

**Hypotesen om uafhængighed,** som går ud på at sandsynlighedsparametrene antager formen

$$p_{rs} = \rho_r \sigma_s$$

hvor

$$((\rho_1, \ldots, \rho_R), (\sigma_1, \ldots, \sigma_S)) \in \Delta_R \times \Delta_S.$$

Bemærk, at der i så fald gælder $\rho_r = p_{r\cdot}$ og $\sigma_s = p_{\cdot s}$. **ML–estimatorerne** i uafhængighedsmodellen er givet ved

$$\hat{\rho}_r = \frac{x_{r\cdot}}{n} \text{ og } \hat{\sigma}_s = \frac{x_{\cdot s}}{n}$$

så estimatorerne for selve sandsynlighedsparametrene under uafhængighedshypotesen bliver

$$\hat{p}_{rs} = \frac{x_{r\cdot} x_{\cdot s}}{n^2}.$$

**Kvotientteststørrelsen** (ved test for uafhængighedshypotesen mod den fulde model) bliver

$$2\left(\sum_{r=1}^{R}\sum_{s=1}^{S} x_{rs}\log x_{rs} - \sum_{r=1}^{R} x_{r\cdot}\log x_{r\cdot} - \sum_{s=1}^{S} x_{\cdot s}\log x_{\cdot s} + n\log n\right)$$

og Pearsons teststørrelse er

$$\sum_{r=1}^{R}\sum_{s=1}^{S}\frac{\left(x_{rs} - \frac{x_{r\cdot} x_{\cdot s}}{x_{\cdot\cdot}}\right)^2}{\frac{x_{r\cdot} x_{\cdot s}}{x_{\cdot\cdot}}}$$

Disse teststørrelser skal vurderes i en $\chi^2$–fordeling med $(R-1)(S-1)$ frihedsgrader.

Størrelserne

$$u_{rs} = \frac{x_{rs} - \frac{x_{r\cdot} x_{\cdot s}}{n}}{\sqrt{\frac{x_{r\cdot} x_{\cdot s}}{n}}}.$$

(hvis kvadratsum netop er Pearsons teststørrelse) kaldes de **normerede residualer**. Hvis uafhængighedsmodellen holder vil de typisk ligge mellem -2 og 2, numerisk store værdier fortæller hvor de væsentlige afvigelser fra uafhængighedsmodellen ligger.

## Kapitel 6: Poisson modeller

**Én poissonfordelt variabel.** Lad $x$ være en observation fra en Poissonfordeling med parameter $\lambda > 0$. **ML–estimatoren** for $\lambda$ er så givet ved $\hat{\lambda} = x$. Test for en **simpel hypotese** af formen $\lambda = \lambda_0$ kan foretages v.h.a. **kvotientteststørrelsen** $2((x\log x - x) - (x\log\lambda_0 - \lambda_0))$ eller ved hjælp af **Pearsons teststørrelse** $\frac{(x-\lambda_0)^2}{\lambda_0}$. Begge skal vurderes i en $\chi^2$–fordeling med 1 frihedsgrad.

Generelt ser vi på følgende

**Model:** $x_1, \ldots, x_k$ er stokastisk uafhængige, Poissonfordelte med parametre $\lambda_1, \ldots, \lambda_k > 0$. I **den fulde model**, hvor parametrene $\lambda_1, \ldots, \lambda_k$ varierer frit, er **ML–estimatorerne** givet ved $\hat{\lambda}_j = x_j$.

**Test imod den fulde model** af en hypotese svarende til en glat delmængde $\Theta_0$ af $]0,+\infty[^k$ af dimension $d_0 < k$ kan foretages ved hjælp af **kvotientteststørrelsen**

$$2\sum_{j=1}^{k}\left(x_j \log\frac{x_j}{\hat\lambda_{0j}} + \hat\lambda_{0j} - x_j\right)\left[= 2\sum_{j=1}^{k} x_j \log\frac{x_j}{\hat\lambda_{0j}}\right]$$

eller **Pearsons teststørrelse**

$$-2\log q \approx \sum_{j=1}^{k}\frac{(x_j - \hat\lambda_{0j})^2}{\hat\lambda_{0j}}.$$

Det simplere udtryk i kantet parentes er gyldigt, når den reducerede model er **invariant under multiplikation med positiv skalar**, dvs. når det for ethvert punkt $(\lambda_1,\ldots,\lambda_k) \in \Theta_0$ og ethvert tal $\beta > 0$ gælder at $(\beta\lambda_1,\ldots,\beta\lambda_k) \in \Theta_0$. Begge teststørrelser skal vurderes i en $\chi^2$–fordeling med $k - d_0$ frihedsgrader. Denne approksimation anses, som tommelfingerregel, for brugbar når alle fittede værdier $\hat\lambda_{0j}$ i den reducerede model er $\geq 5$.

**Successiv testning.** Test for en yderligere reduceret model givet ved en glat delmængde $\Theta_{00} \subset \Theta_0$ af dimension $d_{00} < d_0$ foretages ved hjælp af **kvotientteststørrelsen**

$$2\sum_{j=1}^{k}\left(x_j \log\frac{\hat\lambda_{0j}}{\hat\lambda_{00j}} + \hat\lambda_{00j} - \hat\lambda_{0j}\right)\left[= 2\sum_{j=1}^{k} x_j \log\frac{\hat\lambda_{0j}}{\hat\lambda_{00j}}\right]$$

eller **Pearsons teststørrelse**

$$\sum_{j=1}^{k}\frac{(\hat\lambda_{0j} - \hat\lambda_{00j})^2}{\hat\lambda_{00j}}.$$

Igen kan den simplere formel i kantet parentes kun bruges, når *begge* modeller er invariante under multiplikation med positiv skalar.

**Test for homogenitet.** Ved test for hypotesen $\lambda_1 = \cdots = \lambda_k$ imod den fulde model bliver kvotientteststørrelsen den samme som bruges ved test for homogenitet i polynomialfordelingen, altså

$$-2\log q = 2\left(x_1\log x_1 + \cdots + x_k \log x_k - n\log n + n\log k\right)$$

og den skal også her vurderes i en $\chi^2$–fordeling med $k-1$ frihedsgrader.

**Proportionalitet med given baggrundsvariabel:** Denne model er givet ved $\lambda_i = \beta z_i$, hvor $z_1,\ldots,z_k$ er givne tal. **ML–estimatoren** for

$\beta$ er $\hat\beta = x./z.$ . Test imod den fulde model foretages ved vurdering af kvotientteststørrelsen

$$2\sum x_j \log \frac{x_j}{\hat\beta z_j}$$

eller Pearsons teststørrelse

$$\sum \frac{(x_j - \hat\beta z_j)^2}{\hat\beta z_j}$$

i en $\chi^2$–fordeling med $k-1$ frihedsgrader.

**Den multiplikative model** for en tosidet antalstabel $(x_{rs})$ er givet ved $\lambda_{rs} = \alpha_r \beta_s$. **ML–estimatorerne** i denne model er givet ved at de fittede værdier bliver

$$\hat\lambda_{rs} = \frac{x_{r\cdot} x_{\cdot s}}{x_{\cdot\cdot}} .$$

Testet for denne model imod den fulde model foretages på samme måde som testet for *uafhængighed* i den tilsvarende polynomialfordelingsmodel, se kapitel 5.

## Kapitel 7: Uafhængige id. ford. normale observationer

**Model:** $y_1, \ldots, y_n$ er uafhængige, identisk normalt fordelte med middelværdi $\mu$ og varians $\sigma^2$.

**Kendt varians. ML–estimatet** for middelværdien bliver $\hat\mu = \bar y$. Et **eksakt 95% sikkerhedsinterval** for $\mu$ er givet ved

$$\left[\bar y - 1.96\sqrt{\frac{\sigma^2}{n}}, \bar y + 1.96\sqrt{\frac{\sigma^2}{n}}\right]$$

**Test af simpel hypotese** $\mu = \mu_0$ foretages ved tosidet vurdering af størrelsen $u = \frac{\bar y - \mu_0}{\sqrt{\sigma^2/n}}$ i en normeret normalfordeling.

**Ukendt varians. ML–estimatet** for middelværdien bliver igen $\hat\mu = \bar y$. Som estimat for variansen benyttes det **korrigerede ML–estimat**

$$\hat\sigma^2 = \frac{\mathrm{SSD}_y}{n-1} \quad \text{hvor} \quad \mathrm{SSD}_y = \sum (y_i - \bar y)^2 .$$

Et (eksakt) **95% sikkerhedsinterval** for $\mu$ er givet ved

$$\left[\bar y - t_{n-1}(97.5\%)\sqrt{\frac{\hat\sigma^2}{n}}, \bar y + t_{n-1}(97.5\%)\sqrt{\frac{\hat\sigma^2}{n}}\right]$$

hvor $t_{n-1}(97.5\%)$ betegner 97.5%–fraktilen i en T–fordeling med $n-1$ frihedsgrader.

**Test af simpel hypotese** $\mu = \mu_0$ foretages ved tosidet vurdering af størrelsen $t = \frac{\bar{y}-\mu_0}{\sqrt{\hat{\sigma}^2/n}}$ i en T–fordeling med $n-1$ frihedsgrader.

**Modelkontrol** foretages ved tegning af et histogram over observationerne (som så i passende grad skal ligne en sædvanlig klokkeformet normalfordelingskurve) eller ved tegning af et **fraktil**– eller **probitdiagram**, hvor punkterne

$$\left(y_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right) \ , \quad i = 1, \ldots, n$$

indtegnes i et koordinatsystem; de skal approksimativt ligge på en linie.

Ved **beregningerne** anvendes formlerne

$$\bar{y} = \mathrm{S}_y/n \text{ og } \mathrm{SSD}_y = \mathrm{SS}_y - \mathrm{S}_y^2/n,$$

hvor

$$\mathrm{S}_y = \sum y_i \ , \quad \mathrm{SS}_y = \sum y_i^2 \ .$$

## Kapitel 8: Regressionsanalyse

**Model:** $y_1, \ldots, y_n$ er uafhængige, normalfordelte med samme varians $\sigma^2$ og middelværdier $\mathrm{E}y_i = \mu_i = \alpha + \beta x_i$, hvor $x_1, \ldots, x_n$ er givne (faste) tal. I denne forbindelse kaldes $x$ den **forklarende** eller **uafhængige variabel**, medens $y$ kaldes den **afhængige variabel** eller **responsen**.

Betegnelser for **mellemregningsstørrelser:**

$$\mathrm{S}_x = \sum x_i, \qquad \bar{x} = \frac{1}{n}\mathrm{S}_x, \qquad \mathrm{SS}_x = \sum x_i^2,$$

$$\mathrm{SSD}_x = \sum (x_i - \bar{x})^2 = \mathrm{SS}_x - \frac{1}{n}\mathrm{S}_x^2,$$

$$\mathrm{S}_y = \sum y_i, \qquad \bar{y} = \frac{1}{n}\mathrm{S}_y, \qquad \mathrm{SS}_y = \sum y_i^2,$$

$$\mathrm{SSD}_y = \sum (y_i - \bar{y})^2 = \mathrm{SS}_y - \frac{1}{n}\mathrm{S}_y^2,$$

$$\mathrm{SP}_{xy} = \sum x_i y_i, \qquad \mathrm{SPD}_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \mathrm{SP}_{xy} - \frac{1}{n}\mathrm{S}_x\mathrm{S}_y.$$

**ML–estimaterne** for **middelværdiparametrene** bliver

$$\hat{\beta} = \frac{\mathrm{SPD}_{xy}}{\mathrm{SSD}_x} \ , \qquad \hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} \ .$$

Variansen estimeres ved den **korrigerede ML–estimator**

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}}{n-2}.$$

hvor **residualkvadratsummen**

$$\mathrm{SSD}_{\mathrm{res}} = \sum \big(y_i - (\hat{\alpha} + \hat{\beta} x_i)\big)^2$$

kan beregnes v.h.a. de mellemregningsstørrelser vi har indført som

$$\mathrm{SSD}_{\mathrm{res}} = \mathrm{SSD}_y - \frac{\mathrm{SPD}_{xy}^2}{\mathrm{SSD}_x} = \mathrm{SSD}_y - \hat{\beta}^2 \mathrm{SSD}_x.$$

**Estimatorernes fordeling:** $\hat{\beta}$ er normalfordelt med middelværdi $\beta$ og varians $\sigma^2/\mathrm{SSD}_x$. Heraf følger at **95% sikkerhedsgrænser** for $\beta$ kan udregnes efter formlen

$$\beta = \hat{\beta} \pm t_{n-2}(97.5\%)\sqrt{\frac{1}{\mathrm{SSD}_x}\hat{\sigma}^2}\,.$$

Estimatoren $\hat{\alpha} + \hat{\beta} x_0$ for middelværdien af en hypotetisk observation, svarende til værdien $x_0$ af den forklarende variabel, er normalfordelt med middelværdi $\alpha + \beta x_0$ og varians

$$\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSD}_x}\right)\sigma^2$$

hvoraf følger at **95% sikkerhedsgrænser** for denne parameterfunktion kan udregnes som

$$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2}(97.5\%)\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SSD}_x}\right)\hat{\sigma}^2}\,.$$

Specielt fås (for $x_0 = 0$) **95% sikkerhedsgrænserne**

$$\alpha = \hat{\alpha} \pm t_{n-2}(97.5\%)\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SSD}_x}\right)\hat{\sigma}^2}$$

for afskæringen $\alpha$.

**Test for** $\beta = 0$ kan foretages ved (tosidet) vurdering af

$$t = \frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\mathrm{SSD}_x}}}$$

i en T–fordeling med $n - 2$ frihedsgrader.

**Model:** $y_{gi}$, $g = 1, \ldots, G$, $i = 1, \ldots, n_g$ er uafhængige, normalfordelte med samme varians $\sigma^2$ og middelværdier $\mathrm{E}y_{gi} = \mu_g$.

**ML–estimatorerne** for middelværdierne i de enkelte grupper er givet ved $\hat{\mu}_g = \bar{y}_g.$ . Variansen estimeres ved

$$\hat{\sigma}^2 = \frac{\mathrm{SSD_{res}}}{n - G}$$

hvor **residualkvadratsummen** $\mathrm{SSD_{res}}$ er givet ved

$$\mathrm{SSD_{res}} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g.)^2 = \sum_{g=1}^{G} \mathrm{SSD}_y^g\,,$$

hvor størrelserne $\mathrm{SSD}_y^g$ kan beregnes ud fra tallene i gruppe $g$ på nøjagtig samme måde som $\mathrm{SSD}_y$ i kapitel 7 blev udregnet på basis af hele datasættet. **95% sikkerhedsgrænser** for selve gruppemiddelværdierne er givet ved

$$\mu_g = \bar{y}_g. \pm t_{n-G}(97.5\%)\sqrt{\frac{\hat{\sigma}^2}{n_g}}$$

medens estimater og 95% sikkerhedsintervaller for **kontraster** er givet ved

$$\mu_{g'} - \mu_{g''} = \bar{y}_{g'}. - \bar{y}_{g''}. \pm t_{n-G}(97.5\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)}.$$

**Test for homogenitet**, dvs. for hypotesen $\mu_1 = \cdots = \mu_G$ foretages ved vurdering af teststørrelsen

$$f = \frac{(\mathrm{SSD}_y - \mathrm{SSD_{res}})/(G-1)}{\mathrm{SSD_{res}}/(n-G)}$$

i en F–fordeling med $(G-1, n-G)$ frihedsgrader.

En **parvis sammenligning**, dvs. et test for en hypotese af typen $\mu_{g'} = \mu_{g''}$, foretages ved tosidet vurdering af teststørrelsen

$$t = \frac{\bar{y}_{g'}. - \bar{y}_{g''}.}{\sqrt{\left(\frac{1}{n_{g'}} + \frac{1}{n_{g''}}\right)\hat{\sigma}^2}}$$

i en T–fordeling med $n - G$ frihedsgrader.

**Bartletts test for varianshomogenitet** foretages ved vurdering af teststørrelsen

$$b = \frac{(n-G)\log\frac{\mathrm{SSD}_{\mathrm{res}}}{n-G} - \sum_{g=1}^{G}(n_g-1)\log\frac{\mathrm{SSD}_y^g}{n_g-1}}{1+\frac{1}{3(G-1)}\left(\left(\sum_{g=1}^{G}\frac{1}{n_g-1}\right)-\frac{1}{n-G}\right)}$$

i en $\chi^2$–fordeling med $G-1$ frihedsgrader. For $G=2$ kan man benytte et eksakt test, hvor

$$f = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\mathrm{SSD}_y^1/(n_1-1)}{\mathrm{SSD}_y^2/(n_2-1)}$$

vurderes **tosidet** i en F–fordeling med $(n_1-1, n_2-1)$ frihedsgrader.

## Kapitel 10: Lineære normale modeller

**Model:** $y_1, \ldots, y_n$ er uafhængige, normalfordelte med samme varians $\sigma^2$ og middelværdier $\mu_i$ givet ved et **lineært udtryk**. Et sådant lineært udtryk er i praksis givet som en sum af led, der kan antage følgende former:

$\gamma$: Konstantled.

$\alpha_f$: En effekt af en faktor med niveauer $f = 1, \ldots, F$, også kaldet en **hovedvirkning** af faktoren $F$.

$\alpha_{fg}$: En effekt af **produktet** eller **krydsklassifikationen** af to faktorer med niveauer $f = 1, \ldots, F$ og $g = 1, \ldots, G$, også kaldet en **vekselvirkning af første orden** eller en **to–faktor vekselvirkning**

$\alpha_{fgh}$, effekten af tre faktorer med **anden ordens** eller **tre–faktor vekselvirkning**.

...

$\beta x_i$: lineær effekt af regressionsvariabel $x$.

$\beta_f x_i$: lineær effekt af regressionsvariabel hvor hældningen afhænger af niveauet af en faktor.

$\beta_{fg} x_i$: Hældning som afhænger af niveauerne af to faktorer.

...

**Terminologi.** Hvis der er mange regressionsvariable og kun få faktorer, kaldes modellen ofte en **multipel regressionsmodel**. Hvis der kun er faktorer kalder man den en **variansanalysemodel**. Hvis en variansanalysemodel suppleres med nogle få regressionsvariable taler man undertiden om **kovariansanalyse.**

**På matrixform** kan alle disse modeller skrives

$$\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix} = \begin{bmatrix}x_{11} & x_{12} & \ldots & x_{1p}\\ x_{21} & x_{22} & \ldots & x_{2p}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \ldots & x_{np}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{bmatrix}$$

eller, med betegnelsen $\mu = (\mu_1, \ldots, \mu_n) \in \mathbf{R}^n$ for **middelværdivektoren** og betegnelsen $\beta = (\beta_1, \ldots, \beta_p) \in \mathbf{R}^p$ for **parametervektoren**,

$$\mu = X\beta.$$

Matricen $X$, kaldet **modelmatricen**, kan i praksis dannes på følgende måde ud fra det lineære udtryk som definerer modellen: Hvert led i det lineære udtryk genererer et antal søjler i modelmatricen. Et konstantled $\gamma$ omsættes til en søjle af ettaller. Et led af formen $\beta x_i$ omsættes til en søjle der indeholder værdierne $x_1, \ldots, x_n$. Kun hvis der optræder faktorniveauer som fodtegn genereres mere end én søjle. Antallet af søjler som genereres af et led er produktet af niveauantallene for de faktorer der indgår. Et led af formen $\alpha_f$ genererer lige så mange søjler som der er niveauer af faktoren $F$, og disse søjler bliver "dummy'erne" (eller indikatorerne) for grupperne i den tilsvarende klassifikation. Tilsvarende vil et led af formen $\alpha_{fg}$ generere en søjle for hvert par $(f, g)$ af niveauer for de to faktorer, og disse bliver "dummy'erne" hørende til den tilsvarende krydsklassifikation. Osv. osv. Blandede led af typen $\beta_f x_i$, $\beta_{fg} x_i$, $\ldots$ genererer i første omgang de samme søjler som led af formen $\alpha_f$, $\alpha_{fg}$, $\ldots$, men disse søjler skal så til sidst ganges plads for plads med værdierne $x_i$.

**ML–estimatoren** for parametervektoren $\beta$ er løsning til **normalligningerne**

$$X'y = X'X\beta.$$

Denne løsning er dog kun entydig hvis parametriseringen af middelværdivektoren $\mu$ ved $\beta$ er injektiv — dvs. den lineære afbildning $X : \mathbf{R}^p \to \mathbf{R}^n$ er injektiv, eller matricen $X$ har lineært uafhængige søjler. I så fald er $p \times p$–matricen $X'X$ regulær (faktisk positivt definit), og ML–estimatoren dermed entydigt bestemt som

$$\hat{\beta} = \left(X'X\right)^{-1} X'y.$$

Lad

$$\mathrm{SSD}_{\mathrm{res}} = \sum (y_i - \hat{\mu}_i)^2 = \|y - \hat{\mu}\|^2 = \|y - Py\|^2$$

betegne **residualkvadratsummen**. her er $P$ ortogonalprojektionen på **middelværdiunderrummet** $L = \{X\beta \mid \beta \in \mathbf{R}^p\}$. ML–estimatoren for variansen er $\mathrm{SSD}_{\mathrm{res}}/n$. Som estimator for variansen benyttes imidlertid altid den **korrigerede ML–estimator**

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}}{n-p}.$$

En begrundelse for dette er, at denne estimator er central, hvilket følger af at $\mathrm{SSD}_{\mathrm{res}}$ er $\chi^2$–fordelt med $n - p$ frihedsgrader og skalaparameter

$\sigma^2$. Fordelingen af $\hat{\beta}$ er den $p$–dimensionale normalfordeling med middelværdi $\beta$ (så $\hat{\beta}$ er også central) og kovariansmatrix

$$\operatorname{cov}(\hat{\beta}) = \sigma^2 \left(X'X\right)^{-1}.$$

**F–test for modelreduktion.** Betragt den lineære model givet ved modelmatricen $X$ hvis søjler udspænder middelværdiunderrummet $L$ af dimension $p$, samt en reduceret model givet ved modelmatricen $X_0$ som på samme måde definerer et underrum $L_0 \subset L$ af dimension $p_0 < p$. Test for hypotesen $\mu \in L_0$ kan da foretages ved hjælp af størrelsen

$$F = \frac{(\mathrm{SSD}^0_{\mathrm{res}} - \mathrm{SSD}_{\mathrm{res}})/(p - p_0)}{\mathrm{SSD}_{\mathrm{res}}/(n-p)}$$

som under hypotesen er F–fordelt med $(p - p_0, n - p)$ frihedsgrader. Her betegner $\mathrm{SSD}^0_{\mathrm{res}}$ naturligvis residualkvadratsummen i den reducerede model.

**T–test for modelreduktion.** I specialtilfældet $p_0 = p - 1$ kan hypotesen $\mu = X\beta \in L_0$ altid skrives på formen $\vartheta = 0$ for en passende lineær parameterfunktion $\vartheta = a_1\beta_1 + \cdots + a_p\beta_p$. Ovennævnte F–test er i så fald ækvivalent med et T–test, der kan begrundes og udregnes på følgende måde. Variansen på $\hat{\vartheta} = a_1\hat{\beta}_1 + \cdots + a_p\hat{\beta}_p$ kan, ved hjælp af formlen for $\operatorname{cov}(\hat{\beta})$, udregnes som

$$\operatorname{var}(\hat{\vartheta}) = \operatorname{var}(a_1\hat{\beta}_1 + \cdots + a_p\hat{\beta}_p) = a'\left(\sigma^2(X'X)^{-1}\right)a = c\sigma^2,$$

hvor $c$ er en konstant. Ved i udtrykket

$$U = \frac{\hat{\vartheta}}{\sqrt{c\sigma^2}}$$

(som er normeret normalfordelt under hypotesen, og en naturlig teststørrelse for hypotesen $\vartheta = 0$ i en model med kendt varians) at erstatte den ukendte varians $\sigma^2$ med $\hat{\sigma}^2$ får vi teststørrelsen

$$T = \frac{\hat{\vartheta}}{\sqrt{c\hat{\sigma}^2}}$$

som under hypotesen er T–fordelt med $n - p$ frihedsgrader. Der gælder så $F = T^2$, og i overensstemmelse hermed kan testet for modelreduktion foretages ved tosidet vurdering af $T$ i sin fordeling under hypotesen.

**Udtynding til lineær uafhængighed** af modelmatricens søjler (med henblik på at opnå en injektiv parametrisering) foregår normalt ved hjælp af følgende standard algoritme: Matricens søjler gennemgås en for en. Hver gang man møder en søjle, der kan skrives som linearkombination af de foregående, slettes den pågældende søjle (eller den tilsvarende parameter sættes = 0).

**Modelformler** er omkodninger af sædvanlige specifikationer af lineære modeller ved angivelse af et udtryk for middelværdien (med græske bogstaver som navne for parametre, faktorniveauer som fodtegn osv.) til en form, der er bedre egnet som input til et computerprogram, og som undgår at give parametrene irrelevante navne. Konventionerne i den forbindelse vil ikke blive gentaget her, se noterne side 110–111.

## Kapitel 11: Successiv testning

**Variansanalyseskemaet.** I de fleste statistikpakker indeholder output fra fit af en lineær model som standard et såkaldt variansanalyseskema. Det gælder f.eks. SAS (the type I ANOVA table) og ISUW. Dette skema indeholder blandt andet F–teststørrelser og P–værdier for alle de modelreduktioner, som man får ved successivt at fjerne leddene fra den oprindeligt specificerede modelformel, begyndende med det sidste. I ISUW versionen (som matematisk er den enkleste) ser variansanalyseskemaet sådan ud:

| `Effect` | `S.S.` | `d.f.` | `M.S.` | `F` |
|---|---|---|---|---|
| `led(1)` | $\mathrm{SSD}^0_{\mathrm{res}} - \mathrm{SSD}^1_{\mathrm{res}}$ | $f_0 - f_1$ | $\frac{\mathrm{SSD}^0_{\mathrm{res}} - \mathrm{SSD}^1_{\mathrm{res}}}{f_0 - f_1}$ | $F_1$ |
| $\vdots$ | | | $\vdots$ | |
| `led(j)` | $\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}$ | $f_{j-1} - f_j$ | $\frac{\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}}{f_{j-1} - f_j}$ | $F_j$ |
| $\vdots$ | | | $\vdots$ | |
| `led(k)` | $\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}$ | $f_{k-1} - f_k$ | $\frac{\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}}{f_{k-1} - f_k}$ | $F_k$ |
| `Residual` | $\mathrm{SSD}^k_{\mathrm{res}}$ | $f_k$ | $\hat{\sigma}^2$ | |
| `Total` | $\sum y_i^2$ | $n$ | $\frac{\sum y_i^2}{n}$ | |

Betydningen af de enkelte oplysninger er følgende. Modelformlen antages at bestå af $k$ led, som her (i den rækkefølge de forekommer i modelformlen) benævnes `led(1)` ... `led(k)`. Hver linie i tabellen, bortset fra overskriftslinien og de to sidste linier, vedrører en modelreduktion. Oplysningerne i den linie, som i første søjle `Effect` indeholder navnet på $j$'te led, vedrører den modelreduktion der svarer til fjernelse af leddet `led(j)`. I forbindelse med successiv testning skal tabellen altså læses nedefra.

I søjlen `S.S.` angives den ændring af residualkvadratsummen, som den pågældende modelreduktion fører til, og i søjlen `d.f.` står den tilsvarende ændring af modellens dimension. I søjlen `M.S.` angives forholdet mellem tallene i de to foregående søjler, som jo netop er tælleren i den

F–størrelse man skal benytte. I søjlen `F` står selve F–størrelsen, og i sidste søjle `P` — som her er udeladt af pladshensyn — finder man den tilhørende P–værdi.

Tabellen afsluttes naturligt ved i næstsidste linie `Residual` at angive residualkvadratsummen og dens frihedsgradsantal (antal observationer minus antal parametre i modellen) og forholdet mellem disse, som jo netop er variansestimatet i startmodellen. Det betyder, at summen af alle tallene i søjlen `S.S.` netop bliver den totale kvadratsum af observationerne, medens summen af frihedsgradsantallen i søjlen `d.f.` netop er det samlede antal observationer $n$.

To vigtige varianter er omtalt i noterne. For det første har man traditionelt opfattet et konstantled som noget, der er obligatorisk som første led i enhver model. I SAS specificerer man slet ikke konstantleddet, idet det underforstås at være til stede og stå før alle andre led i modelformlen. Under denne konvention er det naturligt at udelade første linie i variansanalysetabellen (den der hører til `led(1)`, som altså nu er konstantleddet, "`1`" i ISU jargon). Derved bliver summerne i sidste linie i stedet til $\mathrm{SSD}_y = \sum(y_i - \bar{y})^2$ og $n-1$. Den anden variant, som er noget specielt for SAS, vedrører alene de to sidste søjler. Her har man i SAS valgt at benytte $\hat{\sigma}^2$ fra startmodellen som nævner i *alle* F–størrelser, i stedet for at opdatere variansestimatet hver gang en modelreduktion godkendes. Det er sådan set en fejl, i forhold til hvad almindelige principper for successiv testning foreskriver. Men de udregnede størrelser bliver stadig F–fordelte (bare med et andet frihedsgradsantal for nævneren), og man kan argumentere, for at dette i de fleste tilfælde ikke gør så stor forskel.

**Tosidet variansanalyse i det balancerede tilfælde.** Hovedresultatet er her, at der i et tosidet $R \times S$ skema med lige mange observationer per celle gælder, at de fittede værdier under additivitetsmodellen "`R+S`" er givet ved

$$\hat{\mu}_{rsi} = \bar{y}_{r\cdot\cdot} + \bar{y}_{\cdot s\cdot} - \bar{y}_{\cdot\cdot\cdot} \quad ,$$

samt at der mellem residualkvadratsummerne i nogle af modellerne knyttet til dette skema gælder

$$\mathrm{SSD}_{\mathrm{res}}(\texttt{R+S}) = \mathrm{SSD}_{\mathrm{res}}(\texttt{R}) + \mathrm{SSD}_{\mathrm{res}}(\texttt{S}) - \mathrm{SSD}_{\mathrm{res}}(\texttt{1})$$

hvor betegnelsen $\mathrm{SSD}_{\mathrm{res}}(\ldots)$ her anvendes for residualkvadratsummen i modellen bestemt ved modelformlen $\ldots$. Da modelformlerne `R` og `S` svarer til ensidede variansanalysemodeller, medens `1` jo svarer til modellen for identisk fordelte observationer (kap. 7), betyder dette, at alle udregninger i den additive model kan foretages ret enkelt ved hjælp af en lommeregner, altså uden inversion af større matricer. Dette resultats beregningsmæssige betydning var noget større i tidligere tider end det er idag. Men det indebærer også, at de `S.S.`'er som optræder i variansanalyseskemaet bliver de samme uanset hvilken rækkefølge man vælger at fjerne de to hovedvirkninger i. Samt yderligere, at hvis SAS's konvention vedrørende nævnere i F–teststørrelser følges, så bliver selve testene for fjernelse af hovedvirkninger og deres P–værdier også uafhængige af rækkefølgen.

## Kapitel 12: Lineære normalfordelingsmodeller i praksis

Kun ganske få resultater fra dette kapitel egner sig til opsummering:

**Forklaringsgraden** for en given lineær normalfordelingsmodel, er størrelsen

$$R^2 = \frac{\mathrm{SSD}_y - \mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_y} = 1 - \frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_y}\,.$$

Idet $\frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_y}$ naturligt fortolkes som den andel af den samlede variation $\mathrm{SSD}_y$ der *ikke* kan forklares af modellen, fortolkes $R^2$ som den andel af variationen der forklares af modellen. En stor værdi af $R^2$ betyder således at man har en god model. Men her skal man naturligvis være opmærksom på, at udvidelse af en model med en ekstra regressionsvariabel eller faktor altid vil gøre $R^2$ større, uanset om dette nye led overhovedet er relevant, endsige signifikant.

**Residualplottet** benyttes til modelkontrol i multiple regressionsmodeller. Punkterne $(\hat{\mu}_i, y_i - \hat{\mu}_i)$ = "(fitted,residual)" afsættes i et koordinatsystem. Hvis punktskyen har tendens til at ligne en liggende trompet (større lodret spredning i den ene side end i den anden), tyder det på, at en model hvor variansen er en voksende eller aftagende funktion af middelværdien ville være mere relevant. Hvis skyen har antydning af parabelform kan det være, at nogle af de vigtigere forklarende variable burde transformeres.

**Modelkontrol i øvrigt.** En simpel form for modelkontrol består i at danne en passende *udvidet* model og foretage F–testet for reduktion til den oprindelige model. I en simpel regressionsanalysemodel, hvor hver værdi af den forklarende variable $x$ forekommer flere gange, kan man for eksempel som udvidelse vælge den ensidede variansanalysemodel, hvor grupperne er bestemt ved de mulige værdier af $x$. Hvis det sædvanlige $x$–$y$ plot tyder på, at afhængigheden mere ligner en krum kurve end en linie, kan man udvide den simple regressionsmodel $\mu_i = \alpha + \beta x_i$ til en model der beskriver afhængigheden ved et andengradspolynomium:

$$\mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2\,.$$

Tilsvarende metoder kan selvfølgelig bruges for andre modeller. For variansanalysemodeller med mange faktorer vil man typisk foretage modelkontrol ved at prøve at tilføje vekselvirkninger af højere ordener og se om de kan fjernes igen.

Hvis man vil undersøge antagelsen om *normalitet* kan man tegne et histogram (eller et probitdiagram) som viser residualernes fordeling (bemærk: *residualerne*, ikke selve observationerne). Dette fungerer dog kun når de fittede værdier kan antages at ligge rimeligt tæt på de sande middelværdier, dvs. typisk når $n$ er stor og $p$ er lille.

**Multikollinearitet** er det fænomen, at der i en multipel regressionsmodel forekommer to eller flere forklarende variable, som approksimativt er linearkombinationer af hinanden, således at de til en hvis grad kan erstatte hinanden i modellen. Dette kan i ondartede tilfælde gøre det vanskeligt at tage stilling til hvilke modelreduktioner man skal foretage. For to forklarende variable kan det for eksempel forekomme, at de hver for sig er insignifikante (dvs. testene for fjernelse af den enkelte variable fører til godkendelse), men så snart man har fjernet én af dem bliver den anden ekstremt signifikant.

## Kapitel 13: Generaliserede lineære modeller

En løs definition af den generaliserede lineære model går ud på, at den er givet på samme måde som en lineær normalfordelingsmodel, blot med følgende modifikationer:

(1) Fordelingerne af observationerne kan være af andre typer end den normale.

(2) Udtrykket for den enkelte observations middelværdi $\mu_i$ er ikke i sig selv lineært, men har formen $m(\eta_i)$ hvor $m$ er en kendt monoton funktion, kaldet **middelværdifunktionen**, og $\eta_i$ er givet som et lineært udtryk. Anderledes sagt, det er ikke middelværdien selv men $\eta_i = m^{-1}(\mu_i)$ der er lineær i regressionsvariable, faktorer osv. på sædvanlig måde.

Funktionen $m^{-1}$ kaldes i denne forbindelse **linkfunktionen**.

Følgende to vigtige specialtilfælde er behandlet:

**Log–lineære modeller** eller **multiplikative Poissonmodeller.** Disse modeller fås, når observationerne antages at være uafhængige, Poissonfordelte, med middelværdi specificeret som i en generaliseret lineær model med eksponentialfunktionen som middelværdifunktion og (dermed) den naturlige logaritme som link–funktion. Altså

$$\mathrm{E}y_i = \mu_i = \exp(\eta_i)$$

hvor $\eta_i$ er et sædvanligt lineært udtryk. Disse modeller omfatter alle de modeller for Poissonfordelte variable som er gennemgået i noterne, og også (ved argumenter, som involverer betingning med diverse totalsummer) alle modeller for polynomialfordelte variable som er gennemgået i noterne. Samt mange flere, deriblandt en lang række modeller til analyse af flersidede antalstabeller.

Et nyt begreb, som er nødvendigt i forbindelse med specifikationen af modellen "proportionalitet med en given baggrundsvariabel" som en generaliseret lineær model, er begrebet **afsæt.** En afsætvariabel i en model er et led $a_i$ der adderes til det lineære udtryk på forhånd. Altså, om man vil, en regressionsvariabel hvis koefficient på forhånd er sat til værdien 1.

**Logit–lineære modeller**, også kaldet **logistiske regressionsmodeller** er generaliserede lineære modeller for observationer, hvis fordelinger kan karakteriseres som fordelinger af relative hyppigheder $y_i = y_i'/m_i$, hvor $y_i'$ er binomialfordelt med sandsynlighedsparameter $\mu_i = \mathrm{E}y_i$ og antalsparameter $m_i$. Middelværdifunktionen (som i dette tilfælde afbilder hele aksen ind i det åbne enhedsinterval) og dens inverse antages givet ved

$$\mu = \frac{\exp(\eta)}{1+\exp(\eta)}\ , \qquad \eta = \operatorname{logit}(\mu) = \log\frac{\mu}{1-\mu}\ .$$

Denne klasse af modeller omfatter alle de modeller for binomialfordelte observationer, som er gennemgået i noterne (og mange flere).

En variant, som er omtalt i noterne, går ud på at erstatte middelværdifunktionen med $m = \Phi$ (den normale fordelings fordelingsfunktion) og dermed logit med $\operatorname{probit}(\mu) = \Phi^{-1}(\eta)$. Disse modeller kaldes de **probit–lineære modeller**. Begrebsmæssigt kan probit–modellerne være lidt nemmere at forstå, fordi de kan forklares som fremkommet ud fra en lineær normalfordelingsmodel ved "dikotomisering" af responserne, dvs. ved at de oprindelige (normalfordelte) responser er reduceret til binære responser, som kun indeholder oplysningen om hvorvidt den oprindelige respons var større (1) eller mindre (0) end en vis tærskelværdi. I praksis er der dog ikke stor forskel på logit– og probit–modeller.

**Listning af parameterestimater.** Dette er en kommentar der vedrører statistikpakkers håndtering af lineære modeller (kap. 10–12) og generaliserede lineære modeller (kap. 13). En standard facilitet, som følger en hvilken som helst statistikpakke, går ud på, at man ved fit af en lineær eller generaliseret lineær model kan få parameterestimaterne skrevet ud. Typisk vil statistikpakken før udregning af estimatorerne have foretaget en udtynding af modelmatricen til lineær uafhængighed. Man kan eventuelt styre dette ved at vælge hvilke (overflødige) led man vil tage med i modelformlen, bytte om på faktorniveauer eller generere sine dummier selv. Under alle omstændigheder er det nødvendigt for fortolkningen af parameterestimaterne at vide præcis hvordan modellen er parametriseret. En meget almindelig standard facilitet går ud på at tilføje tre søjler til listen, som indeholder en standardafvigelse for det angivne estimat, en U– eller T–størrelse der er udregnet som estimatet divideret med standardafvigelsen, og en tilhørende P–værdi. For de lineære normalfordelingsmodeller er standardafvigelsen udregnet som forklaret i kapitel 10, idet $\hat{\sigma}^2$ er indsat for $\sigma^2$, hvorved de beregnede T–størrelser og tilhørende (tosidede) P–værdier svarer til eksakte tests for, om den enkelte parameter kan sættes lig med 0. Fortolkningen for de generaliserede lineære modeller er den samme, blot er der her tale om approksimative standardafvigelser, og det test der udføres er baseret på approksimativ normalitet af parameterestimaterne. I begge tilfælde skal man være opmærksom på, at mange af disse tests er fuldstændigt irrelevante.