

STATISTISK DATAANALYSE MED PYTHON

Formelsamling

2025 edition

Forord

Denne formelsamling er lavet til at give overblik over vigtige statistiske modeller og formler til kurset *Statistisk Dataanalyse med Python 2025* ved Aarhus Universitet. For en mere detaljeret beskrivelse se **webbogen**. Kontakt 202204949@post.au.dk for forslag eller rettelser.

Contents

1 Cheat sheet	3
1.1 Generelt	3
1.2 Notation	3
1.3 Binomialmodel	3
1.4 Poissonmodel	3
1.5 Multinomialmodel	3
2 Binomialmodellen	4
3 Poissonmodellen	4
4 Multinomialmodeller	5
4.1 Multinomialmodellen	5
4.2 Flere multinomialmodeller (homogenitetstest)	5

5 Normalfordelingsmodeller	6
5.1 Et normalfordelt datasæt	6
5.2 To normalfordelte datasæt	7
5.2.1 Test for ens varians	9
5.2.2 Fra log til ikke-log	9
6 Ophobningsloven	9
7 Lineær regression	10
7.1 Notation i kurset	10
7.2 Den lineære regressionsmodel	10
7.3 Linjens værdi i et punkt	12
7.3.1 Invers regression	12
8 Modeller til variansanalyse	12
8.1 Bartlett's test for ens varians	13
8.2 Oneway ANOVA	13
8.3 Twoway ANOVA	14
8.4 Gruppesspecifik regression	15
9 Multipel regression	15
9.1 Cross-validation fremgangsmåde	16
9.2 Backward selection	16
9.3 Forward selection	17
9.4 Ridge regression	17

1 Cheat sheet

1.1 Generelt

Signifikansniveau: $\alpha = 0.05$

Hypotese: $H : \theta = \theta_0$ mod $H_a : \theta \neq \theta_0$

Beslutningsregel: Afvis H hvis p -værdi $\leq \alpha$

1.2 Notation

Observationer: x_1, \dots, x_n

Antal observationer: n

Antal frihedsgrader: df

Parameterskøn: $\hat{\theta}$

1.3 Binomialmodel

Model: $X \sim \text{binom}(n, p)$

Sandsynlighedsparameter: $\hat{p} = \frac{x}{n}$

Hypotese: $H : p = p_0$ mod $H_a : p \neq p_0$

1.4 Poissonmodel

Model: $X_i \sim \text{poisson}(t\lambda)$

Rateparameter: $\hat{\lambda} = \frac{x_1 + \dots + x_n}{nt}$

1.5 Multinomialmodel

Model: $(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k))$

Sandsynligheder: $\hat{\pi}_j = \frac{a_j}{n}$

Middelværdi: $\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$

Empirisk varians: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Empirisk spredning: $s = \sqrt{s^2}$

Standard error: $\text{std}_s(\hat{\mu}) = \frac{s}{\sqrt{n}}$

2 Binomialmodellen

Lad den stokastiske variabel X angive antal ud af n delforsøg med et bestemt udfald, hvor der er sandsynlighed p for dette udfald. Binomialmodellen skrives:

$$X \sim \text{binom}(n, p), \quad 0 \leq p \leq 1. \quad (1)$$

Skøn over p er givet ved:

$$\hat{p} = \frac{x}{n}, \quad (2)$$

p -værdien for test af hypotesen $p = p_0$ mod alternativet $p \neq p_0$ er givet ved:

$$p\text{-værdi} = P(X \leq np_0 - t) + P(X \geq np_0 + t), \quad (3)$$

hvor $t = |x - np_0|$ og x er det observerede antal med det bestemte udfald. Approximativt 95%-konfidensinterval for p er givet ved:

$$\left[\frac{x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}}{n + u^2}, \frac{x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}}{n + u^2} \right], \quad u = 1.96. \quad (4)$$

3 Poissonmodellen

Lad den stokastiske variabel X angive antal hændelser beskrevet ved rateparameter λ . Poissonmodellen skrives:

$$X \sim \text{poisson}(\lambda), \quad \lambda > 0. \quad (5)$$

Ved n uafhænige eksperimenter og en kendt eksponering t (ofte tid) kan modellen skrives:

$$X_i \sim \text{poisson}(t\lambda), \quad i = 1, \dots, n, \quad \lambda > 0. \quad (6)$$

Skøn over λ er givet ved:

$$\hat{\lambda} = \frac{x_1 + \dots + x_n}{nt}, \quad (7)$$

Et approximativt 95%-konfidensinterval for λ er givet ved:

$$\left[\frac{1}{nt} \left(x + \frac{u^2}{2} - u\sqrt{x + \frac{u^2}{4}} \right), \frac{1}{nt} \left(x + \frac{u^2}{2} + u\sqrt{x + \frac{u^2}{4}} \right) \right], \quad u = 1.96, \quad (8)$$

hvor $x = x_1 + \dots + x_n$.

4 Multinomialmodeller

4.1 Multinomialmodellen

Lad der være k mulige udfald i et forsøg, med sandsynligheder π_1, \dots, π_k , hvor $\pi_1 + \dots + \pi_k = 1$. De stokastiske variable A_1, \dots, A_k angiver antal af hvert udfald i n gentagelser af forsøget. Multinomialmodellen skrives:

$$(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k)), \quad \pi_j \geq 0, \pi_1 + \dots + \pi_k = 1. \quad (9)$$

Skøn over π_j er givet ved:

$$\hat{\pi}_j = \frac{a_j}{n}, \quad j = 1, \dots, k. \quad (10)$$

p -værdien for test af hypotesen $\pi_j = p_j(\theta)$ beregnes ved et G -test:

$$p\text{-værdi} = P(G \geq G_{\text{obs}}) = 1 - \chi^2_{\text{cdf}}(G_{\text{obs}}, df), \quad (11)$$

hvor G -teststørrelsen er givet ved:

$$G_{\text{obs}} = 2 \sum_{j=1}^k a_j \log \left(\frac{a_j}{e_j} \right), \quad (12)$$

$e_j = np_j(\hat{\theta})$ er det forventede antal i udfald j , og df er frihedsgraderne. Er p_j ikke afhængig af en parameter θ , er $df = k - 1$. Afhænger p_j af d parametre (ofte i goodness of fit test), er $df = k - 1 - d$.

OBS! For at G -testen er gyldig, skal det forventede antal e_j i hvert udfald være mindst 5, dvs. $e_j \geq 5$ for alle $j = 1, \dots, k$. Alternativt kan Cochran's regel anvendes: Højst 20% af de forventede antal må være mindre end 5, og ingen af de forventede antal må være mindre end 1.

4.2 Flere multinomialmodeller (homogenitetstest)

Betrægt r populationer, hvor den i 'te population har n_i observationer fordelt på k udfald med sandsynligheder $\pi_{i1}, \dots, \pi_{ik}$. De stokastiske variable A_{i1}, \dots, A_{ik} angiver antal af hvert udfald i population i . Modellen skrives:

$$\begin{aligned} (A_{11}, \dots, A_{1k}) &\sim \text{multinom}(n_1, (\pi_{11}, \dots, \pi_{1k})), \\ &\vdots \\ (A_{r1}, \dots, A_{rk}) &\sim \text{multinom}(n_r, (\pi_{r1}, \dots, \pi_{rk})), \\ \pi_{ij} &\geq 0, \quad \pi_{i1} + \dots + \pi_{ik} = 1, \quad i = 1, \dots, r. \end{aligned} \quad (13)$$

Skøn over π_{ij} er givet ved:

$$\hat{\pi}_{ij} = \frac{a_{ij}}{n_i}, \quad i = 1, \dots, r, \quad j = 1, \dots, k. \quad (14)$$

Til test af homogenitetshypotesen om at alle populationer har samme fordeling, dvs. $\pi_{1j} = \dots = \pi_{rj}$ for alle $j = 1, \dots, k$, kan et G -test anvendes med teststørrelsen:

$$G_{\text{obs}} = 2 \sum_{i=1}^r \sum_{j=1}^k a_{ij} \log \left(\frac{a_{ij}}{e_{ij}} \right), \quad (15)$$

hvor $e_{ij} = n_i \hat{\pi}_j$ er det forventede antal i udfald j i population i , og $\hat{\pi}_j = \frac{a_{1j} + \dots + a_{rj}}{n_1 + \dots + n_r}$ er det samlede skøn over sandsynligheden for udfald j . Frihedsgraderne er givet ved $df = (r - 1)(k - 1)$. p -værdien beregnes som i (11). **OBS!** Her gælder samme betingelser for gyldighed af G -testen som i afsnit 4.1.

5 Normalfordelingsmodeller

5.1 Et normalfordelt datasæt

Betrægt n uafhængige stokastiske variable X_1, \dots, X_n med målinger x_1, \dots, x_n , som antages at være normalfordelte med middelværdi μ og spredning σ . Modellen skrives:

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n, \quad (\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+. \quad (16)$$

Skøn over middelværdi er givet ved:

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \dots + x_n), \quad (17)$$

Idet man ikke kender hele populationen, anvendes den empiriske varians som skøn over variansen:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (18)$$

Og empirisk spredningen er givet ved:

$$s = \sqrt{s^2}. \quad (19)$$

I Python bestemmes disse ved `np.mean(data)`, `np.var(data, ddof=1)` og `np.std(data, ddof=1)`.

Yderligere har vi *standard error* for middelværdien:

$$\text{std}_s(\hat{\mu}) = \frac{s}{\sqrt{n}}. \quad (20)$$

Som løst sagt ikke er hvor meget data spredes sig men hvor meget skønnet over middelværdien spredes sig.

Til test af hypotesen $\mu = \mu_0$ mod $\mu \neq \mu_0$ beregnes *t-teststørrelsen*:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad (21)$$

Og vurderes i en $t(df)$ -fordeling med $df = n - 1$ frihedsgrader. *p*-værdien beregnes som:

$$p\text{-værdi} = 2 \cdot t_{\text{cdf}}(-|t_{\text{obs}}|, df). \quad (22)$$

Et 95%-konfidensinterval for μ er givet ved:

$$\left[\bar{x} - t_0 \frac{s}{\sqrt{n}}, \bar{x} + t_0 \frac{s}{\sqrt{n}} \right], \quad t_0 = t_{\text{ppf}}(0.975, df), \quad (23)$$

hvor $df = n - 1$. Et konfidensinterval for variansen σ^2 er givet ved:

$$\left[\frac{df \cdot s^2}{\chi_{\text{ppf}}^2(0.975, df)}, \frac{df \cdot s^2}{\chi_{\text{ppf}}^2(0.025, df)} \right], \quad df = n - 1. \quad (24)$$

Og for spredningen σ er konfidensintervallet:

$$\left[\sqrt{\frac{df \cdot s^2}{\chi_{\text{ppf}}^2(0.975, df)}}, \sqrt{\frac{df \cdot s^2}{\chi_{\text{ppf}}^2(0.025, df)}} \right], \quad df = n - 1. \quad (25)$$

5.2 To normalfordelte datasæt

Normalfordelingsmodellen for to uafhængige datasæt med forskellige varianser og n_1 og n_2 observationer skrives:

$$\begin{aligned} X_{1i} &\sim N(\mu_1, \sigma_1^2), \quad i = 1, \dots, n_1, \\ X_{2i} &\sim N(\mu_2, \sigma_2^2), \quad i = 1, \dots, n_2, \\ (\mu_1, \mu_2, \sigma_1, \sigma_2) &\in \mathbf{R}^2 \times \mathbf{R}_+^2. \end{aligned} \quad (26)$$

En undermodel med antagelsen om fælles varians skrives:

$$\begin{aligned} X_{1i} &\sim N(\mu_1, \sigma), \quad i = 1, \dots, n_1, \\ X_{2i} &\sim N(\mu_2, \sigma), \quad i = 1, \dots, n_2, \\ (\mu_1, \mu_2, \sigma) &\in \mathbf{R}^2 \times \mathbf{R}_+. \end{aligned} \quad (27)$$

I begge tilfælde er skøn over middelværdierne er givet ved:

$$\hat{\mu}_1 = \bar{x}_1 = \frac{\sum_i x_{1i}}{n_1}, \quad \hat{\mu}_2 = \bar{x}_2 = \frac{\sum_i x_{2i}}{n_2}, \quad (28)$$

og skøn over empiriske varianser er givet ved:

$$s_1^2 = \frac{\sum_i (x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum_i (x_{2i} - \bar{x}_2)^2}{n_2 - 1}. \quad (29)$$

Ved antagelse om fælles varians er skøn over den fælles varians er givet ved:

$$s_r^2 = \frac{\sum_i (x_{1i} - \bar{x}_1)^2 + \sum_i (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}. \quad (30)$$

Til test af hypotesen $\mu_1 = \mu_2$ mod $\mu_1 \neq \mu_2$ beregnes t -teststørrelsen:

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_r^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad (31)$$

Og vurderes i en $t(df)$ -fordeling. Ved antagelse om fælles varians anvendes $df = n_1 + n_2 - 2$ frihedsgrader. Ved forskellige varianser anvendes Welch's approksimation hvor frihedsgraderne er givet ved:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}. \quad (32)$$

I begge tilfælde beregnes p -værdien som:

$$p\text{-værdi} = 2 \cdot t_{\text{cdf}}(-|t_{\text{obs}}|, df). \quad (33)$$

Et 95%-konfidensinterval for forskellen i middelværdi $\mu_1 - \mu_2$ når der antages fælles varians er givet ved:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_0 \sqrt{s_r^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, (\bar{x}_1 - \bar{x}_2) + t_0 \sqrt{s_r^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right], \quad (34)$$

Og i tilfælde af forskellige varianser er konfidensintervallet givet ved:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_0 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_0 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right], \quad (35)$$

hvor $t_0 = t_{\text{ppf}}(0.975, df)$.

Man kan komme ud for at skulle beregne en effektstørrelse. Denne er givet ved:

$$\frac{|\bar{x}_1 - \bar{x}_2|}{s_r}, \quad (36)$$

og effektstørrelser større end 1 betragtes som store.

5.2.1 Test for ens varians

Til test af hypotesen $\sigma_1^2 = \sigma_2^2$ mod $\sigma_1^2 \neq \sigma_2^2$ anvendes undersøger man om forholdet s_1^2/s_2^2 afgiver signifikant fra 1. F -teststørrelsen er givet ved:

$$F_{\text{obs}} = \frac{s_1^2}{s_2^2}, \quad (37)$$

som vurderes i en $F(df_1, df_2)$ -fordeling, hvor p -værdien beregnes som:

$$p\text{-værdi} = 2 \cdot \min(F_{\text{cdf}}(F_{\text{obs}}, df_1, df_2), 1 - F_{\text{cdf}}(F_{\text{obs}}, df_1, df_2)), \quad (38)$$

hvor $df_1 = n_1 - 1$ og $df_2 = n_2 - 1$ er frihedsgraderne. Yderligere er et 95%-konfidensinterval for forholdet σ_1^2/σ_2^2 givet ved:

$$\left[\frac{F_{\text{obs}}}{F_{\text{ppf}}(0.975, df_1, df_2)}, \frac{F_{\text{obs}}}{F_{\text{ppf}}(0.025, df_1, df_2)} \right], \quad (39)$$

hvor $df_1 = n_1 - 1$ og $df_2 = n_2 - 1$. Hvis 1 er indeholdt i konfidensintervallet, kan man ikke afvise hypotesen om ens varians.

5.2.2 Fra log til ikke-log

Lad ν_1 og ν_2 være middelværdien for logaritmiske transformerede data fra to grupper, hvor der kan antages fælles varians. Forholdet mellem middelværdierne μ_1 og μ_2 i de oprindelige data er givet ved:

$$\frac{\mu_1}{\mu_2} = \exp(\nu_1 - \nu_2). \quad (40)$$

Et 95%-konfidensinterval for forholdet findes ved at tage intervaldepunkterne fra konfidensintervallet for $\nu_1 - \nu_2$, beregnet ved (34) og indsætte i (40).

6 Ophobningsloven

Betrægt en parameter som er funktion af flere variable $\theta = f(\mu_1, \dots, \mu_k)$ og skøn over disse variable $\hat{\mu}_1, \dots, \hat{\mu}_k$. Skøn over parameteren er givet ved:

$$\hat{\theta} = f(\hat{\mu}_1, \dots, \hat{\mu}_k). \quad (41)$$

Simpel notation for den partielle afledte af f med hensyn til μ_j er givet ved:

$$\frac{\partial f(\mu_1, \dots, \mu_k)}{\partial \mu_j} = f'_{\mu_j}, \quad j = 1, \dots, k, \quad (42)$$

og udregnet i skønnet noteres dette som \hat{f}'_{μ_j} . Kendes standard error for hver af de enkelte skøn, dvs. $\text{std}_s(\hat{\mu}_j)$ for $i = j, \dots, k$, kan standard error for skønnet over parameteren approksimativt beregnes ved:

$$\text{std}_s(\hat{\theta}) = \sqrt{\sum_{j=1}^k \left(\hat{f}'_{\mu_j}\right)^2 \text{std}_s(\hat{\mu}_j)^2 + 2 \sum_{j=1}^{k-1} \sum_{m=j+1}^k \hat{f}'_{\mu_j} \hat{f}'_{\mu_m} \text{Cov}_s(\hat{\mu}_j, \hat{\mu}_m)}, \quad (43)$$

Og i tilfælde af uafhængige målinger forenkles dette til:

$$\text{std}_s(\hat{\theta}) = \sqrt{\sum_{j=1}^k \left(\hat{f}'_{\mu_j}\right)^2 \text{std}_s(\hat{\mu}_j)^2}. \quad (44)$$

Et 95%-konfidensinterval for parameteren er givet ved:

$$[\hat{\theta} - 1.96 \cdot \text{std}_s(\hat{\theta}), \hat{\theta} + 1.96 \cdot \text{std}_s(\hat{\theta})]. \quad (45)$$

7 Lineær regression

7.1 Notation i kurset

Fra gymnasiet kender vi den lineære regressionsformel som:

$$y = ax + b, \quad (46)$$

hvor b er skæringspunktet med y -aksen og a er hældningskoefficienten. I statistisk dataanalyse med Python anvendes dog en anden notation:

$$x = \alpha + \beta t \quad (47)$$

hvor t er den forklarende variabel og x er responsvariablen. Her er α skæringspunktet med x -aksen og β er hældningskoefficienten. Mere præcist siger man at middelværdien af X givet i t er:

$$E(X_i) = \alpha + \beta t_i. \quad (48)$$

7.2 Den lineære regressionsmodel

Betrægt n uafhængige stokastiske variable X_1, \dots, X_n med tilhørende forklarende variable t_1, \dots, t_n .

Den lineære regressionsmodel skrives:

$$X_i \sim N(\alpha + \beta t_i, \sigma^2), \quad i = 1, \dots, n, \quad (\alpha, \beta, \sigma) \in \mathbf{R} \times \mathbf{R} \times \mathbf{R}_+. \quad (49)$$

Fra mindste kvadraters metode er skøn over parametrene givet ved:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(t_i - \bar{t})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \quad \hat{\alpha} = \bar{x} - \hat{\beta}\bar{t}, \quad (50)$$

hvor $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ og $\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$. Skøn over variansen er givet ved:

$$s_r^2 = \frac{SSD(M_r)}{df(M_r)}, \quad df(M_r) = n - 2, \quad (51)$$

hvor:

$$SSD(M_r) = \sum_{i=1}^n (x_i - (\hat{\alpha} + \hat{\beta}t_i))^2, \quad (52)$$

er summen af kvadrerede afvigelser for regressionsmodellen M_r . Standard error for skønnnet over hældningskoefficienten er givet ved:

$$\text{std}_s(\hat{\beta}) = \frac{s_r}{\sqrt{SSD_t}}, \quad SSD_t = \sum_{i=1}^n (t_i - \bar{t})^2. \quad (53)$$

og standard error for skønnnet over skæringspunktet er givet ved:

$$\text{std}_s(\hat{\alpha}) = s_r \sqrt{\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}}. \quad (54)$$

Til test af hypoteserne $\beta = \beta_0$ mod $\beta \neq \beta_0$ og $\alpha = \alpha_0$ mod $\alpha \neq \alpha_0$ beregnes t -teststørrelserne:

$$t_{\beta,\text{obs}} = \frac{\hat{\beta} - \beta_0}{\text{std}_s(\hat{\beta})}, \quad t_{\alpha,\text{obs}} = \frac{\hat{\alpha} - \alpha_0}{\text{std}_s(\hat{\alpha})}, \quad (55)$$

og vurderes i en $t(df)$ -fordeling med $df = n - 2$ frihedsgrader. I begge tilfælde beregnes p -værdierne som:

$$p\text{-værdi} = 2 \cdot t_{\text{cdf}}(-|t_{\text{obs}}|, df). \quad (56)$$

Et 95%-konfidensinterval for hældningen β er givet ved:

$$[\hat{\beta} - t_0 \cdot \text{std}_s(\hat{\beta}), \hat{\beta} + t_0 \cdot \text{std}_s(\hat{\beta})], \quad t_0 = t_{\text{ppf}}(0.975, df), \quad (57)$$

og for skæringspunktet α er konfidensintervallet:

$$[\hat{\alpha} - t_0 \cdot \text{std}_s(\hat{\alpha}), \hat{\alpha} + t_0 \cdot \text{std}_s(\hat{\alpha})], \quad t_0 = t_{\text{ppf}}(0.975, df), \quad (58)$$

hvor $df = n - 2$. Et 95%-konfidensinterval for variansen σ^2 er givet ved:

$$\left[\frac{df(M_r) \cdot s_r^2}{\chi_{\text{ppf}}^2(0.975, df(M_r))}, \frac{df(M_r) \cdot s_r^2}{\chi_{\text{ppf}}^2(0.025, df(M_r))} \right], \quad df(M_r) = n - 2, \quad (59)$$

og for spredningen σ er konfidensintervallet kvadratroden af intervallet i (59).

7.3 Linjens værdi i et punkt

For en given værdi $t = t_*$ er linjens værdi givet ved:

$$\xi_* = \hat{\alpha} + \hat{\beta}t_*. \quad (60)$$

Standard error for linjens værdi i punktet er givet ved:

$$\text{std}_s(\xi_*) = s_r \sqrt{\frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}}, \quad (61)$$

hvor SSD_t er givet ved (53). Et 95%-konfidensinterval for linjens værdi i punktet er givet ved:

$$[\xi_* - t_0 \cdot \text{std}_s(\xi_*), \xi_* + t_0 \cdot \text{std}_s(\xi_*)], \quad t_0 = t_{\text{ppf}}(0.975, df), \quad (62)$$

hvor $df = n - 2$. For en *kommende observation* i punktet $t = t_*$ er standard error givet ved:

$$\text{std}_{\text{præ}} = s_r \sqrt{1 + \frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}}, \quad (63)$$

og et 95%-prædiktionsinterval for en kommende observation i punktet er givet ved:

$$[\xi_* - t_0 \cdot \text{std}_{\text{præ}}, \xi_* + t_0 \cdot \text{std}_{\text{præ}}], \quad t_0 = t_{\text{ppf}}(0.975, df), \quad (64)$$

hvor $df = n - 2$.

7.3.1 Invers regression

Ved invers regression ønskes en værdi $t = t_*$ givet en eller flere responsværdi $x = x_*$. Udregning er ikke tydeligt gennemgået i webbogen, men i Python bestemmes t_* ved `inversReg` fra `pytFunktioner`.

8 Modeller til variansanalyse

Faktor er en forklarende variabel, som kan antage nogle niveauer. Hvor vi før har betragtet at responsvariablen X kan afhænge af én forklarende variabel t , kan vi i variansanalyse have flere faktorer.

8.1 Bartlett's test for ens varians

For at kunne udføre en variansanalyse (ANOVA), må det antages at alle grupper har samme varians. Denne hypotese $\sigma_1 = \dots = \sigma_k$ testes ved Bartlett's test. Hertil beregnes Bartlett's teststørrelse:

$$Ba = \frac{1}{C} \left(df \cdot \log s^2 - \sum_{g=1}^k dg_g \cdot \log(s_g^2) \right), \quad C = 1 + \frac{1}{3(k-1)} \left(\sum_{g=1}^k \frac{1}{df_g} - \frac{1}{df} \right), \quad (65)$$

Hvor det fælles variansområdet og totale antal frihedsgrader er givet ved:

$$s^2 = \frac{\sum_{g=1}^k df_g s_g^2}{df}, \quad df = \sum_{g=1}^k df_g, \quad df_g = n_g - 1, \quad (66)$$

og hvor s_g^2 er den empiriske varians i gruppe g med n_g observationer. Bartlett's teststørrelse vurderes i en $\chi^2(df = k-1)$ -fordeling, og p -værdien beregnes som:

$$p\text{-værdi} = 1 - \chi_{\text{cdf}}^2(Ba, df). \quad (67)$$

8.2 Oneway ANOVA

Udgangspunktet er en enkelt faktor G med k niveauer. Hver gruppe har sin egen middelværdi og varians. Den grundlæggende enkeltfaktor model skrives:

$$\begin{aligned} X_i &\sim N(\mu_{G_i}, \sigma_{G_i}^2), \quad i = 1, \dots, n, \\ (\mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k) &\in \mathbf{R}^k \times \mathbf{R}_+^k. \end{aligned} \quad (68)$$

Hvis hypotesen om ens varians ikke kan afvises, kan en variansanalyse udføres. Til dette benyttes modellen:

$$\begin{aligned} X_i &\sim N(\mu_{G_i}, \sigma^2), \quad i = 1, \dots, n, \\ (\mu_1, \dots, \mu_k, \sigma) &\in \mathbf{R}^k \times \mathbf{R}_+. \end{aligned} \quad (69)$$

Skøn over middelværdierne er givet ved:

$$\hat{\mu}_g = \bar{x}_g = \frac{1}{n_g} \sum_{i \in I_g} x_i, \quad g = 1, \dots, k, \quad (70)$$

Empirisk variansområdet indenfor grupperne i denne model er givet ved:

$$s^2(M_1) = \frac{1}{n-k} \sum_i (x_i - \bar{x}_{G_i})^2, \quad (71)$$

Undermodellen hvor hypotesen om ens middelværdi $\mu_1 = \dots = \mu_k$ gælder svarer til den simple normalfordelingsmodel i (16) benævnt M_2 . Til test af hypotesen om ens middelværdi anvendes variansskønnet indenfor grupperne i (71) og variansskønnet mellem grupperne:

$$s^2(M_1, M_2) = \frac{1}{k-1} \sum_{g=1}^k n_g (\bar{x}_g - \bar{x})^2, \quad (72)$$

Test af hypotesen udføres ved F -teststørrelsen:

$$F_{\text{obs}} = \frac{s^2(M_1, M_2)}{s^2(M_1)}, \quad (73)$$

som vurderes i en $F(df_1, df_2)$ -fordeling, hvor $df_1 = k - 1$ og $df_2 = n - k$ er frihedsgraderne. p -værdien beregnes som:

$$p\text{-værdi} = 1 - F_{\text{cdf}}(F_{\text{obs}}, df_1, df_2). \quad (74)$$

8.3 Twoway ANOVA

I denne model betragtes en faktor G og en faktor H , som hver især kan antage henholdsvis k og m niveauer. Den grundlæggende to-faktor model skrives:

$$\begin{aligned} X_i &\sim N(\mu_{G_i, H_i}, \sigma_{G_i, H_i}^2), \quad i = 1, \dots, n, \\ (\mu_{11}, \dots, \mu_{km}, \sigma_{11}, \dots, \sigma_{km}) &\in \mathbf{R}^{k \cdot m} \times \mathbf{R}_+^{k \cdot m}. \end{aligned} \quad (75)$$

Der ønskes at teste for ens varians på tværs af grupperne. Dette gøres ved Bartlett's test som beskrevet tidligere. I tilfælde af ens varians kan modellen for to-faktor ANOVA skrives:

$$\begin{aligned} X_i &\sim N(\mu_{G_i, H_i}, \sigma^2), \quad i = 1, \dots, n, \\ (\mu_{11}, \dots, \mu_{km}, \sigma) &\in \mathbf{R}^{k \cdot m} \times \mathbf{R}_+. \end{aligned} \quad (76)$$

En undermodel til to-faktor ANOVA er den additive model, hvor der ikke er nogen interaktion mellem faktorerne, således middelværdien kan udtrykkes som $\mu_{G_i, H_i} = \zeta_{G_i} + \eta_{H_i}$. Denne model skrives:

$$\begin{aligned} X_i &\sim N(\zeta_{G_i} + \eta_{H_i}, \sigma^2), \quad i = 1, \dots, n, \\ (\zeta_1, \dots, \zeta_k, \eta_1, \dots, \eta_m, \sigma) &\in \mathbf{R}^{k+m} \times \mathbf{R}_+. \end{aligned} \quad (77)$$

Man kan under den additive model teste for om en faktor har en effekt på responsvariablen ved at teste hypotesen $\zeta_1 = \dots = \zeta_k$ mod at mindst én af disse middelværdier er forskellig.

8.4 Gruppespecifik regression

For n uafhængige stokastiske variable X_1, \dots, X_n med tilhørende forklarende variable t_1, \dots, t_n og en faktor G med k niveauer, kan en gruppespecifik regressionsmodel med egen varians skrives:

$$X_i \sim N(\alpha_{G_i} + \beta_{G_i} t_i, \sigma_{G_i}^2), \quad i = 1, \dots, n, \quad (78)$$

$$(\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \sigma_1, \dots, \sigma_k) \in \mathbf{R}^{2k} \times \mathbf{R}_+^k.$$

Bartlett's test kan anvendes til at teste for ens varians på tværs af grupperne, ved at lave regressionsanalyse i hver gruppe. I tilfælde af ens varians kan modellen skrives:

$$X_i \sim N(\alpha_{G_i} + \beta_{G_i} t_i, \sigma^2), \quad i = 1, \dots, n, \quad (79)$$

$$(\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \sigma) \in \mathbf{R}^{2k} \times \mathbf{R}_+.$$

Herunder tester man for om der er forskel på hældningskoefficienterne på tværs af grupperne ved at teste hypotesen $\beta_1 = \dots = \beta_k$. En undermodel til gruppespecifik regression er den additive model, hvor alle grupper har samme hældningskoefficient:

$$X_i \sim N(\alpha_{G_i} + \beta t_i, \sigma^2), \quad i = 1, \dots, n, \quad (80)$$

$$(\alpha_1, \dots, \alpha_k, \beta, \sigma) \in \mathbf{R}^{k+1} \times \mathbf{R}_+.$$

Man kan ligeledes teste for om der er forskel på skæringspunkterne på tværs af grupperne ved at teste hypotesen $\alpha_1 = \dots = \alpha_k$. Hvis begge hypoteser om ens hældningskoefficienter og ens skæringspunkter accepteres, reduceres modellen til den simple lineære regressionsmodel i (49).

9 Multipel regression

For n uafhængige stokastiske variable X_1, \dots, X_n , som er afhængige af d forklarende variable t_{i1}, \dots, t_{id} , kan den multiple regressionsmodel skrives:

$$X_i \sim N(\alpha + \beta_1 t_{i1} + \dots + \beta_d t_{id}, \sigma^2), \quad i = 1, \dots, n, \quad (81)$$

$$(\alpha, \beta_1, \dots, \beta_d, \sigma) \in \mathbf{R}^{d+1} \times \mathbf{R}_+.$$

Man prøver altså at modellere hvordan flere parametre tilsammen påvirker en respons. Skøn over den i 'te forventede værdi er givet ved:

$$\hat{\xi}_i = \hat{\alpha} + \hat{\beta}_1 t_{i1} + \dots + \hat{\beta}_d t_{id}, \quad (82)$$

Og variansskønnet er givet ved:

$$s^2 = \frac{SSD(M)}{df(M)}, \quad SSD(M) = \sum_i (x_i - \hat{\xi}_i)^2 \quad df(M_r) = n - (d + 1), \quad (83)$$

Og prædikterede værdi samt prædiktionsinterval for en kommende observation kan beregnes som i den simple lineære regression.

Det kan være vanskeligt at vurdere hvilke koefficienter der rent faktisk er vigtige i modellen. Til dette introduceres der 3 simple algoritmer: *backward selection*, *forward selection* og en maskinlæringsalgoritme kaldet *ridge regression*.

9.1 Cross-validation fremgangsmåde

Specifikt til forward selection og ridge regression bruges ikke p -værdier til at vurdere vigtigheden af koefficienter, men derimod *cross-validation* (CV). Her deles datasæt tilfældigt op i et træningssæt med datapunkter $(t_{i1}, \dots, t_{id}, x_i)$ hvor $i = 1, \dots, n$ og et testsæt med datapunkter $(\tilde{t}_{i1}, \dots, \tilde{t}_{id}, \tilde{x}_i)$ hvor $i = 1, \dots, m$. Fremgangsmåden er som følger:

- Træningssættet bruges til at estimere koefficienterne $\hat{\beta}_1, \dots, \hat{\beta}_d$.
- Beregn en prædikteret værdi ved $\hat{\xi}_i^P = \hat{\alpha} + \hat{\beta}_1 \tilde{t}_{i1} + \dots + \hat{\beta}_d \tilde{t}_{id}$ for hver observation i testsættet.
- Beregn prædiktionsfejlen $\tilde{x}_i - \hat{\xi}_i^P$ for hver observation i testsættet.
- Beregn prædiktionsspredningen $s_{cv}^2 = \sqrt{\frac{1}{m} \sum_{i=1}^m (\tilde{x}_i - \hat{\xi}_i^P)^2}$.

I dette kursus anvendes ét datapunkt i testsættet ad gangen, hvilket kaldes *leave-one-out cross-validation* (LOOCV). Her er $m = 1$ og prædiktionsspredningen forenkles til:

$$s_{cv} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\xi}_i^{(-i)})^2}, \quad (84)$$

og gentages for alle datapunkter i datasættet.

9.2 Backward selection

Algoritmen starter med at inkludere alle forklarende variable i modellen.

- Beregn p -værdier for alle forklarende variable i modellen.

- Fjern den forklarende variabel med den højeste p -værdi, hvis denne er større end et forudbestemt signifikansniveau (ofte 0.05).
- Gentag processen ved at fjerne en forklarende variabel ad gangen.

Ofte afsluttes der med et F -test for reduktion fra den fulde model til den endelige model.

9.3 Forward selection

Algoritmen starter med en model uden forklarende variable.

- Beregn spredningsskøn for hver enkelt forklarende variabel, hvis denne tilføjes i modellen.
- Vælg den forklarende variabel som giver lavest spredningsskøn og tilføj denne til modellen.
- Gentag processen ved at tilføje en forklarende variabel ad gangen.
- For hvert af ovenstående trin, beregnes prædiktionsspredningen ved LOOCV. Den model som giver lavest prædiktionsspredning vælges som den endelige model.

9.4 Ridge regression

I ridge regression tilføjes en reguleringsparameter λ . Store værdier af λ medfører at koefficienterne skubbes mod 0, hvilket mindsker risikoen for overfitting, men kan også medføre at vigtige forklarende variable fjernes fra modellen. Der ønskes at vælge en λ således s_{cv} minimeres. Dette gøres ved at prøve forskellige værdier af λ og beregne prædiktionsspredningen ved LOOCV for hver værdi. Den værdi af λ som giver lavest prædiktionsspredning vælges som den endelige model. Man kan godt vælge en lidt større λ hvis prædiktionsspredningen kun stiger lidt, da dette giver en mere simpel model.

Skøn over koefficienterne i ridge regression findes ved at minimere:

$$\sum_{i=1}^n (x_i - \alpha - \beta_1 t_{i1} - \dots - \beta_d t_{id})^2 + \lambda \sum_{j=1}^d \beta_j^2. \quad (85)$$

Fra den fulde model kan man beregne den prædikterede værdi som i (82) og prædiktionsinterval for en kommende observation ved:

$$\hat{\xi} \pm 1.96 s_{cv}. \quad (86)$$