Oppsummering av STK2120

Geir Storvik

Vår 2011

Hovedtemaer

- Generelle inferensmetoder
- Spesielle modeller/metoder
- Bruk av R
  - Vil ikke bli testet på kommandoer, men må forstå generelle utskrifter

Generelle inferensmetoder

- Estimering
  - Maksimum likelihood
- Konfidensintervaller
  - Normaltilnærming
  - Bootstrapping
- Hypotesetesting
  - Likelihood ratio test
- Grunnlag: Sannsynlighetsregning

Spesielle modeller/metoder

- Variansanalyse
- Regresjon
  - Lineær
  - Ikke-lineær
  - Logistisk
- Kategoriske data/føyningstest
- Sjekk av modell-antagelser
  - Transformasjoner

Maksimum likelihood/Sannsynlighetsmaksimering

- $y_1, ..., y_n \overset{uif}{\sim} f(y;\theta)$
- $L(\theta;\mathbf{y}) = f(y_1, ..., y_n;\theta) = \prod_i f(y_i;\theta)$
- $\log L(\theta;\mathbf{y}) = \sum_i \log f(y_i;\theta)$
- $\hat{\theta} = \mathrm{argmax}_\theta L(\theta;\mathbf{y})$
  - Konsistent, asymptotisk effisient
  - Analyttiske løsninger for lineære/Gaussiske modeller
    Ett-/to- utvalgs modeller, variansanalyse, lineær regresjon
  - Generelt: Numerisk optimering
- For stor n:

$$\hat{\theta} \approx N(\theta, I(\hat{\theta})^{-1}) \approx N(\theta, J(\hat{\theta};\mathbf{y})^{-1})$$

$$J(\theta;\mathbf{y}) = -\frac{\partial^2}{\partial\theta\partial\theta^T}\log L(\theta;\mathbf{y})$$

$$I(\theta) = E[J(\theta;\mathbf{y})] \quad \text{Alltid pos. (semi)definit}$$

Konfidensintervaller

- Normaltilnærming: $\hat{\theta}_j \pm z_{\alpha/2}\text{SE}(\hat{\theta}_j)$.
  - $\text{SE}(\hat{\theta}_j)$: Normaltilnærming ($\sqrt{I(\theta)_{jj}}$) eller bootstrapping
- Bootstrap intervaller: $\hat{\theta}^*_{j,1}, ..., \hat{\theta}^*_{j,B}$ bootstrap simuleringer
  - Normaltilnærming: $\text{SE}(\hat{\theta}_j) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{j,b} - \bar{\hat{\theta}}^*)^2}$
  - Standard bootstrap intervaller: $(\hat{\theta}_j - \delta_U, \hat{\theta}_j - \delta_L)$

$$\hat{\delta}_L = \hat{\theta}^*_{j,(k_1)} = k_1 \text{ minste } \hat{\theta}^*_{j,b}$$

$$\hat{\delta}_U = \hat{\theta}^*_{j,(k_2)} = k_2 \text{ minste } \hat{\theta}^*_{j,b}$$

$$k_1 = B * \alpha/2, k_2 = B * (1 - \alpha/2)$$

Numerisk optimering

- Sentrale begreper
  - Likelihood $L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$
  - Skår funksjonen $s(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \mathbf{y})$
  - Observert informasjon $J(\theta; \mathbf{y}) = -\frac{\partial}{\partial \theta \theta^T} \log L(\boldsymbol{\theta}; \mathbf{y})$
  - Forventet (Fisher) informasjon $I(\boldsymbol{\theta}) = E[J(\boldsymbol{\theta}; \mathbf{y})]$
- Newton-Raphson

$$\boldsymbol{\theta}^{s+1} = \boldsymbol{\theta}^{s} + J(\boldsymbol{\theta}^{s}; \mathbf{x})^{-1} s(\boldsymbol{\theta}; \mathbf{x})$$

  - Scoring: $J(\boldsymbol{\theta}^s; \mathbf{x}) \rightarrow I(\boldsymbol{\theta}^s)$.
  - Mindre hopp
  - Reparametrisering
  - Dimensjonsreduksjon

Bootstrapping

- Av interesse: Egenskaper til $\hat{\theta} = \hat{\theta}(\mathbf{y})$ *ved gjentatt bruk av denne*
- Bootstrap idé: Simuler $\hat{\theta}^* = \hat{\theta}(\mathbf{y}^*)$ der $\mathbf{y}^* = (y_1^*, ..., y_n^*)$ er bootstrap simuleringer av $\mathbf{y}$.
- Ikke-parametrisk bootstrapping: Trekk $y_1^*, ..., y_n^*$ *med tilbakelegging* fra $\{y_1, ..., y_n\}$.
- Parametrisk bootstrapping: Anta $y_i \sim f(y; \theta)$. Simuler $y_i^* \sim f(y; \hat{\theta})$
- Semi-parametrisk bootstrapping: Mellomting
  Eksempel: Regresjon

$$\begin{aligned} y_i =& g(x_i; \beta) + \varepsilon_i \\ y_i^* =& g(x_i; \hat{\beta}) + \varepsilon_i^* \end{aligned}$$

  ε_i^* trelles med tilbakelegging fra $\{\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n\}$, $\hat{\varepsilon}_i = y_i - g(x_i; \hat{\beta})$.
- Forventningsskjevhet, usikkerhet, konfidensintervaller
- Egenskaper: STK4170

Hypotesetesting

Antar data $y_1, ..., y_n \overset{uif}{\sim} f(y;\theta)$
Ønsker å teste $H_0 : \theta \in \Omega_0$ mot $H_a : \theta \in \Omega_a$
Prosedyre

- Spesifiser en test-observator
- Bestem et forkastningsområde for gitt signifikansnivå
- Beregn test-observator og forkastningsområde numerisk og konkluder
  - Hvis test-observator i forkastningsområde, forkast H_0 på det gitte signifikansnivå
  - Ellers, konkluder med at det ikke er grunnlag i data for å forkaste H_0 på det gitte signifikansnivå.
    Merk: Dette er *ikke* det samme som å påstå at H_0 er riktig!
- Ofte vanlig å rapportere P-verdi som angir hvor mye bevis det ligger i data.
- Merk: Bør skille mellom *statistisk signifikant* og *praktisk signifikant*
  Ved mye data kan en ende opp med å forkaste $H_0 : \theta = \theta_0$ selv om $\hat{\theta}$ er svært lik θ_0.

Likelihood ratio test

Antar data $y_1, ..., y_n \overset{uif}{\sim} f(y; \theta)$
Ønsker å teste $H_0 : \theta \in \Omega_0$ mot $H_a : \theta \in \Omega_a$

- Neyman-Pearson: $H_0 : \theta = \theta_0$ mot $H_a; \theta = \theta_a$

$$LR = \frac{L(\theta_0; \mathbf{y})}{L(\theta_a; \mathbf{y})}$$

  optimal testobservator

- Generell likelihood ratio

$$LR = \frac{\max_{\theta \in \Omega_0} L(\theta; \mathbf{y})}{\max_{\theta \in \Omega} L(\theta; \mathbf{y})}, \quad \Omega = \Omega_0 \cup \Omega_a$$

- $-2 \log LR \overset{H_0}{\approx} \chi^2_{\text{df}}$, $\text{df} = |\Omega| - |\Omega_0|$
- P-verdi: $\Pr(\chi^2_{\text{df}} > -2 \log LR)$
- Ofte: LR må beregnes numerisk.

Variansanalyse

- Enveis variansanalyse

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \sum_i \alpha_i = 0$$

  - $H_0 : \alpha_i = 0, \quad F = \frac{SSTr/(I-1)}{SSE/I(J-1)}$
- Testing av mange hypoteser
  - Tukey's metode

$$H_0 : \alpha_i = \alpha_j, |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > Q_{\alpha,I,I(J-1)}\sqrt{MSE/J}$$

- Toveis variansanalyse

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

$$H_0 : \alpha_i = 0 \quad F = \frac{SSA/(I-1)}{SSE/IJ(K-1)}$$

$$H_0 : \beta_j = 0 \quad F = \frac{SSB/(J-1)}{SSE/IJ(K-1)}$$

$$H_0 : \delta_{ij} = 0 \quad F = \frac{SSAB/(I-1)(J-1)}{SSE/IJ(K-1)}$$

Lineær regresjon

- Modell $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$
- Antagelser
  - $E[\varepsilon_i] = 0$
  - $Var[\varepsilon_i] = \sigma^2$
  - Uavhengighet
  - Normalfordelte
- Estimering
  - $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
  - Forventningsrett
  - $\mathrm{COV}(\widehat{\boldsymbol{\beta}}) = \sigma^2[\mathbf{X}^T\mathbf{X}]^{-1}$
  - $\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_i (y_i - \hat{y}_i)^2$
  - $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}, \mathbf{H} = \mathbf{X}[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T$
  - Projeksjon ned i plan spent ut av x-ene.
- Konfidensintervaller
  $\widehat{\beta}_j \pm t_{\alpha/2;n-k-1} s_{\widehat{\beta}_j}$

Ikke-lineær regresjon

- $Y_i = g(\mathbf{x}_i; \boldsymbol{\beta}) + \varepsilon_i$
- Vanlige antagelser på $\{\varepsilon_i\}$.
- Numerisk optimering for å finne ML-estimater
- Egenskaper/konfidensintervall ved normaltilnærming eller bootstrapping

Logistisk regresjon

- Respons $Y_i \in \{0, 1\}$.
- $Y_i \sim \text{Binom}(1, p(x_i))$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- Numerisk optimering for å finne ML-estimater
- Egenskaper/konfidensintervall ved normaltilnærming (eller bootstrapping)
- Eksempel på *Generaliserte lineære modeller*, tema i STK3100.

Analyse av kategoriske data

- Gruppering av data i kategorier, data er antall innen hver kategori
- Sentral fordeling: Multinomisk fordeling
- Enveis gruppering
- Toveis gruppering
  - Test av homogenitet
  - Test av uavhengighet

En-veis gruppering

- En populasjon, utvalg på n, N_i antall i kategori i
- Antar $(N_1, ..., N_k) \sim \text{Multinom}(n, p_1, ..., p_k)$
- $H_0 : p_i = p_{i0}, i = 1, ..., k$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} \overset{H_0}{\approx} \chi^2_{k-1}$$

- $H_0 : p_i = \pi_i(\theta), i = 1, ..., k$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i(\hat{\theta}))^2}{n\pi_i(\hat{\theta})} \overset{H_0}{\approx} \chi^2_{k-1-m}$$

- $\hat{\theta}$ er ML-estimat.
- Kan brukes til testing av fordelingsantagelser

To-veis gruppering

- Testing av homogenitet
  - I populasjoner, utvalg n_i fra populasjon i, n_{ij} fra kateg. j
  - $(N_{i1}, ..., N_{iJ}) \sim \text{Multinom}(n_i; p_{i1}, ..., p_{iJ}), i = 1, ..., I.$
  - $H_0 : p_{ij} = p_j$
  - Pearson's χ^2 test, $df = (I - 1) * (J - 1)$
- Testing av uavhengighet
  - 1 populasjon, utvalg n, n_{ij} fra kateg. (i, j)
  - $(N_{11}, ..., N_{ij}, ..., N_{IJ}) \sim \text{Multinom}(n; p_{11}, ..., p_{ij}, ..., p_{IJ}).$
  - $H_0 : p_{ij} = p_{i\cdot} * p_{\cdot j}$
  - Pearson's χ^2 test, $df = (I - 1) * (J - 1)$

Veien videre

- STK2120 dekker de *generelle* prinsipper.
- Kan takle mange ulike situasjoner (også mange vi ikke har diskutert!)
- Mange aspekter som krever mer.
- Illustrasjon relasjon lengde fisk og alder

Fiske data

- Oblig: $\{(l_i, a_i), i = 1, ..., n\}$
- I praksis $\{(l_{b,i}, a_{b,i}, x_b), b = 1, .., , B, i = 1, ..., n_b\}$, b båt.
- Av interesse: $E[l_{b,i}|a_{b,i}, x_b]$, $Pr(A_{b,i} = a|x_b)$.
- Regresjonsmodeller for multinomiske data, tema i STK3100

$$\Pr(A_{b,i} = a) = \frac{\exp(\alpha_{a,0} + \alpha_{a,1}x_b)}{\sum_{a'} \exp(\alpha_{a',0} + \alpha_{a',1}x_b)}$$

- Fisk fra samme båt likere enn fisk fra forskjellige båter.
- Mulig modell:

$$l_{b,i} = \beta_0 + \eta_b + \beta_1 \log(a_{b,i}) + \varepsilon_{b,i}$$

  der $\eta_b \sim N(0, \sigma_\eta^2)$.
- η_b er en *tilfeldig effekt* som modellerer *korrelasjoner innen båter*
  Tema i STK3100
- Også aktuelt å modellere korrelasjoner i fordeling for alder
  Tema i STK3100
- Tidsstrukturer når data er samlet inn over flere år:
  STK3100/4060/4110

Fiske data

$$l_{b,i} = \beta_0 + \eta_b + \beta_1 \log(a_{b,i}) + \varepsilon_{b,i}$$

- Ofte: Alder mangler, men lengde kan gi informasjon om alder
  Missing data: Simuleringstilnærming STK4050, andre tilnærminger i andre kurs.
- Romlig struktur: Båter som fisker i nærheten av hverandre vil ha likere lengde/alder
  Bygger inn korrelasjoner mellom η_b-ene.
  Romlig statistikk: STK4150
- Mange mulige modeller, hvordan velge mellom disse? STK4160
  Bootstrapping i kompliserte situasjoner: STK4170
- Vet mye om hvordan fisk vokser fra tidligere studier, dvs har gode gjett på β_0, β_1.
  Apriori informasjon: Bayesiansk statistikk, STK4020
- Kompliserte beregninger: Monte Carlo metoder STK4050