UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1100 — Statistiske metoder og dataanalyse 1 - Løsningsforslag
Eksamensdag:	Mandag 30. november 2015.
Tid for eksamen:	14.30 – 18.00.
Oppgavesettet er på 6 sider.	
Vedlegg:	Tabell over ??
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

(a) Vi har

$$L(\theta) = f(x_1, ..., x_n; \theta) \overset{uavh}{=} \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta^2} x_i e^{-x_i/\theta}$$

$$l(\theta) = \sum_{i=1}^{n} [-2\ln(\theta) + \ln(x_i) - x_i/\theta] = -2n\ln(\theta) + \sum_{i=1}^{n} \ln(x_i) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

(b) Maksimum likelihood prinsippet er å estimere parameteren ved den verdi som maksimerer likelihood funksjonen, som vil tilsvare å maksimere log-likelihood funksjonen. Dette svarer til å velge den parameterverdi som gjør dataene mest "sannsynlige".

Vi har her at

$$\frac{d}{d\theta} l(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i$$

Hvis vi setter dette lik null og løser mhp θ, får vi $\hat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} x_i$.

(Fortsettes side 2.)

Vi bør i prinsippet også sikre oss at dette er et maks.punkt (men dette har vi ikke lagt mye vekt på i kurset). Vi har at

$$\frac{d^2}{d\theta^2}l(\theta) = \frac{2n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i$$

Hvis vi setter inn verdien $\hat{\theta}$, får vi $-2n/\hat{\theta}^2$ som er negativ og dermed blir det et makspunkt

(c) Vi har at momentestimatoren er løsningen av

$$\frac{1}{n}\sum_{i=1}^{n} x_i = E(X) = 2\theta$$

som gir $\hat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} x_i$. Videre er

$$E[\hat{\theta}] = E[\tfrac{1}{2n}\sum_{i=1}^{n} x_i] = \tfrac{1}{2n}\sum_{i=1}^{n} E[x_i] = \tfrac{1}{2n}\sum_{i=1}^{n} 2\theta = \theta$$

og er dermed forventningsrett.

(d) Vi får $\hat{\theta} = 166.40/(2 * 20) = 4.16$. Vi har videre at

$$V(\hat{\theta}) = V(\tfrac{1}{2n}\sum_{i=1}^{n} x_i) = \tfrac{1}{4n^2}V(\sum_{i=1}^{n} x_i) \overset{uavh}{=} \tfrac{1}{4n^2}\sum_{i=1}^{n} V(x_i) = \tfrac{1}{4n^2}\sum_{i=1}^{n} 2\theta^2 = \tfrac{1}{2n}\theta^2$$

og dermed blir standardfeilen $\sigma_{\hat{\theta}} = \theta/\sqrt{2n}$. Estimert standardfeil blir da $\hat{\sigma}_{\hat{\theta}} = 4.16/\sqrt{40} = 0.658$.

(e) Vi har at

$$\Pr(-z_{\alpha/2} < \tfrac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} < z_{\alpha/2}) \approx 1-\alpha$$

$$\Updownarrow$$

$$\Pr(\hat{\theta} - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}) \approx 1-\alpha$$

som gir at $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ er et tilnærmet $100(1-\alpha)\%$ konfidensintervall. Setter vi inn data får vi 95% intervallet til å bli $[2.87, 5.45]$.

Alternativt kan en bruke at man har en formel for standard feilen til $\hat{\theta}$

(Fortsettes side 3.)

slik at

$$\begin{aligned}
\Pr(-z_{\alpha/2} < \tfrac{\hat{\theta}-\theta}{\theta/\sqrt{2n}} < z_{\alpha/2}) &\approx 1-\alpha \\
&\Updownarrow \\
\Pr(-z_{\alpha/2}\sqrt{2n} < \tfrac{\hat{\theta}-\theta}{\theta} < z_{\alpha/2}/\sqrt{2n}) &\approx 1-\alpha \\
&\Updownarrow \\
\Pr(-z_{\alpha/2}/\sqrt{2n} < \tfrac{\hat{\theta}}{\theta} - 1 < z_{\alpha/2}/\sqrt{2n}) &\approx 1-\alpha \\
&\Updownarrow \\
\Pr(1 - z_{\alpha/2}/\sqrt{2n} < \tfrac{\hat{\theta}}{\theta} < 1 + z_{\alpha/2}/\sqrt{2n}) &\approx 1-\alpha \\
&\Updownarrow \\
\Pr(\frac{\hat{\theta}}{1 - z_{\alpha/2}/\sqrt{2n}} > \theta > \frac{\hat{\theta}}{1 + z_{\alpha/2}/\sqrt{2n}}) &\approx 1-\alpha
\end{aligned}$$

som gir $\frac{\hat{\theta}}{1\pm z_{\alpha/2}/\sqrt{2n}}$ som et tilnærmet $100(1-\alpha)\%$ konfidensintervall. Da får man $[3.18, 6.03]$ som et tilnærmet 95% konfidensintervall

(f) I bootstrapping så estimerer vi egenskaper til en estimator ved hjelp av simuleringer, basert på at egenskaper som varians, konfidensinterval etc er relatert til hvordan estimatoren oppfører seg ved *gjentatte forsøk*. Problemet i å utføre dette er imidlertid at fordelingen $F(x)$ er ukjent. I ikke-parametrisk bootstrapping bruker vi $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$ som et estimat på $F(x)$ og simulering fra denne svarer til å trekke fra data med tilbakelegging.

Fra simuleringene kan vi da få et Bootstrap persentil intervall som er gitt ved $[2.98, 5.40]$. Dette intervallet er ganske likt det vi fikk ved normaltilnærming, noe som også virker rimelig gitt at histogrammet av bootstrap simuleringene ser ganske normalfordelte ut.

Oppgave 2.

(Fortsettes side 4.)

(a) Vi har at

$$\begin{aligned} L(\beta_0,\beta_1,\sigma^2) =& f(y_1,...,y_n;\beta_0,\beta_1,\sigma^2) \overset{uavh}{=} \prod_{i=1}^{n} f(y_i;\beta_0,\beta_1,\sigma^2) \\ =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-(\beta_0-\beta_1 x_{i1}))^2} \\ l(\beta_0,\beta_1,\sigma^2) =& \sum_{i=1}^{n} [-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i-(\beta_0-\beta_1 x_{i1}))^2] \\ =& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0-\beta_1 x_{i1}))^2 \end{aligned}$$

som viser at maksimering av $l(\beta_0,\beta_1,\sigma^2)$ mhp β_0,β_1 er ekvivalent med minimering av $\sum_{i=1}^{n}(y_i-(\beta_0-\beta_1 x_{i1}))^2$ som svarer til minste kvadraters prinsippet.

(b) Den vanlige test-observator er

$$T = \frac{\hat{\beta}_1 - 0}{s_{\hat{\theta}}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \frac{\frac{\hat{\beta}_1-0}{\sigma/\sqrt{\sum_i (x_i-\bar{x})^2}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}}.$$

Vi har videre at telleren er standard normalfordelt og at $(n-2)S^2/\sigma^2$ er χ^2-fordelt med $n-2$ frihetsgrader. Ved å bruke at $\hat{\beta}_1$ og S^2 er uavhengige, får vi at $T \sim t_{n-2}$.

Fra utskriften har vi at $t = 10.55$ og tilhørende P-verdi er svært liten som gir klart grunnlag for å forkaste H_0 (på ethvert fornuftig signifikansnivå).

(c) Vi får en økning i forklaringsgrad R^2 fra 0.54 til 0.58. Imidlertid kan denne være noe optimistisk så man bør heller bruke den justerte R^2. Denne gir også en liten forbedring (fra 0.53 til 0.58) slik at den nye modell er å foretrekke.

(d) Vi har at

$$\begin{aligned} \boldsymbol{E} =& \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ =& \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \ [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y} \end{aligned}$$

der $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Dermed er residualene lineære kombinasjoner av Y-ene og dermed

(Fortsettes side 5.)

normalfordelte. Videre er

$$\begin{aligned}
E(\boldsymbol{E}) =& E([\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{Y}) \\
=& [\boldsymbol{I}-\boldsymbol{H}]E(\boldsymbol{Y}) \\
=& [\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{X}\boldsymbol{\beta} \\
=& [\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{X}\boldsymbol{\beta} \\
=& \boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\
=& \boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\boldsymbol{\beta}=0 \\
\mathrm{Cov}(E) =& \mathrm{Cov}([\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{Y}) \\
=& [\boldsymbol{I}-\boldsymbol{H}]\mathrm{Cov}(\boldsymbol{Y})[\boldsymbol{I}-\boldsymbol{H}]^T \\
=& \sigma^2[\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T][\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\
=& \sigma^2[I-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T+\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\
=& \sigma^2[I-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]=\sigma^2[\boldsymbol{I}-\boldsymbol{H}]
\end{aligned}$$

(e) Standardiserte residualer er definert ved

$$e_i^* = \frac{e_i}{\sqrt{1-h_{ii}}}$$

Disse vil alle ha varians σ^2, noe som gjør det lettere å vurdere disse. (Hvis de også deler på s i definisjonen av e_i^* så er det ok.)

Residualplottene ser ut til å virke rimelige ut ifra antagelsene, bortsett fra at det ser ut til å være en stigende tendens i residualene med indeks nummer som indikerer at det er en slags tid-korrelasjon her.

Oppgave 3.

(a) Vi har at $E(\hat{p}_K) = p_K$ og $V(\hat{p}_K) = p_K(1-p_K)/n_k$ Dermed får vi at $E(Z) = 0$ og $V(Z) = 1$. For n_k stor, er Y_k tilnærmet normalfordelt. Siden Z er en lineærkombinasjon av Y_k, blir denne også tilnærmet normalfordelt.

(b) Vi må først bestemme en testobservator. Her vil det være fornuftig å bruke

$$Z = \frac{\hat{p}_K - 0.339}{\sqrt{0.339(1-0.339)/n_k}}$$

som vil være tilnærmet standard normalfordelt når n_k er stor (som her). Vi får i dette tilfellet at $z = -5.78$ som (i absolutt verdi) er klart større en $z_{0.025} = 1.96$. Vi forkaster derfor H_0 på $\alpha = 0.05$ signifikansnivå.

(Fortsettes side 6.)

(c) Vi har nå at

$$\begin{aligned} V(\hat{p}_K - \hat{p}_M) =& \frac{p_K(1-p_K)}{n_K} + \frac{p_M(1-p_M)}{n_M} \\ \overset{H_0}{=}& p(1-p)(n_K^{-1} + n_M^{-1}) \end{aligned}$$

som gir at

$$Z = \frac{\hat{p}_K - \hat{p}_M}{\sqrt{\hat{p}(1-\hat{p})(n_K^{-1} + n_M^{-1})}}$$

er tilnærmet standard normalfordelt når n-ene er store. Her har vi brukt at $\hat{p} = (515{+}27)/(1520{+}191) = 0.317$. Da blir $z = -5.528$ (5.544 ved hvis en kun bruker de 3 desimaler) som også er klart signifikant. Da vi har at $|z| > z_{0.000001}$ vet vi i allefall at P-verdien er mindre enn $2 * 0.000001 = 0.000002$.

(d) Dette vil være en bedre test da vi tar hensyn til usikkerheten i $\hat{p}_M$. Da det er såpass mye data for å estimere $\hat{p}_M$, gjør det ikke den store forskjellen her.

(e) Hvis vi definerer en forklaringsvariabel x_i til å være lik 1 hvis individ i er en kvinne og 0 hvis det er en mann, vil en logistisk regresjonsmodell svare til at

$$\begin{aligned} \Pr(Y_i = 1|x_i) =& \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \\ =& \begin{cases} \frac{e^{\beta_0}}{1+e^{\beta_0}} & x_i = 0 \\ \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} & x_i = 1 \end{cases} \end{aligned}$$

Hvis vi nå definerer $p_M = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ og $p_K = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$, så ser vi at $p_K = p_M$ svarer til at β_1 er 0, slik at vi kan teste det innenfor logistisk regresjon.

SLUTT