UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse **Løsningsforslag**
Eksamensdag:	Tirsdag 2. desember, 2025
Tid for eksamen:	15.00 – 19.00.
Oppgavesettet er på 4 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator Formelsamling for STK1110

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

Oppgave 1

(a) $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ er forventningsrett for λ siden $\mathrm{E}(\hat{\lambda}) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\lambda = \frac{1}{n}n\lambda = \lambda$

Dessuten $\hat{\lambda}$ er momentestimator for λ siden den er løsningen av $\lambda = \mathrm{E}(X_i) = \bar{X}$.

Videre er $\mathrm{Var}(\hat{\lambda}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\lambda = \frac{\lambda}{n}$.

(b) Likelihooden for dataene blir $L(\lambda) = \prod_{i=1}^{n}\frac{\lambda_i^X}{X_i!}\exp(-\lambda) = \frac{\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}\exp(-n\lambda)$.

Dermed blir log-likelihood $l(\lambda) = \ln(\lambda)\sum_{i=1}^{n} X_i - n\lambda - \ln(\prod_{i=1}^{n} X_i!)$

og videre score-funksjone $s(\lambda) = l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n$.

Når vi setter $s(\hat{\lambda}) = 0$ får vi løsningen $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$.

Dette er et maksimum for $l(\lambda)$ (siden $s'(\lambda) < 0$ for alle λ) blir altså $\hat{\lambda}$ også maximum likelihood estimator for λ.

(c) La $p(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda)$ være punktsannsynlighetene i Poissonfordelingen.

Da kan den forventede (Fisher) informasjonen i en observasjon X_i gis som $I(\lambda) = -\mathrm{E}\left(\frac{\partial^2 \ln(p(X_i;\lambda))}{\partial\lambda^2}\right) = -\mathrm{E}(-\frac{X_i}{\lambda^2}) = \frac{\mathrm{E}(X_i)}{\lambda^2} = \frac{1}{\lambda}$

I store utvalg gis variansen til maksimum likelihoood estimatoren $\hat{\lambda}$ som $\frac{1}{nI(\lambda)}$ som blir lik $\frac{\lambda}{n} = \mathrm{Var}(\hat{\lambda})$ som utledet i punkt a).

I henhold til Cramer-Rao's nedre grense finnes det ikke forventningsrette estimatorer med mindre varians enn $\frac{1}{nI(\lambda)}$.

(Fortsettes på side 2.)

(d) For Anencefalidataene blir $\hat{\lambda} = 260/132 \approx 1.97$ og estimatoren har standardfeil $se(\hat{\lambda}) = \sqrt{\hat{\lambda}/132} = 0.122$

$\hat{\lambda} \pm 1.96 se(\hat{\lambda}) = (1.73, 2.21)$ blir et tilnærmet 95%-konfidensintervall for λ siden $\hat{\lambda}$ er tilnærmet normalfordelt pga. maksimum likelihood estimatorens egenskaper i store utvalg (evt. direkte pga. sentralgrenseteoremet) der forventningen i normalfordelinger λ og variansen λ/n. Dermed har vi

$$0.95 \approx \mathrm{P}(-1.96 < Z = \frac{\hat{\lambda} - \lambda}{se(\hat{\lambda})} < 1.96) = \mathrm{P}(\hat{\lambda} - 1.96 se(\hat{\lambda}) < \lambda < \hat{\lambda} + 1.96 se(\hat{\lambda}))$$

(e) Vi har at aposteriorifordelingen kan skrives som $\pi(\lambda|x_1, \ldots, x_{132}) = c\pi(\lambda)L(\lambda)$ der $L(\lambda)$ er likelihooden definert i punkt b) og c en normaliseringskonstant som bestemmes av at $\int_0^\infty \pi(\lambda|x_1, \ldots, x_n)d\lambda = 1$. Her får vi da

$$\pi(\lambda|x_1, \ldots, x_{132}) = c\pi(\lambda)L(\lambda) = c\exp(-\lambda)\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\exp(-n\lambda) = c'\lambda^{\sum_{i=1}^n x_i + 1 - 1}\exp(-(n+1)\lambda)$$

for en annen konstant c'. Aposteriorifordelingen har dermed samme form som en gamma-fordeling med parametre $\alpha = 1 + \sum_{i=1}^{132} x_i$ og $\beta = 1/(n+1)$ og må dermed være denne gammafordelingen.

I gammafordelingen er forventningen lik $\alpha\beta$ og dermed blir aposteriori forventning $\frac{1+\sum_{i=1}^{132} x_i}{132+1} = \frac{261}{133} = 1.96$ og altså bare marginalt forskjellig fra $\hat{\lambda} = 1.97$,

Oppgave 2

(a) Vi kan skrive testobservatoren

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{1/12 + 1/12}}\frac{\sigma}{S_p} = Z\frac{\sigma}{S_p}$$

der Z er standardnormalfordelt (under nullhypotesen) og uavhengig av S_p. Samtidig blir (eksakt) $\frac{S_p^2}{\sigma^2}(24-2) \sim \chi^2_{22}$ fordelt. Men dette medfører at $T \sim t_{22}$ under nullhypotesen.

Man forkaster hvis absoluttverdien $|T|$ er stor nok, dvs. med nivå α når $|T| >$ $(1-\alpha/2)100$-persentilen i t_{22}.

(b) Vi har observert verdi $T = -2.4618$. Dermed blir P-verdien, som angir sannsynligheten for et mer ekstremt utfall enn det observerte gitt at nullhypotesen holder , lik $\mathrm{P}(|T_{22}| > 2.4618) = 2\mathrm{P}(T_{22} > 2.4618) = 0.022 < 5\%$ når T_{22} er en t-fordelt stokastisk variabel med 22 frihetsgrader. Vi forkaster altså nullhypotesen med et nivå på 5% og det er (en viss) grunn til å påstå at nitrogen påvirker avlingens størrelse.

(Fortsettes på side 3.)

95% konfidensintervallet er regnet ut med formelen $\bar{Y}_1 - \bar{Y}_2 \pm t_{0.025} S_p \sqrt{\frac{1}{12} + \frac{1}{12}}$ der $t_{0.025}$ er 97.5-persentilen i t_{22} fordelingen (og lik 2.074 som man f.eks. kan regne ut fra det oppgitte intervallet).

(c) Denne to-utvalgssituasjonen kan reformuleres som et regresjonsproblem med en forklaringsvariabel x_{1i} der vi har kodet $x_{1i} = -0.5$ hvis det ikke er tilført nitrogen og $x_{1i} = 0.5$ hvis det nitrogen er tilført ved at man skriver $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$ der $\varepsilon_i \sim \mathrm{N}(0, \sigma)$ og uavhengige. Man ser nemlig at

$$\mathrm{E}(Y_i) = \begin{cases} \beta_0 - \frac{1}{2}\beta_1 = \mu_1 & \text{hvis } x_{1i} = -0.5 \\ \\ \beta_0 + \frac{1}{2}\beta_1 = \mu_2 & \text{hvis } x_{1i} = +0.5 \end{cases}$$

Dermed får β_1 fortolkning som forskjellen i avling med og uten nitrogen samtidig som β_0 vil svare til det totale gjennomsnittet. Man har da også at $5.617 = \hat{\beta}_1 = \bar{Y}_2 - \bar{Y}_1 = 57.683 - 52.067$ og $54.875 = \hat{\beta}_0 = (\bar{Y}_1 + \bar{Y}_2)/2 = (52.067 + 57.683)/2$.

Estimatet S_p for σ i to-utvalgstesten vil være det samme som estimatet av σ i den lineære regresjonen, altså $S_p = 5.588$.

Oppgave 3

(a) Den multiple regresjonsmodellen med konstantledd β_0 samt regresjonskoeffisienter β_1 og β_2 svarende til forklaringsvariablene x_{1i} og x_{i2} kan skrives opp ved $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ der $\varepsilon_i \sim \mathrm{N}(0, \sigma)$ er uavhengige.

Vi finner $\hat{\beta}_0 = 54.875$, $\hat{\beta}_1 = 5.617$ og $\hat{\beta}_2 = -3.983$ samt estimatet for σ som $\hat{\sigma} = 5.309$.

Det betyr at $\hat{\beta}_0$ og $\hat{\beta}_1$ er identiske med dem i den enkle lineære regresjonen, mens vi selvfølgelig ikke hadde med noen $\hat{\beta}_2$ i denne. Vi finner også at estimert σ i den nye regresjonen er litt mindre enn i den enkle. Dessuten finner vi en lnoe større forklart andel av variasjonen $R^2 = 0.3246$ (mot 0.216 i den enkle) noe som samsvarer med at standardfeil for $\hat{\beta}_1$ er mindre, $t_1 = \hat{\beta}_1 / se_1 = 2.591$ større og P-verdi for $\beta_1 = 0$ litt mindre.

Det å inkludere x_{2i} i modellen har altså ført til mindre residual variasjon og dermed større utsagnskraft.

(b) Modellmatrisen $\mathbf{X}$ er en $24x3$ matrise med alle verdier i første kolonne lik 1, verdiene av x_{1i} i andre kolonne og x_{2i} i tredje. Eksplisitt ser den ut som

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \cdots & \cdots & \cdots \\ 1 & x_{1,24} & x_{2,24} \end{bmatrix}$$

Matrisemultiplikasjon gir da

$$\mathbf{X}^\intercal \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{bmatrix} = \begin{bmatrix} n & 0 & 0 \\ 0 & \sum_{i=1}^n x_{1i}^2 & 0 \\ 0 & 0 & \sum_{i=1}^n x_{2i}^2 \end{bmatrix}$$

(Fortsettes på side 4.)

pga. at sentrering $\sum_{i=1}^n x_{ji} = 0$ og ukorrelerte x_{i1} og x_{2i} som gir $\sum_{i=1}^n x_{1i}x_{2i} = 0$

Minste kvadraters estimatorene blir da gitt som $\hat{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}$ der $\mathbf{Y}$ er kolonnen med $Y_1, \ldots, Y_{24}$. Siden $\mathbf{X}^\intercal\mathbf{X}$ er diagonal blir dens invers

$$(\mathbf{X}^\intercal\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{\sum_{i=1}^n x_{1i}^2} & 0 \\ 0 & 0 & \frac{1}{\sum_{i=1}^n x_{2i}^2} \end{bmatrix}$$

Det følger da at $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$, $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_{1i}Y_i}{\sum_{i=1}^n x_{1i}^2}$ og tilsvarende $\hat{\beta}_2 = \frac{\sum_{i=1}^n x_{2i}Y_i}{\sum_{i=1}^n x_{2i}^2}$.

For spesialtilfellet i punkt a) blir $\hat{\beta}_1$ differansen av gjennomsnitt mellom Y_i med $x_{1i} = 0.5$ og $x_{1i} = -0.5$ (som vi observerte) og tilsvarende blir $\hat{\beta}_2$ lik differansen i gjennomsnitt av Y_i med $x_{2i} = 0.5$ og $x_{2i} = -0.5$

(c) Vi ser at $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$. Dermed blir

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})^2 = \hat{\beta}_1^2 \sum_{i=1}^n x_{1i}^2 + \hat{\beta}_2^2 \sum_{i=1}^n x_{2i}^2$$

siden $\sum_{i=1}^n \hat{\beta}_1 x_{1i} \hat{\beta}_2 x_{2i} = \hat{\beta}_1\hat{\beta}_2 \sum_{i=1}^n x_{1i}x_{2i} = 0$. Da følger

$$\begin{aligned} R^2 = \frac{SSR}{SST} &= \frac{\hat{\beta}_1^2 \sum_{i=1}^n x_{1i}^2 + \hat{\beta}_2^2 \sum_{i=1}^n x_{2i}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^n x_{1i}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} + \frac{\hat{\beta}_2^2 \sum_{i=1}^n x_{2i}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \frac{(\sum_{i=1}^n x_{1i}Y_i)^2}{\sum_{i=1}^n x_{1i}^2 \sum_{i=1}^n (Y_i - \bar{Y})^2} + \frac{(\sum_{i=1}^n x_{2i}Y_i)^2}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n (Y_i - \bar{Y})^2} = r_1^2 + r_2^2 \end{aligned}$$

som skulle vises.

SLUTT