UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse
Eksamensdag:	28. november - 2023
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) Den øverste delen av utskriften gir en tabell relatert til regresjonskoeffisientene, en rad for hver β_j. Første kolonne gir estimat, 2. kolonne gir standard feil, 3 kolonne gir T-observatoren for å teste $H_0 : \beta_j = 0$ og siste kolonne gir tilhørende P-verdi.

Nederste del gir estimat på σ (Residual standard error), multippel R^2 og dens justerte verdi samt F-observatoren for å sjekke om noen av forklaringsvariable (i dette tilfelle bare en) har signfikant betydning for å vårklare variasjonen i responsen y.

(b) Vi får en stor $|T| = 8.680$ verdi med tilhørende svræt liten P-verdi som indikerer at `Alder` er en svært viktig forklaringsvariabel. Samtidig er R^2 ganske lav slik at det er kun en liten andel av variasjonen. Dette kan indikere at det er andre forklaringvariable som burde være med for å forklare responsen.

Et 95% konfidensintervall for β_1 (bruker at $T \sim t_{n-2}$ og $t_{0.025;155} = 1.975$):

$$-0.078601 \pm *1.975 * 0.009056 = [-0.0965, -0.0607]$$

(c) Når vi tester hypotesen på at $H_j : \beta_j = 0$ så gjør vi det *betinget på* at de øvrige forklaringsvariable er med i modellen. Det betyr at når Alder^2 ikke ser ut til å være viktig å ta med så er det som et tillegg til at Alder allerede er med. Tilsvarende så er det ikke viktig å ta med Alder hvis Alder^2 allerede er med.

Ser man på F-observatoren og tilhørende P-verdi så er denne klart signifikant som indikerer at *enten* Alder *eller* Alder^2 er viktig å ta med.

(d) Da prediksjonsintervallene også tar hensyn til usikkerheten i støyleddene, som har et standard avvik på 1.801, så blir intervallene ganske store.

(Fortsettes på side 2.)

Selv om regresjonskoeffisienten tilhørende `Alder`^{2} er svært liten, så vil den ha større effekt når alder er stor, dermed slår det mer ut i forskjell på intervallene. Et annet moment er også at gjennomsnittelig alder ser ut til å ligge nærmere 20 enn 90, noe som også påvirker usikkerheten i "forventningsdelen" $\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$.

Oppgave 2

(a) Vi har at

$$\begin{aligned} L(\beta) = p(x, y|\alpha, \beta) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \frac{1}{\beta^{3\alpha}\Gamma(3\alpha)} y^{3\alpha-1} e^{-y/\beta} \\ &= \frac{1}{\beta^{4\alpha}\Gamma(\alpha)\Gamma(3\alpha)} x^{\alpha-1} y^{3\alpha-1} e^{-[x+y]/\beta} \end{aligned}$$

og

$$\begin{aligned} \ell(\alpha, \beta) = &- 4\alpha \log(\beta) - \log\Gamma(\alpha) - \log\Gamma(3\alpha) - (\alpha - 1)\log(x) - \\ &(3\alpha - 1)\log(y) - \frac{1}{\beta}(x + y) \\ = &\text{konst} - 4\alpha \log(\beta) - \frac{1}{\beta}(x + y) \end{aligned}$$

der `konst` består av ledd som ikke ikke inneholder β.

Merk at selv om $X + Y \sim \text{Gamma}(4\alpha, \beta)$, så vil det ikke være helt riktig å ta utgangspunkt i dette direkte. Da kaster man bort noe informasjon (hvor stor X er i forhold til Y). I dette tilfellet når vi kun er interessert i å estimere β, så får vi det samme svaret, men det hadde ikke vært tilfelle hvis vi også skulle estimere α.

(b) Vi har

$$\frac{\partial}{\partial\beta}\ell(\alpha, \beta) = -\frac{4\alpha}{\beta} + \frac{1}{\beta^2}(x + y)$$

og vi finner estimatet som den verdi som gir den deriverte lik 0, noe som gir

$$\hat{\beta} = \frac{x + y}{4\alpha}.$$

Vi kan sjekke at det faktisk blir en maksimumsverdi ved at

$$\frac{\partial^2}{\partial\beta^2}\ell(\alpha, \beta) = +\frac{4\alpha}{\beta^2} - \frac{2}{\beta^3}(x + y)$$

som innsatt $\hat{\beta}$ gir

$$\frac{(4\alpha)^3}{(x + y)^2} - 2\frac{(4\alpha)^3}{(x + y)^2} = -\frac{(4\alpha)^3}{(x + y)^2} < 0$$

og dermed at vi har et maksimumspunkt.

(Fortsettes på side 3.)

Vi har at

$$E[\hat{\beta}] = \frac{1}{4\alpha}[E[X] + E[Y]] = \frac{1}{4\alpha}[\alpha\beta + 3\alpha\beta] = \beta;$$
$$V[\hat{\beta}] = \frac{1}{(4\alpha)^2}[V[X] + V[Y]] = \frac{1}{(4\alpha)^2}[\alpha\beta^2 + 3\alpha\beta^2] = \frac{\beta^2}{4\alpha}.$$

(c) Når α er heltall, så kan vi bruke at X er en sum av $X_1, ..., X_\alpha$ som hver er Gamma$(1, \beta)$ og uavhengige. Tilsvarende er Y sum av $Y_1, ..., Y_{3\alpha}$ som alle også er Gamma$(1, \beta)$ og uavhengige. Dermed er $\hat{\beta}$ et gjennomsnitt av 4α uif variable og av sentralgrenseteoremet tilnærmet normalfordelt.

Vi får $\hat{\beta} = 0.497$. Vi har at

$$I(\beta) = -E[\frac{\partial^2}{\partial\beta^2}\ell(\alpha, \beta)] = -\frac{4\alpha}{\beta^2} + \frac{2}{\beta^3}(\alpha\beta + 3\alpha\beta) = \frac{4\alpha}{\beta^2}$$

Dermed vil $V[\hat{\beta}] \approx \frac{\beta^2}{4\alpha} \approx \frac{\hat{\beta}^2}{4\alpha} = 0.0031$.

Oppgave 3

(a) Vi har at under $H_0 : \sigma_1^2 = \sigma_2^2$ så er $S_1^2/S_2^2 \sim F_{m-1,n-1}$. Her er $m = n = 15$ og $S_1^2/S_2^2 = 0.596/0.273 = 2.18315$. Fra tabell så ser vi at denne verdien er under 0.05 kvantilen som tilsier (siden vi har en to-sidig test) at P-verdien er større enn 0.1. Dermed ingen grunn til å forkaste H_0.

(b) Anta $X_i \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_i \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$. Hvis vi antar $\sigma_1 = \sigma_2$ så kan vi bruke

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/m + 1/n}} \sim t_{m+n-2}$$

der

$$S_p^2 = [(m-1)S_1^2 + (n-1)S_2^2]/(m+n-2) = (0.596 + 00.273)/2 = 0.4345$$
$$m+n-2 = 28$$

Vi har da

$$P(-t_{\alpha/2;28} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/m + 1/n}} \leq t_{\alpha/2;28}) = 1 - \alpha$$

Ved å gange med nevner og flytte $\bar{X} - \bar{Y}$ over på andre side av ulikehetstegnene, så får vi konfidensintervaller

$$\bar{X} - \bar{Y} \pm t_{\alpha/2;28} S_p\sqrt{2/15}$$

For $\alpha = 0.05$ blir $t_{\alpha/2;28} = 2.048$ og intervallet blir [0.060,1.046].

For $\alpha = 0.01$ blir $t_{\alpha/2;28} = 2.763$ og intervallet blir [-0.112,1.218].

(Fortsettes på side 4.)

(c) Konklusjon hvis $\alpha = 0.05$: H_0 *forkastes* på et $\alpha = 0.05$ signifikansnivå.
Konklusjon hvis $\alpha = 0.01$: H_0 forkastes *ikke* på et $\alpha = 0.01$ signifikansnivå.
Dette samsvarer med konfidensintervallene i (b) siden 0 ikke er med i 95% KI men er med i 99% intervallet.

(d) Når vi gjør mindre antagelser, så vil vi også typisk få svakere konklusjoner, i dette tilfelle høyere P-verdi.

Kvantilplottene gir indikasjon på at det kan være noe avvik fra normalitet, noe som tilsier at det er bedre å bruke den ikke-parametriske metoden. Dermed er det noe tvilsomt om en kan konkludere at det er forskjell mellom de to verdensdeler (det er det egentlig også hvis vi bruker T-testen da vi får forskjellige konklusjoner avhengig av om vi velger $\alpha = 0.05$ eller $\alpha = 0.01$).