UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Torsdag 2. desember 2010.
Tid for eksamen:	09.00 – 13.00.
Oppgavesettet er på 3 sider.	
Vedlegg:	Tabeller for normal- og t-fordeling.
Tillatte hjelpemidler:	Lærebok: Moore & McCabe "Introduction to the practice of

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

Oppgave 1

(a) Antagelser: Enkelt randomisert utvalg av størrelse n_1 fra $N(\mu_1, \sigma_1)$ og et uavhengig randomisert utvalg av størrelse n_2 fra $N(\mu_2, \sigma_2)$.

I dette tilfellet svarer de to utvalgene til to aldersgrupper. Kvantilplott av hver av aldersgruppene viser at det ikke er urimelig å anta data er normalfordelte da punktene ligger rimelig langs en linje og innenfor de to grensekurvene. Også likheten mellom gjennomsnitt og median antyder at data er rimelig symmetrisk fordelte.

Da fiskene er tilfeldig utvalgte fra den totale fangsten fra en enkelt båt, vil populasjonene være alle fisk av alder 6 og 7 år henholdsvis innenfor denne fangsten.

(b) Null-hypotese $H_0 : \mu_1 = \mu_2$ mot $H_a : \mu_1 < \mu_2$.

Da vi kan anta at fisk iallefall ikke blir mindre med alder er et ensidig alternativ fornuftig her.

(c) Testobservator:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Da vi har et en-sidig alternativ, er P-verdien gitt ved $P(T < t)$. Denne beregnes fra t-fordelingen med 84 frihetsgrader. Frihetsgrader er her beregnet av software (alternativ er $\min(n_1 - 1, n_2 - 1)$).

Vi konkluderer dermed med at vi kan forkaste H_0 og at forventet lengde for alder 6 år er signifikant mindre enn forventet lengde ved alder 7 år (P-verdi=0.000 med 84 frihetsgrader).

(Fortsettes på side 2.)

Oppgave 2

(a) Modell: $y = \beta_0 + \beta_1 x + \varepsilon$ der $\varepsilon \sim N(0, \sigma^2)$.

Parameter	Estimater
β_0	2.91226
β_1	0.71642
σ	0.09749

β_0: Skjæringspunkt (forventet verdi når $x = 0$). β_1: Stigningskoeffesient. Måler forventet endring i respons når forklaringsvariabel øker med en enhet. σ: Standardavvik til støyleddet. Måler variabilitet rundt regresjonslinjen.

(b) `R-Sq` eller R^2 er den multiple regresjonskoeffisienten og angir hvor mye av variabiliten i responsen som kan forklares av forklaringsvariabelen.

I dette tilfellet har vi at $R^2 = 74.16$ som betyr at en ganske stor andel av variabiliteten blir forklart og at vi derfor har en ganske god modell.

(c) Vi får et 95% konfidensintervall ved

$$[\text{estimat} \pm t^* SE_{\text{estimat}}]$$

der t^* er den verdi slik at $P(T > t^*) = 0.025$ i t-fordelingen med $n-2 = 131$ frihetsgrader. Da Tabell D i boka kun har 100 frihetsgrader, bruker vi denne og får at $t^* = 1.984$. Fra Minitab-utskriften har vi at estimatet er $b_1 = 0.716$ og $SE_{b_1} = 0.037$. Dette gir konfidensintervallet

$$[0.716 \pm 1.984 * 0.0370] = [0.643, 0.789]$$

Da intervallet ikke dekker 0, svarer dette til å forkaste H_0 med signifikansnivå $\alpha = 1 - C = 0.05$. Dette samsvarer med at P-verdien i utskriften er mindre enn 0.05.

(d) Konfidensintervallet omhandler estimering av $\mu_y = \beta_0 + \beta_1 x$ mens prediksjonsintervallet omhandler prediksjon av $y = \beta_0 + \beta_1 x + \varepsilon$.

Estimatet av μ_y, $\hat{\mu}_y$ og prediksjonen av y, $\hat{y}$ er begge lik $b_0 + b_1 x$ da beste gjett på ε for en ny observasjon er 0.

Prediksjonsintervallet blir større da vi her må ta hensyn til den ekstra variabiliteten som ligger i ε.

Vi får intervaller på opprinnelig skala ved å ta den inverse transformasjonen av endepunktene. Dette gir oss konfidensintervallet $[85.968, 91.728]$ og prediksjonsintervallet $[73.026, 107.985]$.

(e) Kvantilplott av residualer og histogram av residualer kan begge brukes til å sjekke normalantagelsen på støyleddene ε_i. At punktene ligger rimelig i nærheten av den rette linjen i kvantilplottet og at histogrammet ser rimelig symmetrisk ut uten outliere antyder at denne antagelsen er rimelig.

(Fortsettes på side 3.)

Plot av predikert verdi $\hat{y}_i$ mot residual kan brukes til å sjekke om konstant standardavvik for alle observasjoner er rimelig. Dette plottet kan kanskje antyde at variasjonen er noe mindre for store verdier av $\hat{y}_i$ (som svarer til store aldersgrupper), men her er det ikke mye data.

Residualer mot observasjonsnummer kan brukes til å sjekke uavhengighetsantagelsen. Her ser det ikke ut til å være noen struktur og dermed ingen indikasjon på at denne antagelsen ikke er rimelig.

Residualer mot forklaringsvariabel kan også brukes til å sjekke antagelsen om konstant standardavvik, tilsvarende $\hat{y}_i$ mot residual, men i tillegg sjekke om det er ikke-lineære strukturer tilstede. Plottet gir ingen indikasjon på ikke-linearitetar, men også her kan det se ut som om spredningen er mindre for høye aldersgrupper.

Oppgave 3

(a) Da summen av sannsynlighetene for de mulige utfallene må være 1, blir $P(\text{Alder} \leq 3) = 1 - 0.10 - 0.14 - 0.33 - 0.42 = 0.01$.

(b) La hendelsen C svare til at en fisk anslås til å ha alder 5 år og la A_i svare til at fisken har alder i. Da er ifølge Bayes regel

$$\begin{aligned} P(A_5) =& \frac{P(C|A_5)P(A_5)}{P(C|A_{\leq 3}P(A_{\leq 3})+P(C|A_4)P(A_4)+P(C|A_5)P(A_5)+P(C|A_6)P(A_6)+P(C|A_{\geq 7})P(A_{\geq 7})} \\ =& \frac{0.95 * 0.14}{0 * 0.01 + 0 * 0.10 + 0.95 * 0.14 + 0.05 * 0.33 + 0 * 0.42} \\ =& 0.89 \end{aligned}$$

Alternativt kan en argumentere for at siden det kun er mulig å anslå riktig alder eller ett år for lite, så må mulige aldre være 5 eller 6 år. En kan da forenkle til at

$$\begin{aligned} P(A_5) =& \frac{P(C|A_5)P(A_5)}{P(C|A_5)P(A_5)+P(C|A_6)P(A_6)} \\ =& \frac{0.95 * 0.14}{0.95 * 0.14 + 0.05 * 0.33} \\ =& 0.89 \end{aligned}$$