UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Løsningsforslag: Statistiske metoder og dataanalys
Eksamensdag:	Fredag 9. desember 2011
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over normalfordeling, tabell over t-fordeling og tabell over χ^2-fordeling
Tillatte hjelpemidler:	Formelsamling STK1100/STK1110 og godkjent kalkulator

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

Dette løsningsforslaget er mer detaljert enn det vil være rimelig å vente av en eksamensbesvarelse.

Oppgave 1

a) Residualene et definert som $y_i - \hat{\beta}_0 - \hat{\beta}_1 \text{lengde}_i - \hat{\beta}_2 \text{alder}_i$, der $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$ er estimatene for koeffesientene i forventningene til responsvariablene. De er med andre ord differansene mellom observasjonene og de tilpassede verdiene. For en veltilpasset modell skal de ikke vise noen systematikk, og plottene av residualene mot tilpassede verdier og uavhengige variable kan avsløre slike systematiske avvik.

1. Mangel på lineær sammenheng med de uavhengige varaiblene kan vise seg i plott av residualene mot disse variablene. Det kan gjøre det nødvendig å ta med kvadratiske ledd for eksmpel.
2. Ikke konstant varians kan vise seg i plott av residualene mot de uavhengige variablene, men også i plott mot tilpasset verdi. Her vil man kunne få et inntrykk av om varianasen øker med forventningen.
3. Outliers kan vise seg som residualer med stor absoluttverdi.
4. Hvis normalfordelingsplottet ikke er en rett linje, indikerer det at antagelsen om normalfordelte restledd ikke er berettiget.

I dette tilfellet ser det ikke ut til å være noen systematikk i residualplottene, og normalfordelingsplottet ser også nokså rettlinjet ut. Det er en residual som er noe større enn de andre. Dessuten kan det se ut som om variansen er mindre for høy alder, men ingen av disse er tydelige nok til å endre inntrykket av god tilpasning. Det er dessuten få observasjoner, så eventuelle avvik vil være vanskelige å oppdage.

I tillegg er en de r^2-verdi på 0.97 uttrykk for at en stor del av variasjonen i responsverdiene forklares ved variasjoen i de uavhengige variablene lengde og alder.

b) Et 95% konfidensintervall for koeffesienten for alder β_2 har grenser $\hat{\beta}_2 \pm t_{0.025,6} s_{\hat{\beta}_2}$ der $s_{\hat{\beta}_2}$ er den estimerte standardfeilen til $\hat{\beta}_2$. Den kritiske verdien, $t_{\alpha,n}$, i t-fordelingen er gitt ved $P(T > t_{\alpha,n}) = \alpha$ der den tilfeldige variabelen T er t-fordelt med n frihetsgrader. I dette tilfellet er antallet observasjoner 9 og antallet parametre i den lineære delen 3, slik at antallet frihetsgarder er 6. Fra tabellen er $t_{0.025,6} = 2.447$. Fra utskriften er $\hat{\beta}_2 = 7.1498$ og $s_{\hat{\beta}_2} = 4.0309$ slik at konfidensintervallet har grenser $7.1498 \pm 2.447 \times 4.0309$ som gir konfidensintervallet $(-2.713425, 17.013089)$

c) Her er det rimelig å forkaste nullhypotesen for store verdier av $\hat{\beta}_2$. Siden $t = \hat{\beta}_2 / s_{\hat{\beta}_2}$ er t-fordelt med 6 frihetsgrader under H_0 har en test med nivå 5% forkastningsområde $\{t_{obs} > t_{0.05,6}\}$. Fra tabellen er $t_{0.05,6} = 1.943$ og $t_{obs} = 1.774$, slik at nullhypotesen ikke forkastes. P-verdien i utskriften, 0.126468, gjelder den tosidige testen, og er lik $P(|T_6| > t_{obs})$ der den tilfeldige variabelen T_6 er t-fordelt med 6 frihetsgrader. For den ensidige testen er P-verdien $P(T_6 > t_{obs}) = P(|T_6| > t_{obs})/2$, altså $0.126468/2 = 0.063234$.

Oppgave 2

a) Under alternativet er sannsynligheten for at et av forsøkene er en suksess mindre en 1/2. Man kan derfor forvente at det trengs flere forsøk før første suksess enn tilfellet er hvis suksessannsynligheten er 1/2. Fra formelsamlingen er $E(X) = 1/p$. (Dette følger også av at $E(X) = \sum_{i=1}^{\infty} xp(x-1)^{x-1} = p\sum_{i=1}^{\infty} x(1-p)^{x-1}$. Leddvis derivasjon av $1/p = \sum_{i=0}^{\infty}(1-p)^x$ gir at $-1/p^2 = -\sum_{i=0}^{\infty} x(1-p)^{x-1} = -\sum_{i=1}^{\infty} x(1-p)^{x-1}$. Innsatt gir dette $E(X) = 1/p$.) Siden $1/p$ er avtagende i p, er forventningen er større for verdier under alternativet, slik at store verdier av X tyder på at nullhypotesen er gal og at den skal forkastes. Dette gir forkastningsområde av formen $\{x|x > x_0\}$ der x_0 velges slik at nivåkravet holder.

b) Fra formelen for en geometrisk fordelinger er generelt $P(X \geq x) = \sum_{i=x}^{\infty} p(1-p)^{i-1} = (1-p)^{x-1}$, $x = 1, 2, \ldots$. Derfor er $P(X \geq x_0+1|p = 1/2) = (\frac{1}{2})^{x_0}$. At nivået er 0.10 medfører derfor at $(\frac{1}{2})^{x_0} \leq 0.1$ eller $2^{x_0} \geq 10$, dvs $x_0 = 4$. Vi forkaster derfor hvis antallet forsøk som er nødvendig før første suksess er 5 eller mer. Dette kan også regnes ut direkte $P(X > 3) = 1-P(X \leq 3) = 1-(1/2+1/4+1/8) = 1/8 = 0.125$ og $P(X > 4) = 1-P(X \leq 4) = 1-(1/2+1/4+1/8+1/16) = 1/16 = 0.0625$.

c) Feil av type II består i at nullhypoteser som er gale ikke forkastes. I dette tilfellet betyr det at nullhypotesen ikke forkastes hvis $p < 1/2$. Spesielt vil dette bety for $p = 1/4$ at $P(\text{ikke forkaste } H_0|p = 1/4) = P(X \leq 4|p = 1/4) = 1 - P(X \geq 5|p = 1/4) = 1 - (\frac{3}{4})^4 = 0.6836$.

$P(\text{forkaste } H_0|p) = P(X \geq 5|p) = (1-p)^4$. Dette er en avtagende funksjon i p. Derfor er $(1-p)^4 < (\frac{1}{2})^4 < 0.1$ når $p \geq 1/2$ som viser at testen også har nivå 0.1 for å teste nullhypotesen $H_0 : p \geq \frac{1}{2}$ mot den alternative hypotesen $H_a : p < \frac{1}{2}$.

Oppgave 3

a) Likelihooden til observasjonene kan skrives

$$\prod_{i=1}^{9} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2)$$

slik at logarimen til likelihooden, log L, er

$$\log L = -(n/2)\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{9}(y_i - \beta_0 - \beta_1 x_i)^2$$

Denne størrelsen maksimeres med hensyn på β_0 og β_1 hvis $Q = \sum_{i=1}^{9}(y_i - \beta_0 - \beta_1 x_i)^2$ minimeres. Verdiene som gir maksimum av $\log L$ definererer sannsynlighetsestimatorene og verdiene som minimerer Q minste kvadraters estimatorene, og disse må derfor være de samme.

Derivasjon av $\log L$ med hensyn på σ gir

$$-n/\sigma - \frac{1}{2\sigma^3}(-2)\sum_{i=1}^{9}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Ved å sette inn for sannsynlighetsestimatorene for β_0 og β_1, sette utrykket lik null og løse med hensyn på σ fås at sannsynlighetsestimatoren for σ^2 blir $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{9}(y_i - \beta_0 - \beta_1 x_i)^2$. Fra pensum og formelsamlingen STK1100/STK1110 følger at $n\hat{\sigma}^2/\sigma^2$ er χ^2-fordelt med $n-2$ frihetsgrader. Forventningen til en χ^2-fordelt variabel er lik antall frihetsgrader, dvs at $E(n\hat{\sigma}^2/\sigma^2) = n-2$ eller $E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2$, slik at $\hat{\sigma}^2$ ikke er forventningsrett.

b) Siden $\hat{\beta}_1$ er forventningsrett vil store og små verdier av $\hat{\beta}_1 - 5.0$ tyde på at H_0 er gal og skal forkastes. Det betyr at $\hat{\beta}_1 - 5$ er et rimelig utganngspunkt som testobservator. Fra pensum vet vi at $\hat{\beta}_1$ er $N(\beta, \sigma^2/\sum_{i=1}^{9}(x_i - \bar{x})^2)$-fordelt og $n\hat{\sigma^2}/\sigma^2 = (n-2)S^2/\sigma^2$ er $\chi^2_{(n-2)}$ fordelt. Nå er $n = 9$, slik at $(n-2)S^2/\sigma^2$ er χ^2-fordelt med 7 frihetsgrader. Uttrykk

$$t = \frac{\hat{\beta}_1 - 5}{S/\sqrt{\sum_{i=1}^{9}(x_i - \bar{x})^2)}} = \frac{\frac{\hat{\beta}_1 - 5}{\sigma/\sqrt{\sum_{i=1}^{9}(x_i - \bar{x})^2)}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2}/(n-2)}}.$$

Siden S^2 og $\hat{\beta}_1$ er uavhengige, betyr det at t er t- eller Studentfordelt med 7 frihetsgrader under H_0. Vi forkaster hvis den observerte verdien av t observatoren, t_{obs} enten er mindre enn $-t_{\alpha/2,7}$ eller storre enn $t_{\alpha/2,7}$ der α er nivået.

c) Fra utskiften er den estimerte standardfeilen til $\hat{\beta}_1$, dvs. $s_{\hat{\beta}_1} = S/\sqrt{\sum_{i=1}^{9}(x_i - \bar{x})^2)}$ lik 0.5249 som gir at den observerte verdien av testobservatoren blir $t_{obs} = (5.9967 - 5.0)/0.5249 = 1.842$. Fra tabellen er $t_{0.025,7} = 2.365$ slik at nullhypotesen ikke forkastes.

P-verdien er sannsynligheten for å observere noe mer ekstremt dvs mindre enn -1.842 eller større enn 1.842. Fra tabellen ser man at $P(t_7 > 1.895) = 0.05$ og $P(t_7 > 1.415) = 0.1$. P-verdien ligger derfor mellom 0.1 og 0.2 og like over 0.1.

d) En forventningsrett estimator for forventet vekt av en fisk med lengde $50cm = 500mm$ er $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \times 500$. Fra uttrykkene til $\hat{\beta}_0$ og $\hat{\beta}_1$ er $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \times 500 = \bar{Y} + \hat{\beta}_1(500 - \bar{x}) = \sum_{i=1}^{9} Y_i[1/9 + (x_i - \bar{x})(500 - \bar{x})/S_{xx}]$, der $\bar{x}$ er gjennomsnittet av de ni vektene og $S_{xx} = \sum_{i=1}^{9}(x_i - \bar{x})^2$. Siden observasjonene $Y_1, \ldots, Y_9$ er uavhengige er derfor $\hat{\mu}$ en lineær kombinasjon av uavhengige normalfordelte variable, og derfor selv normalfordelt med varians $V(\hat{\mu}) = \sum_{i=1}^{9} V(Y_i)[1/9 + (x_i - \bar{x})(500 - \bar{x})/S_{xx}]^2 = \sigma^2(1/9 + (500 - \bar{x})^2 \sum_{i=1}^{9}(x_i - \bar{x})^2/S_{xx}^2 + 2(500 - \bar{x}) \sum_{i=1}^{9}(x_i - \bar{x})/S_{xx}) = \sigma^2(1/9 + (500 - \bar{x})^2/S_{xx}) = \sigma_{\hat{\mu}}^2$. Standardfeilen til $\hat{\mu}$ estimeres med $S\sqrt{(1/9 + (500 - \bar{x})^2/S_{xx})}$. Det betyr at $(\hat{\mu} - \mu)/\sigma_{\hat{\mu}}$ er standard normalfordelt og $(\hat{\mu} - \mu)/s_{\hat{\mu}}$ t-fordelt med $n - 2$ frihetsgrader.

Et 95% konfidensintervall har derfor grensen $\hat{\mu} \pm t_{0.025,7} s_{\hat{\mu}}$ der $t_{0.025,7}$ er den kritiske verdien definert ved $P(T > t_{\alpha,k}) = \alpha$ når T er t-fordelt med k frihetsgrader.

Estimatet for μ er $-1828.09 + 5.966 \times 500 = 1155.31$. I dette tilfellet er også $S_{xx} = \sum_{i=1}^{9}(x_i - \bar{x})^2 = \sum_{i=1}^{9}(x_i^2) - 9 \times \bar{x}^2 = 1714916 - 9 \times (3924/9)^2 = 4052$ og $s_{\hat{\mu}} = 33.41\sqrt{1/9 + (500 - (3924/9)^2/4025} = 35.39$. Fra tabellen er $t_{0.025,7} = 2.365$ slik at grensene til konfidensinterlvallet blir $1155.31 \pm 2.365 \times 35.39$ som gir intervallet $(1071.63, 1238.99)$.

e) Et estimat for en ny fisk med lengde 50cm er fortsatt $\hat{Y} = \hat{\mu}$. Da er prediksjonsfeilen $Y - \hat{Y}$. Ved å bruke uttryket for $\hat{\mu}$ fra forrige punkt kan dette skrives $Y - \hat{Y} = Y - \sum_{i=1}^{9} Y_i[1/9 + (x_i - \bar{x})(500 - \bar{x})/S_{xx}]$. Her er $Y, Y_1, \ldots, Y_7$ avhengige normalfordelte variable. Siden $Y - \hat{Y}$ er en lineær kombinasjon av dem, er $Y - \hat{Y}$ også normalfordelt med forventning $E(Y - \hat{Y}) = [\beta_0 + \beta_1 \times 500] - [\beta_0 + \beta_1 \times 500] = 0$. Variablene er uavhengige slik at variansen er $V(Y - \hat{Y}) = V(Y) + V(\hat{Y}) = \sigma^2 + \sigma^2(1/9 + (500 - \bar{x})^2/S_{xx}) = \sigma^2 + \sigma_{\hat{\mu}}^2$. Da er $(Y - \hat{Y})/\sigma\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})}$ standard normalfordelt og $(Y - \hat{Y})/S\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})}$ t-fordelt med 7 frihetsgrader. Derfor er

$$0.95 = P(-t_{0.025,7} < (Y - \hat{Y})/S\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})} < t_{0.025,7}) =$$
$$P(\hat{Y} - t_{0.025,7}S\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})} < Y < \hat{Y} + t_{0.025,7}S\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})}$$

slik at et 95% prediksjonsintervall har grense $\hat{\mu} \pm t_{0.025,7} s\sqrt{1 + (1/9 + (500 - \bar{x})^2/S_{xx})} = \hat{\mu} \pm t_{0.025,7} s\sqrt{1 + (1/9 + s_{\hat{\mu}}^2}$.

Tolkningen er at hvis man gjentatte ganger observerer 9 fisk med samme lengder som i denne undersøkelsen og hver gang beregner

prediksjonsintervall som ovenfor, vil i 95% av tilfellene en tiende fisk som er 50cm lang ha en vekt som ligger innen prediksjonsintervallene.

I dette tilfellet blir grensene $1155.31 \pm 2.365 \times 33.41\sqrt{1 + 1/9 + (500 - (3924/9)^2/4025}$ som gir intervallet $(1040.23, 1270.40)$.

Med stor grad av sikkerhet vil derfor en ny fisk med lengde 50cm veie mellom 1.040 og 1.270 kg.

f) Det som kalles "Multiple R-squared"i utskriften eller r^2 har tolkning som andelen av variasjonen i responsen som forklares ved variasjonen i de uavhengige variable. I dette tilfellet andelen av variasjonen i vekten til fiskene som kan forklares ved variasjonen i lengdene. Tallet 0.9486 viser at en stor slik andel forklares.

r^2 er definert som $1 - SSE/SST$ der SST er den totale variasjonen $SST = \sum_{i=1}^{9}(y_i - \bar{y})^2$ og SSE er den residuale kvadratsumen $SSE = \sum_{i=1}^{9}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{9}(y_i - \hat{y}_i)^2$.

Men $SSE = \sum_{i=1}^{9}(y_i - \hat{y})^2 = \sum_{i=1}^{9}[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)]^2 = \sum_{i=1}^{9}[(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i)]^2 = \sum_{i=1}^{9}[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}))]^2 = \sum_{i=1}^{9}(y_i - \bar{y})^2 - 2\hat{\beta}_1\sum_{i=1}^{9}(y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2\sum_{i=1}^{9}(x_i - \bar{x})^2 = SST - 2\hat{\beta}_1(\hat{\beta}_1 S_{xx}) + \hat{\beta}_1^2 S_{xx} = SST - \hat{\beta}_1^2 S_{xx}$.

Herav følger at $SST - SSE = \hat{\beta}_1^2 S_{xx}$ slik at $r^2 = 1 - SSE/SST = \hat{\beta}_1^2 S_{xx}/S_{yy} = S_{xy}^2/S_{xx}S_{yy}$ siden $\hat{\beta}_1 = \sum_{i=1}^{9} y_i(x_i - \bar{x}))/\sum_{i=1}^{9}(x_i - \bar{x})^2 = S_{xy}/S_{xx} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}$. Men $S_{xy}^2/S_{xx}S_{yy}$ er ikke annet en kvadratet av den empiriske korrlasjonskoeffesienten. Betegner vi den med r_e har vi altså $r^2 = (r_e)^2$.

SLUTT