EKSAMENSOPPGAVER
STAT100 Vår 2011

Løsningsforslag

Oppgave 1 (Med referanse til Tabell 1)

a) De 3 fiskene på 2 år hadde lengder på henholdsvis 48, 46 og 35 cm. Finn de manglende tallene i Tabell 1. Test om forventet lengde på 4 år gamle torsker er mer enn 50 cm. Bruk 5% testnivå på testen.

Tabell 1

alder	mean	sd	n
2	43.0	7.0	3
3	45.2	5.0	9
4	52.2	4.9	50

La X være lengde på 4 år gammel torsk. Antar $X \sim N(\mu_4, \sigma)$
Tester:
H_0: $\mu_4 = 50$ mot H_1: $\mu_4 > 50$

Testobservator blir
$T = \frac{52.2-50.0}{4.9/\sqrt{50}} = 3.175$
Forkaster H_0 hvis $T > t_{0.05,49} \approx t_{0.05,50} = 1.676$
Vi forkaster nullhypotesen og hevder at forventet lengde på 4 år gamle torsker er større enn 50 cm.

b) Test på 5% nivå om forventet lengde på 4 år gamle torsker er større enn for 3 år gamle torsker. Hvilke antagelser gjør du?

La Y være lengde på 3 år gammel torsk. Antar $Y \sim N(\mu_3, \sigma)$
La X være lengde på 4 år gammel torsk. Antar $X \sim N(\mu_4, \sigma)$
Antar samme populasjonsstandardavvik i begge populasjoner.
Tester
H_0: $\mu_4 = \mu_3$ mot H_1: $\mu_4 > \mu_3$

Finner felles standardavvik:

$$S_p = \sqrt{\frac{(9-1)5.0^2 + (50-1)4.9^2}{9+50-2}} = 4.914$$

Testobservator:

$$T = \frac{52.2 - 45.2}{4.914\sqrt{\frac{1}{9} + \frac{1}{50}}} = 3.934$$

Tabellverdi: $t_{0.05,57} \approx t_{0.05,50} = 1.676$
Vi forkaster nullhypotesen og hevder at forventet lengde for 4 åringer er større enn for 3 åringer.

Dersom man antar at også 2 år gamle torsker har samme populasjonsstandardavvik som 3 og 4 åringer, kan det lønne seg å kjøre en ANOVA med tre grupper og en kontrast for å teste forskjell på 3 og 4 åringer. En R-utskrift fra ANOVA blir:

```
Df Sum Sq Mean Sq F value   Pr(>F)
alder        2  554.5 277.249  11.006 8.67e-05 ***
Residuals   59 1486.3  25.191
```

Med kontrastanalyse i R:

```
                   Estimate  Std.Error t value  Pr(>|t|)
alder c=( 0 -1 1 ) 6.937778   1.817370 3.817482 0.000325
```

Da blir estimat for felles standardavvik lik $S_p = \hat{\sigma} = \sqrt{MS_E} = \sqrt{25.191} = 5.019$
Testobservator for konstrasten blir 3.82 som gir forkastning av hypotese om at forventningene er like for 3 og 4 åringer.

Oppgave 2

En regresjonsmodell med Y=lengde og X=alder på torsk ble tilpasset i R Commander. Resultatet av kjøringen er gitt i Tabell 2 nedenfor og i form av Figur 1. Bruk resultatene i den grad du finner det nødvendig for å besvare spørsmålene a) – c). (Merk at noen tall er utelatt og erstattet med en stjerne i Tabell 2.)

Tabell 2.

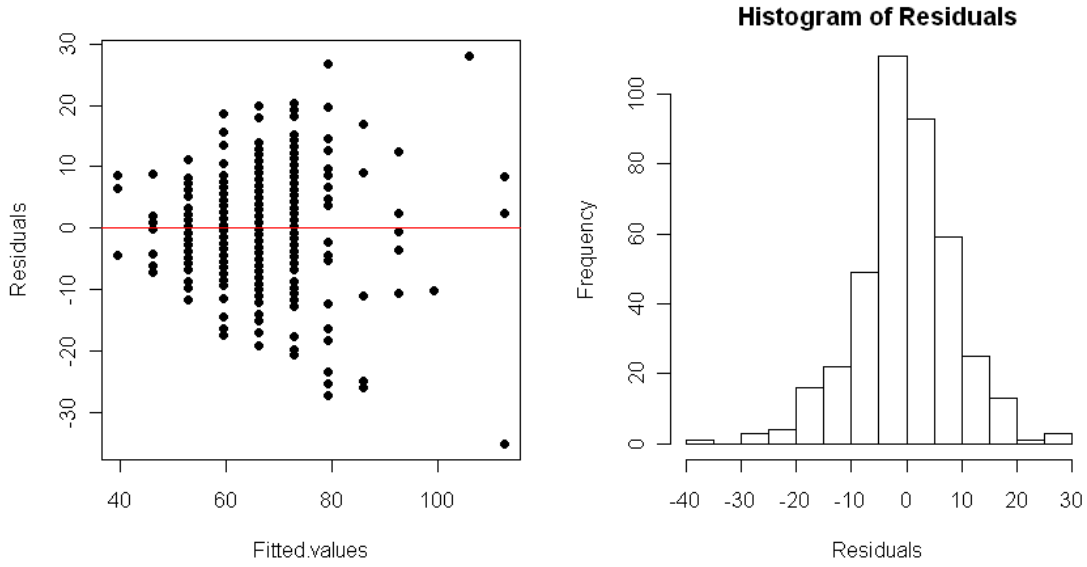
```
Call:
lm.default(formula = lengde ~ alder, data = Fisk)

Residuals:
    Min      1Q  Median      3Q     Max
-35.100  -4.437  -0.066   5.240  28.066

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1973     1.7663   14.83
alder         6.6447     0.2915      *

Residual standard error: 8.865 on 398 degrees of freedom
Multiple R-squared: *,   Adjusted R-squared: 0.5651
F-statistic: 519.5 on 1 and 398 DF,  p-value: *

Analysis of Variance Table
           Df Sum Sq Mean Sq F value    Pr(>F)
alder       1  40828   40828  519.54 < 2.2e-16
Residuals 398  31277      79
Total     399      *
```



Figur 1. Residualer plottet mot tilpassede responsverdier (venstre) og histogram over residualer (høyre)

a) Test på 1% nivå om det er en positiv sammenheng mellom alder og lengde på torsker. Finn determinasjonskoeffisienten R^2 og gi en tolkning av denne.

Modellen som er tilpasset er en enkel regresjonsmodell
$Y_i = \alpha + \beta X_i + \epsilon_i$

Tester:
H_0: $\beta = 0$ mot H_1: $\beta > 0$

Testobservator:

$$T = \frac{6.6447}{0.2915} = 22.79$$

Forkaster dersom $T > t_{0.01,398} \approx t_{0.01,100} = 2.36$
Vi forkaster nullhypotesen og hevder det er en positiv sammenheng mellom alder og lengde på torsk.

$R^2 = SS_R/(SS_R + SS_E) = 40828/(40828 + 31277) = 0.5662$, dvs
56% av variasjonen i fiskelengder kan forklares ved den lineære sammenhengen til alder.

b) Hvilke antagelser ligger til grunn for regresjonsmodellen? Vurder om antagelsene ser ut til å være oppfylte.

Antagelser og sjekk
- Lineært forhold mellom Y og X: Residualplottet viser ingen tegn til ikke-lineæritet da residualene ligger jevnt spredt rundt 0 for alle verdier av tilpasset respons. (Antagelse om lineæritet OK)

- Konstant varians: Residualplottet viser tegn til økende varians med økende tilpasset Y. (Brudd på antagelsen om konstant varians)
- Normalfordeling: Histogrammet viser at fordelingen til residualene er rimelig klokkeformet, men med antydning til tung hale mot venstre (Likevel synes normalforedelingsantagelsen å være OK)
- Antagelsen om uavhengige residualer: Kan vi sjekkes fra den tilgjengelige utskriften.

c) Finn et 95% konfidensintervall for forventet lengde på 6 år gammel fisk (Tilleggsopplysning: $\bar{x} = 5.87$)

Formel:

$$\hat{y} \pm t_{0.025,398} \cdot s\sqrt{\frac{1}{n} + \left(\frac{x-\bar{x}}{s/SE(\hat{\beta})}\right)^2}$$

Setter inn følgende som kan finnes fra utskriften.

$\hat{y} = 26.1973 + 6.6447 \cdot 6 = 66.065$
$t_{0.025,398} \approx t_{0.025,100} = 1.984$
$s = 8.865$
n=400
x=6
$\bar{x} = 5.87$
$SE\hat{\beta} = 0.2915$

Finner da et 95% konfidensintervall lik [**65.18, 66.95**]. Dette intervallet vil med 95% sannsynlighet dekke den sanne forventede lengden på 6 år gammel torsk.

Oppgave 3

Torskepopulasjonen varierer fra sesong til sesong, og vi skal videre anta at observasjonene gjort i ulike sesonger er uavhengige av hverandre. Nedenfor finner du en R utskrift som kan brukes til å besvare spørsmålene i oppgave 3.

a) Skriv opp modellen som ligger til grunn for analysen i Tabell 3, fyll inn de manglende verdiene markert med tallene (1) – (5) i ANOVA-tabellen og estimér alle ukjente parametre i modellen.

Tabell 3.

```
Anova Table

Response: lengde
           Df    SS    MS  F-value
sesong      3  3493  1164.33    6.7204
Residuals 396  68612  173.26
Total     399  72105


sesong    mean        sd        n
1         68.07      12.1      100
2         67.48      12.7      100
3         64.51      15.6      100
4         60.61      11.9      100
```

b) Sett opp hypoteser for å teste om det er forskjell i forventet fiskelengde mellom de ulike sesongene og utfør testen.

La Y_{ij} være lengde målt på fisk nr *j* (*j=1,2,…,*n_i) i sesong nr *i* (*i=1,2,3,4*).
Antar modellen
$Y_{ij} = \mu_i + \epsilon_{ij}$
Der støyleddene antas uavhengige av hverandre og $\epsilon_{ij} \sim N(0, \sigma)$

Manglende verdier, se utskriften over
Forventningsestimater, $\hat{\mu}_i$ for gruppene er gitt ved utvalgsstandardavvikene i Tabell 3.
Estimat for σ er $s = \hat{\sigma} = \sqrt{MS_E} = \sqrt{173.26} = 13.16$

b) Sett opp hypoteser for å teste om det er forskjell i forventet fiskelengde mellom sesongene og utfør testen på 5% nivå.

Hypoteser:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ mot H_1: Minst to forventninger er ulike

Testobservator fra utskriften blir F=6.7204.
Tabellverdi: $F_{0.05,3,396} \approx F_{0.05,3,120} = 2.68$
(Nærmeste mindre tall i F-tabellen for n-k frihetsgrader er 120)
Forkaster nullhypotese om at forventet fiskelengde er den samme i alle sesonger siden testobservatoren er større enn 2.68.

c) Sesong 1 og 2 er fra to år på 1990-tallet, mens sesong 3 og 4 er fra to år på 2000-tallet. Test om det har skjedd en generell endring i forventet fiskelengde mellom de to tiårene.

Setter opp følgende kontrast for å teste dette:
$\theta = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4)$
Tester
$H_0 : \theta = 0$ mot $H_1 : \theta \neq 0$

Estimeres ved å sette inn estimatene for gruppeforventningene og vi finner
$\hat{\theta} = 5.213$
Finner standardfeilen til denne ved

$$SE(\hat{\theta}) = \sqrt{MSE \sum_{i=1}^{4} \frac{c_i^2}{n_i}} = \sqrt{173.26 \cdot 4 \frac{(1/2)^2}{100}} = 1.316$$

Dette gir testobservatoren
$T = 5.213/1.316 = 3.96$
Forkaster nullhypotesen på 5% nivå dersom
$|T| > t_{0.025,396} \approx t_{0.025,100} = 1.984$
Vi forkaster H_0 og hevder at forventet torskelengde er forskjellig i de to tiårene.

d) Finn et 95% konfidensintervall for forventet forskjell i fiskelengde mellom sesong 3 og 4. Bruk intervallet til å teste om det er signifikant forskjell mellom de to sesongene på 5% testnivå.

Konstruerer en kontrast mellom sesong 3 og 4 og finner et konfidensintervall for denne.
$\theta = \mu_3 - \mu_4$
Tester:
$H_0 : \theta = 0$ mot $H_1 : \theta \neq 0$

Denne estimeres til
$\hat{\theta} = 3.895$
Med standardfeil
$SE(\hat{\theta}) = 1.862$

Et 95% konfidensintervall for θ blir
$[\hat{\theta} \pm t_{0.025,396} SE(\hat{\theta})] \approx [3.895 \pm 1.984 \cdot 1.862] = [0.20, 7.59]$

Siden $\theta = 0$ ikke ligger i dette intervallet kan vi ved testnivå 5% forkaste nullhypotesen og hevde at det forskjell på forventet fiskelengde i sesong 3 og 4.

e) Anta at utvalgsgjennomsnittene og utvalgsstandardavvikene for de fire sesongene er som gitt i Tabell 3, men at det bare var 31 observasjoner i hver sesong (altså ikke 100 fra hver sesong). Kan du påvise signifikante forskjeller mellom sesongene på 5% nivå da? Hva med 10% testnivå?

Må finne verdier av MS_G , MS_E og testobservatoren F:
$SS_G = \sum_{i=1}^{4} n_i(\bar{Y}_i - \bar{Y})^2 = 31(68.07 - 65.17)^2 + \cdots + 31(60.61 - 65.17)^2 = 1084.23$
$MS_G = SS_G/(k-1) = 1084.23/(4-1) = 361.41$
$MS_E = S^2_{pooled} = 173.26$ (som før siden det fortsatt er like mange i hver gruppe vil denne være uforandret, men kan altså finnes ved

$$MS_E = S^2_{pooled} = \frac{(31-1)12.1^2 + (31-1)12.7^2 + (31-1)15.6^2 + (31-1)11.9^2}{(31+31+31+31)-4}$$

F = 361.41/173.26 = 2.086
$F_{0.05,\,3,\,120} = 2.68$ (Ikke forkastning)
$F_{0.1,\,3,\,120} = 2.13$ (Ikke forkastning, men som vi ser vil vi få forkastning for testnivå som er marginalt større enn 10%)