

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 – Statistiske metoder og dataanalyse

Eksamensdag: Onsdag 4. desember, 2024

Tid for eksamen: 15.00 – 19.00.

Oppgavesettet er på 4 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Godkjent kalkulator  
Formelsamling for STK1110

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

### Oppgave 1

Anta at  $X_1, X_2, \dots, X_n$  er uavhengige og identisk fordelte stokastiske variable med tetthet  $f(x; \alpha) = \alpha x^{\alpha-1}$  for  $0 < x < 1$  (altså over intervallet  $[0, 1]$ ) og der  $\alpha > 0$ .

- (a) Man kan lett vise at  $E(X_i) = \frac{\alpha}{\alpha+1}$  (det skal du ikke gjøre).

Finn på denne bakgrunn en momentestimator for parameteren  $\alpha$ .

- (b) Men tilsvarende har man også at  $Y_{ri} = X_i^r$  er uavhengige og identisk fordelte med forventning  $E(Y_{ri}) = \frac{\alpha}{\alpha+r}$  (for vilkårlig  $r > -\alpha$ . Det skal du heller ikke vise).

Finn momentestimatorer for  $\alpha$  basert på  $Y_{ri}$ -ene.

Kommenter på en svakhet ved momentestimering knyttet til dette.

- (c) Vis at maximum likelihood estimatoren (MLE) for  $\alpha$  er gitt ved  $\hat{\alpha} = \frac{n}{-\sum_{i=1}^n \ln(X_i)}$

- (d) Angi uten bevis tilnærmet fordeling for MLE  $\hat{\alpha}$  når  $n$  er tilstrekkelig stor.

Vis at variansen i denne tilnærmede fordelingen gis ved  $\frac{\alpha^2}{n}$ .

- (e) La  $Z = \frac{\hat{\alpha} - \alpha}{\sqrt{\hat{\alpha}/n}}$  og merk at  $Z$  da er tilnærmet standard normalfordelt når  $n$  er stor.

Utled et tilnærmet 95% konfidensintervall for  $\alpha$  basert på tilnærmet fordeling for  $Z$ .

Beregn intervallet når  $n = 20$  og  $\sum_{i=1}^n \ln(X_i) = -9.82$ .

(Fortsettes på side 2.)

- (f) Merk at hvis  $\alpha = 1$  så er  $X_i$ -ene uniformt fordelt på  $[0, 1]$ . Benytt intervallet i (e) til å teste nullhypotesen om at  $X_i$ -ene er uniformt fordelt mot alternativet at de ikke er det med nivå 5%.

Finn også P-verdien for tilsvarende test. P-verdien må her angis ved et intervall den må ligge innenfor bestemt av kvantiler  $q_\gamma$  i standardnormalfordelingen i tabellen under som angitt  $P(Z \leq q_\gamma) = \gamma$  når  $Z \sim N(0, 1)$ .

Sannsynlighet $\gamma$	0.001	0.005	0.010	0.015	0.020	0.025	0.050	0.100	0.200
Kvantil $q_\gamma$	-3.09	-2.58	-2.33	-2.17	-2.05	-1.96	-1.64	-1.28	-0.84

## Oppgave 2

En enkel lineær regresjonsmodell er spesifisert ved at responser  $Y_i, i = 1, \dots, n$  er uavhengige og normalfordelte med samme varians  $\sigma^2$  og med forventninger  $E(Y_i) = \beta_0 + \beta_1 x_i$  der  $x_i$  er kjente forklaringsvariable.

- (a) Minste kvadraters estimatoren for  $\beta_1$  kan gis ved

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

der  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  og  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  (Dette skal du ikke vise).

Vis at  $\hat{\beta}_1$  er forventningsrett for  $\beta_1$ .

Alternativ kan man skrive opp  $\hat{\beta}_1$  på formen (det skal du heller ikke vise)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Utled fra denne formelen at variansen til  $\hat{\beta}_1$  er gitt ved uttrykket

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- (b) I en studie av 600 barnehagebarn i Oslo registrerte man  $x_i$  = barnas høyde (i cm) samt deres lungefunksjon  $Y_i$  via målet "forced vital capacity" (FVC, som er mengden luft en person klarer å blåse ut, her målt i liter). Den (litt editerte) utskriften på neste side gir resultater fra en enkel lineær regresjon med respons  $Y_i$  og forklaringsvariabel  $x_i$ .

Angi og fortolk estimatene for modellens parametre.

Angi en formel for t-verdien knyttet til høyde og forklar hvilken nullhypotese og hvilken alternativ hypotese dette er en testobservator for.

Hva er fordelingen til testobservatoren under nullhypotesen?

Formuler en konklusjon på testen.

(Fortsettes på side 3.)

Call:

```
lm(formula = fvc ~ hoyde, data = lunge)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.293767	0.115758	-19.82	<2e-16 ***
hoyde	0.031443	0.001057	29.74	<2e-16 ***
---				

Residual standard error: 0.197 on 598 degrees of freedom

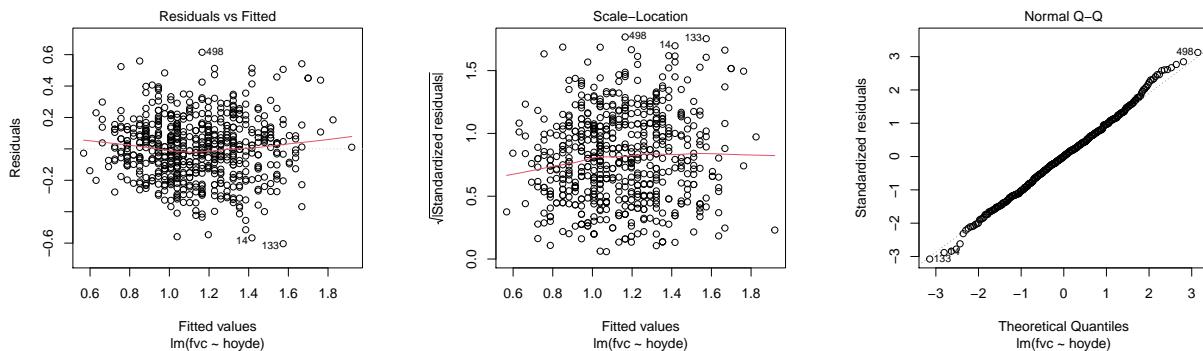
Multiple R-squared: 0.5966, Adjusted R-squared: 0.5959

F-statistic: 884.5 on 1 and 598 DF, p-value: < 2.2e-16

- (c) Definer tilpassede (predikerte) verdier  $\hat{y}_i$  og residualer  $e_i$ .

Diskuter hvordan  $\hat{y}_i$  og  $e_i$  kan benyttes til å sjekke modellforutsetninger grafisk i lineære regresjonsmodeller. Du kan i denne sammenheng gjøre bruk av residualplott fra den enkle lineære regresjonsmodellen i punkt (b). (Ignorer at to av plottene benytter standardiserte residualer i stedet for vanlige residualer  $e_i$ ).

Konkluder om modellen som ble brukt synes å passe med data. Begrunn svaret.



- (d) En multippel lineær regresjonsmodell med to forklaringsvariable  $x_{i1}$  og  $x_{i2}$  gis ved følgende modell for responsen  $Y_i, i = 1, \dots, n$ :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

der  $\varepsilon_i \sim N(0, \sigma)$  er uavhengige og  $\beta_j$  er regesjonsparametre.

Regresjonsparametrene estimeres vanligvis ved å minimere kvadratsummen  $g(b_0, b_1, b_2) = \sum_{i=1}^n (Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$ . Vis at en finner estimatene  $\hat{\beta}_0, \hat{\beta}_1$  og  $\hat{\beta}_2$  ved å løse tre lineære ligninger med tre ukjente.

Sett opp disse ligningene på komponentform og skriv så dette om til matriseform.

Utled løsningen for  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$  på matriseform.

Angi betingelsen for at det eksisterer en entydig løsning.

(Fortsettes på side 4.)

- (e) I en videre analyse av barnehagebarnas lungekapasitet  $Y_i$  er det benyttet multippel lineær regresjon med to forklaringsvariable  $x_{i1}$  = høyde (i cm) og  $x_{i2}$  = vekt (i kg). Resultater fra analysen er gjengitt i utskrift fra R under.

Gi en (kvalitativ) forklaring på endringen i estimert regresjonskoeffisient for høyde sammenlignet med estimatet fra punkt (b). Du kan benytte at  $x_{i1}$  og  $x_{i2}$  er positivt korrelert i disse dataene.

Angi også et uttrykk for størrelsen **Multiple R-squared** og forklar hvorfor denne kan fortolkes som andel forklart variasjon (evt. varians).

Sammenlign med den tilsvarende størrelsen i den enkle lineære regresjonen. Kommenter.

For enkel lineær regresjon er størrelsen  $R^2$  knyttet til korrelasjonskoeffisienten mellom respons og forklaringsvariabel. Angi eksakt hvordan.

Uttrykk en tilsvarende sammenheng ved den multiple lineære regresjonen.

Call:

```
lm(formula = fvc ~ hoyde + vekt, data = lunge)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.037726	0.122648	-16.614	< 2e-16 ***
hoyde	0.026209	0.001416	18.504	< 2e-16 ***
vekt	0.017394	0.003220	5.403	9.49e-08 ***
<hr/>				

Residual standard error: 0.1925 on 597 degrees of freedom

Multiple R-squared: 0.6154, Adjusted R-squared: 0.6141

F-statistic: 477.7 on 2 and 597 DF, p-value: < 2.2e-16

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensversjonen er: STK1110 – Statistiske metoder og dataanalyse  
Korrigert versjon

Eksamensdag: 28. november - 2023

Tid for eksamen: 15.00 – 19.00.

Oppgavesettet er på 5 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Godkjent kalkulator  
Formelsamling for STK1110

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

Tabell over øvre kvantiler for Normal, noen T-fordelinger og noen F-fordelinger:

Fordeling	Kvantiler				
	0.5	0.05	0.025	0.01	0.005
Normal	0.000	1.645	1.960	2.326	2.576
$T_{155}$	0.000	1.655	1.975	2.351	2.608
$T_{154}$	0.000	1.655	1.975	2.351	2.608
$T_{30}$	0.000	1.697	2.042	2.457	2.750
$T_{29}$	0.000	1.699	2.045	2.462	2.756
$T_{28}$	0.000	1.701	2.048	2.467	2.763
$T_{27}$	0.000	1.703	2.052	2.473	2.771
$T_{24.569}$	0.000	1.709	2.061	2.488	2.791
$T_3$	0.000	2.353	3.182	4.541	5.841
$T_2$	0.000	2.920	4.303	6.965	9.925
$T_1$	0.000	6.314	12.706	31.821	63.657
$F_{13,13}$	1.000	2.577	3.115	3.905	4.573
$F_{14,14}$	1.000	2.484	2.979	3.698	4.299
$F_{15,15}$	1.000	2.403	2.862	3.522	4.070

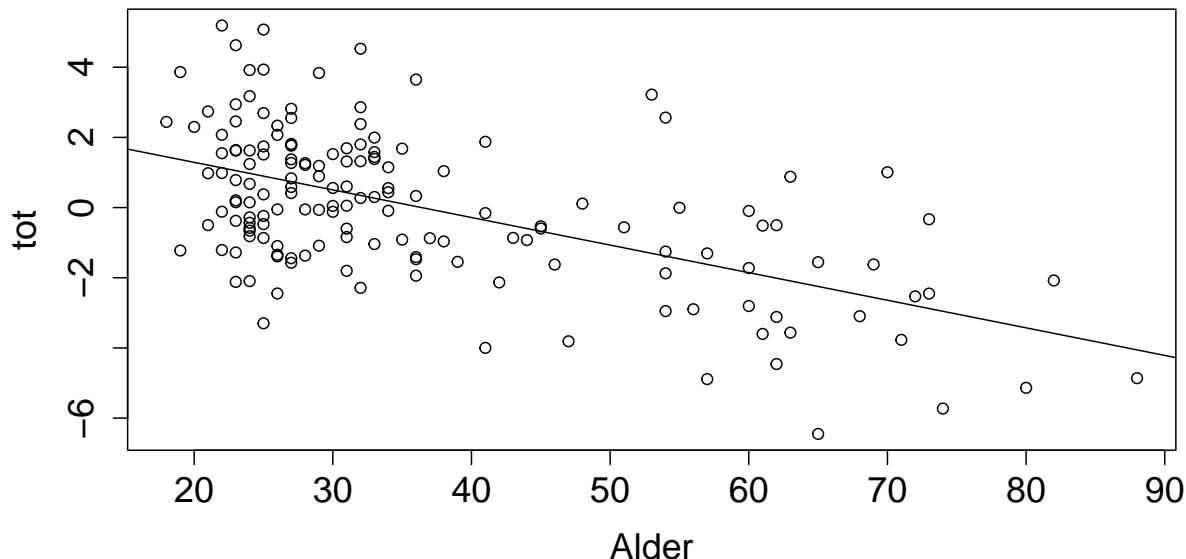
### Oppgave 1

Plottet nedenfor viser data fra et studie på funksjonalitet av nyrer på 157 frivillige individer. Individenes alder er gitt på  $x$ -aksen mens  $y$ -aksen gir et sammensatt mål  $\text{tot}$  for generell funksjon. Funksjonen avtar generelt med alder, noe som tydelig sees av plottet. Den rette linjen viser en tilpasning basert på lineær regresjon:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

der  $\hat{\beta}_0, \hat{\beta}_1$  er beregnet utifra minste kvadraters prinsippet. Her er  $x = \text{Alder}$  og  $Y = \text{tot}$ .

(Fortsettes på side 2.)



- (a) Utskriften nedenfor viser oppsummering av tilpasningen:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.860673	0.359561	7.956	3.49e-13
Alder	-0.078601	0.009056	-8.680	5.14e-15

Residual standard error: 1.801 on 155 degrees of freedom

Multiple R-squared: 0.3271, Adjusted R-squared: 0.3227

F-statistic: 75.34 on 1 and 155 DF, p-value: 5.137e-15

Forklar hva de ulike delene i denne utskriften beskriver.

- (b) Gir utskriften ovenfor indikasjon på at Alder er en viktig forklaringsvariabel? Begrunn svaret.  
Lag et 95% konfidensintervall for  $\beta_1$ .  
Kommentér også verdien på  $R^2$ .
- (c) En alternativ modell er å ta med et kvadratisk ledd for alder. Utskriften nedenfor gir en oppsummering for tilpasning en modell

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.5867806	1.1051749	2.341	0.0205

(Fortsettes på side 3.)

Alder	-0.0644796	0.0546218	-1.180	0.2396
Alder^2	-0.0001523	0.0005808	-0.262	0.7935

Residual standard error: 1.806 on 154 degrees of freedom  
 Multiple R-squared: 0.3274, Adjusted R-squared: 0.3186  
 F-statistic: 37.48 on 2 and 154 DF, p-value: 5.477e-14

Du får her også oppgitt at estimert korrelasjon mellom  $\hat{\beta}_1$  og  $\hat{\beta}_2$  er -0.986.

Prøv å forklare hvorfor tilsvynelatende hverken Alder eller Alder<sup>2</sup> er signifikant forskjellige fra null, men at man likevel får en svært signifikant P-verdi i siste linje av utskriften.

- (d) Anta nå vi ønsker å predikere responsen tot for Alder=20 og Alder=90. Tabellen nedenfor viser prediksjonsestimatorer med 95% prediksjonsintervaller for modellen med kun lineært ledd (Modell 1) og modellen med kvadratisk ledd (Modell 2). Her er lwr og upr nedre og øvre grenser, henholdsvis.

Forklar hvorfor intervallene er ganske brede selv om usikkerheten i  $\hat{\beta}_j$  estimatene er ganske små.

Forklar hvorfor det er rimelig at resultatene er ganske like for Alder=20 mens de er mer forskjellige for Alder=90.

Modell	Alder	Estimat	lwr	upr
1	20	1.289	-2.292	4.870
1	90	-3.427	-7.081	0.226
2	20	1.236	-2.377	4.850
2	90	-3.546	-7.318	0.226

## Oppgave 2

Anta vi har to uavhengige stokastiske variable,  $X$  og  $Y$  der

$$X \sim \text{Gamma}(\alpha, \beta);$$

$$Y \sim \text{Gamma}(3\alpha, \beta).$$

Anta  $\alpha$  er kjent, mens  $\beta$  er ukjent.

- (a) Sett opp likelihoodfunksjonen for  $\beta$  basert på observasjonen  $X$  og  $Y$  og vis at log-likelihood funksjone (der noen ledd som ikke avhenger av  $\beta$  er fjernet) er

$$\ell(\beta) = -4\alpha \log(\beta) - \frac{1}{\beta}(x + y).$$

Merk: I ditt likelihood uttrykk skal alle ledd være med!

- (b) Utled en formel for maksimum likelihood estimatet for  $\beta$ ,  $\hat{\beta}_{ML}$ .

Beregn forventning og varians for ML estimatet til  $\hat{\beta}_{ML}$ .

(Fortsettes på side 4.)

- (c) Hvis  $\alpha$  er et stort heltall, argumenter hvorfor  $\hat{\beta}_{ML}$  er tilnærmet normalfordelt.

Hvis  $\alpha = 20$ ,  $x = 12.907$  og  $y = 26.863$ , beregn en tilnærmet verdi for variansen til  $\hat{\beta}$ .

## Oppgave 3

Tabellen nedenfor viser årlig skillsmisserate for noen europeiske og asiatiske land (FN, 1994). Vi ønsker å teste om det er forskjell i skillsmissemisjonsfrekvensen mellom de to verdensdelene representert ved disse spesifikke landene.

Europa	0.80	2.10	1.30	1.90	2.70	2.50	2.40	1.10
	2.90	1.90	0.90	2.40	2.20	2.80	0.60	
Asia	1.30	1.60	1.20	0.80	1.00	0.80	1.30	1.40
	1.60	1.90	1.50	0.80	2.80	0.90	1.30	

Vi har også følgende oppsummerende mål:

	Mean	$S^2$
Europa	1.900	0.596
Asia	1.347	0.273

- (a) Anta dataene er uavhengige og normalfordelte med henholdsvis  $N(\mu_1, \sigma_1^2)$  for data fra Europa og  $N(\mu_2, \sigma_2^2)$  for data fra Asia.

Utfør en test for å sjekke om variansene i de to utvalg er forskjellige. Formuler en konklusjon på testen.

- (b) Basert på formler i formelsamlingen, *utled* hvordan konfidensintervall for  $\mu_1 - \mu_2$  generelt ser ut.

Beregn både et 95% konfidensintervall og et 99% konfidensintervall for  $\mu_1 - \mu_2$ .

Spesifiser hvilken metode du bruker for å lage disse konfidensintervallene og hvilke ekstra antagelser du gjør.

- (c) En kan vise at P-verdien for en test av hypotesen  $H_0 : \mu_1 = \mu_2$  mot alternativet  $H_a : \mu_1 \neq \mu_2$ , basert på normalfordelingsantagelsen, ligger mellom 0.01 og 0.05.

Hvordan vil du konkludere med hensyn på å teste de to hypotesene mot hverandre?

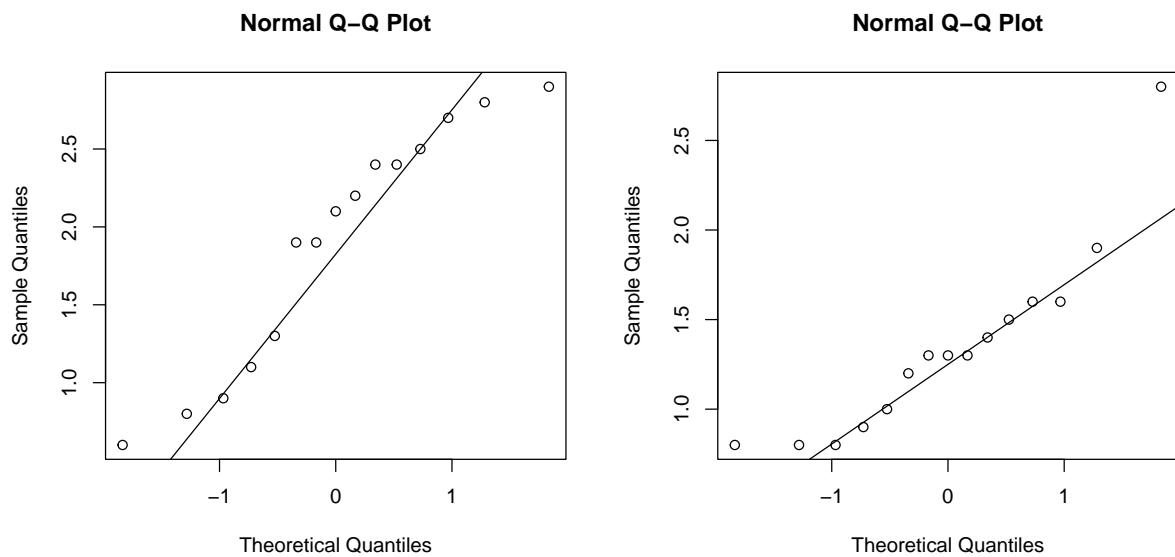
I forhold til konfidensintervallene du fikk i (b), virker det rimelig at P-verdien ligger i intervallet [0.01, 0.05]?

- (d) I kurset har vi lært om også lært om Wilcoxon signed rank-test for ett utvalg. Denne metoden kan imidlertid også brukes for testing av forskjell mellom to utvalg (detaljene er ikke viktige her, bortsett fra at denne utvidelsen også bygger på at fordelingene er symmetriske). For det konkrete datasettet får vi da en P-verdi på 0.061.

Hvorfor er det rimelig at man nå får en høyere P-verdi?

(Fortsettes på side 5.)

Plottene nedenfor viser normal kvantilplot for data fra Europa (venstre) og Asia (høyre). Basert på disse plottene, hvilken test vil du foretrekke å bruke?



SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder og dataanalyse

Eksamensdag: Torsdag 8. desember 2022

Tid for eksamen: 09.00–13.00

Oppgavesettet er på 6 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Godkjent kalkulator  
Formelsamling for STK1110

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

Nedenfor er det gitt kritiske verdier  $t_{\alpha,\nu}$  for t-fordelingen med  $\nu$  frihetsgrader for noen verdier av  $\alpha$  og  $\nu$ . Du vil få bruk for tabellen i Oppgave 1 og 3.

$\alpha :$	0.10	0.05	0.025	0.0125	0.01	0.005
$t_{\alpha,13}$	1.350	1.771	2.160	2.533	2.650	3.012
$t_{\alpha,14}$	1.345	1.761	2.145	2.510	2.624	2.977
$t_{\alpha,15}$	1.341	1.753	2.131	2.490	2.602	2.947
$t_{\alpha,16}$	1.337	1.746	2.120	2.473	2.583	2.921
$t_{\alpha,17}$	1.333	1.740	2.110	2.458	2.567	2.898
$t_{\alpha,18}$	1.330	1.734	2.101	2.445	2.552	2.878

### Oppgave 1

Under er det vist gjennomsnittlig sjøtemperatur i aprilmåned, målt utenfor Ekofisk-feltet i løpet av 15 år i perioden 1995 til 2022 (målingene er oppgitt i  $^{\circ}C$ ):

6.6 4.6 6.3 7.1 6.8 5.8 6.5 6.9 4.8 7.9 7.7 7.4 7.5 6.4 7.2

La  $X_i$  være  $i$ -te måling av temperaturen. Det antas at  $X_1, \dots, X_{15} \stackrel{uif}{\sim} N(\mu, \sigma^2)$  (målingene av temperaturen er gjort med nokså lange mellomrom, slik at antakelsen om uavhengighet er rimelig).

#### a

- Utled et 95% konfidensintervall for forventet sjøtemperatur.
- Beregn intervallet for dataene over, når du får vite at observert snitt og standardavvik er  $\bar{x} = (1/15) \sum_{i=1}^{15} x_i = 6.63$  og  $s = \sqrt{(1/14) \sum_{i=1}^{15} (x_i - \bar{x})^2} = 0.967$ .

(Fortsettes på side 2.)

Temperaturen har tidligere vært på gjennomsnittlig  $6.2^\circ C$ , men forskere tror den kan ha endret seg. De vil derfor teste hypotesene

$$H_0 : \mu = 6.2 \text{ mot } H_a : \mu \neq 6.2.$$

**b**

- Utled en hypotesetest for disse hypotesene med signifikansnivå 5%.
- Hva blir konklusjonen ut fra de observasjonene som ble gjort?

**c**

- Forklar hva en P-verdi generelt betyr, altså definér hva den er.
- Finn et uttrykk for P-verdien til testen i Oppgave **b**.
- Bruk tabellen over kritiske verdier for t-fordelingen til å si noe om størrelsesordenen for P-verdien.

**Oppgave 2**

La  $X_1, \dots, X_n$  være uavhengige stokastiske variabler fra  $\text{Gamma}(4, \beta)$ -fordelingen, dvs. med sannsynlighetstetthet

$$f(x; \beta) = \frac{1}{6\beta^4} x^3 e^{-\frac{x}{\beta}}, \quad x > 0,$$

med  $\beta > 0$  som ukjent parameter. Anta så at vi har observasjoner  $x_1, \dots, x_n$  av disse.

**a**

Sett opp likelihood- og log-likelihood-funksjonen, og vis at maksimum likelihood-estimatoren blir  $\hat{\beta}_{ML} = \frac{1}{4}\bar{X}$ .

**b**

Vis at Fisher-informasjonen i én observasjon er  $I(\beta) = \frac{4}{\beta^2}$ .

**c**

Begrunn at  $\hat{\beta}_{ML}$  er tilnærmet normalfordelt  $N(\beta, \sigma_{\hat{\beta}}^2)$ -fordelt for store  $n$ , og finn et uttrykk for  $\sigma_{\hat{\beta}}^2$ .

La nå  $\psi = \frac{1}{\beta}$ .

**d**

Begrunn at maksimum likelihood-estimatoren for  $\psi$  er  $\hat{\psi}_{ML} = 4/\bar{X}$ .

Vi ønsker nå i stedet å bruke Bayes-estimatoren til å estimere  $\psi$ , og spesifiserer derfor apriorifordelingen  $\psi \sim \text{Gamma}(\alpha_0, \beta_0)$ . Sannsynlighetstettheten til hvert enkelt datapunkt kan da reparametrises til

$$f(x|\psi) = \frac{\psi^4}{6}x^3e^{-\psi x}, \quad x > 0.$$

#### e

- Hva vil det si at en apriorifordeling er konjugert med fordelingen til dataene?
- Vis at apriorifordelingen til  $\psi$  er konjugert med fordelingen til  $X_1, \dots, X_n|\psi$ , og nærmere bestemt at  $[\psi|X_1 = x_1, \dots, X_n = x_n] \sim \text{Gamma}\left(\alpha_0 + 4n, \frac{1}{\beta_0 + \sum_{i=1}^n x_i}\right)$ .
- Finn Bayes-estimatoren for  $\psi$ , og sammenlikn denne med maksimum likelihood-estimatoren fra Oppgave d.

## Oppgave 3

En ønsker å undersøke hvordan nivået på årlige importen til et land avhenger av det årlige forbruket. Dataene nedenfor viser årlig import og forbruk i Frankrike i millioner FRF (franske franc) i løpet av de 18 årene fra 1949 til 1966:

```
import forbruk
15.9  108.1
16.4  114.8
19.0  123.2
19.1  126.9
18.8  132.1
20.4  137.7
22.7  146.0
26.5  154.1
28.1  162.3
27.6  164.3
26.3  167.6
31.1  176.8
33.3  186.6
37.0  199.7
43.3  213.9
49.0  223.8
50.3  232.0
56.6  242.9
```

Vi tilpasser først en enkel lineær regresjonsmodell med import som responsvariabel  $Y$  og forbruk som forklaringsvariabel  $x_1$ , dvs.

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \epsilon_i, \quad i = 1, \dots, 18, \quad (1)$$

der  $\bar{x}_1 = (1/18) \sum_{i=1}^{18} x_{i1}$ , og vi antar at  $\epsilon_1, \dots, \epsilon_{18} \stackrel{\text{uif}}{\sim} N(0, \sigma^2)$ . Resultatet av denne analysen er gitt i R -utskriften på neste side.

(Fortsettes på side 4.)

R-utskrift fra tilpasning til modell (1):

Call:

```
lm(formula = import ~ I(forbruk - mean(forbruk)))
```

Residuals:

Min	1Q	Median	3Q	Max
-3.8435	-1.4407	-0.5032	1.6780	4.1982

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
(Intercept)	30.07778	0.52732	57.04 < 2e-16 ***
I(forbruk - mean(forbruk))	0.29559	0.01305	22.65 1.39e-13 ***

---Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.237 on 16 degrees of freedom

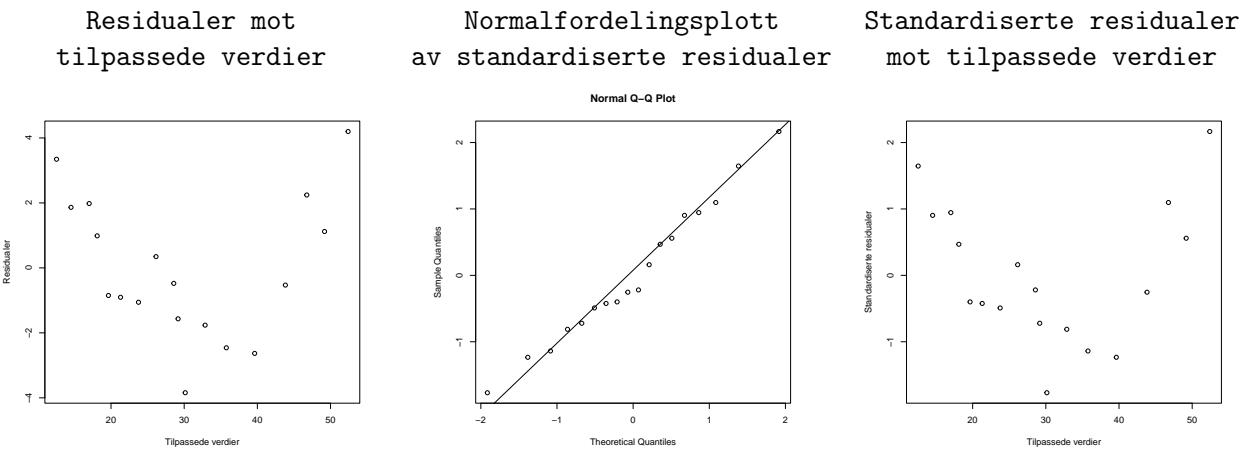
Multiple R-squared: 0.9698, Adjusted R-squared: 0.9679

F-statistic: 513.1 on 1 and 16 DF, p-value: 1.392e-13

**a**

- Gi en fortolkning av estimatene  $\hat{\beta}_0$  og  $\hat{\beta}_1$ .
- Lag så et 95% konfidensintervall for  $\beta_1$  (du trenger ikke å utlede det, men skriv opp formelen du bruker).
- Kan en forvente at importen øker med omrent 0.25 millioner FRF når forbruket øker med 1 million FRF?

Residualplott for modell (1) er vist under:



**b**

- Hvilke modellantakelsjer har en gjort i modell (1)?
- Hvordan sjekker en hver av modellantakelsene med ved hjelp av residualplottene over?
- Benytt residualplottene over til å vurdere gyldigheten av modellantakelsene i modell (1).

(Fortsettes på side 5.)

Ut fra resultatene i residualplottene i Oppgave b, velger vi å tilpasse følgende multiple lineære regresjonsmodell i stedet:

$$Y_i = \gamma_0 + \gamma_1(x_{i1} - \bar{x}_1) + \gamma_2x_{i2} + \varepsilon_i, \quad i = 1, \dots, 18, \quad (2)$$

der  $x_{i2} = (x_{i1} - \bar{x}_1)^2$  og  $\varepsilon_1, \dots, \varepsilon_{18} \stackrel{uif}{\sim} N(0, \tau^2)$ .

### c

- Vis at forventet import nå er en kvadratisk funksjon av forbruket.
- Hvor mye øker forventet import når forbruket øker med én enhet fra  $x_1^*$  (altså fra  $x_1^*$  til  $x_1^* + 1$ )?
- Hva er nå fortolkningen av konstantleddet  $\gamma_0$ , og hva med fortolkningen av  $\gamma_1$ ?

Under og på neste side vises

- R-utskriften fra en test av modell (2) mot modell (1).
- R-utskriften fra tilpasningen til modell (2)
- residualplot for modell (2)

### d

- Hvilke hypoteser er det som testes i R-utskriften under?
- Hvilken av modellene (1) og (2) vil du velge basert på hypotesetesten og R-utskriften fra tilpasningen? Begrunn svaret ditt.
- Bruk residualplottene til å sjekke modellantakelsene for modell (2), og sammenlikn med Oppgave b.

R-utskrift fra hypotesetest av modell (2) mot modell (1):

#### Analysis of Variance Table

```

Model 1: import ~ I(forbruk - mean(forbruk))
Model 2: import ~ I(forbruk - mean(forbruk)) + I((forbruk - mean(forbruk))^2)
Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1      16 80.082
2      15 16.397  1     63.685 58.261 1.528e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

R-utskrift fra tilpassning til modell (2):

Call:

```
lm(formula = import ~ I(forbruk - mean(forbruk)) + I((forbruk -
mean(forbruk))^2))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.79590	-0.52376	-0.04364	0.47666	1.92453

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.034259	0.363875	77.044	< 2e-16 ***
I(forbruk - mean(forbruk))	0.277111	0.006561	42.233	< 2e-16 ***
I((forbruk - mean(forbruk))^2)	0.001251	0.000164	7.633	1.53e-06 ***

---

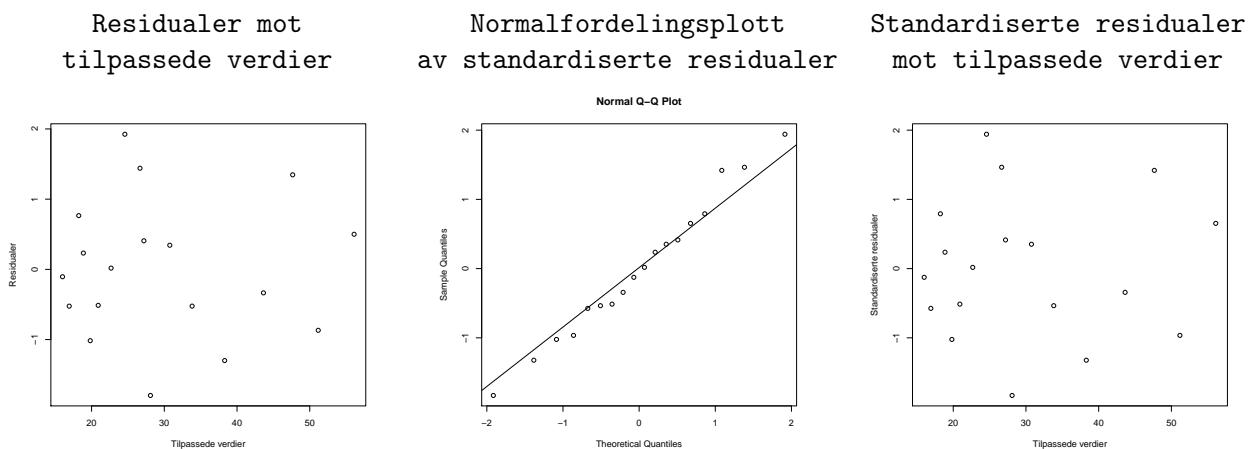
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.046 on 15 degrees of freedom

Multiple R-squared: 0.9938, Adjusted R-squared: 0.993

F-statistic: 1204 on 2 and 15 DF, p-value: < 2.2e-16

Residualplott for modell (2):



END

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder og dataanalyse

Eksamensdag: Fredag 3. desember 2021

Tid for eksamen: 09.00–13.00

Oppgavesettet er på 5 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Alle hjelpeemidler tillatt

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

Nedenfor er det gitt kritiske verdier  $t_{\alpha,\nu}$  for t-fordelingen med  $\nu$  frihetsgrader for noen verdier av  $\alpha$  og  $\nu$ . Du vil få bruk for tabellen i Oppgave 1 og 3.

$\alpha :$	0.05	0.025	0.0125	0.01	0.005	0.001	0.0001	0.00001
$t_{\alpha,15}$	1.753	2.131	2.490	2.602	2.947	3.733	4.880	6.109
$t_{\alpha,16}$	1.746	2.120	2.473	2.583	2.921	3.686	4.791	5.959
$t_{\alpha,17}$	1.740	2.110	2.458	2.567	2.898	3.646	4.714	5.832
:				:				:
$t_{\alpha,69}$	1.667	1.995	2.291	2.382	2.649	3.213	3.929	4.580
$t_{\alpha,70}$	1.667	1.994	2.291	2.381	2.648	3.211	3.926	4.576
$t_{\alpha,71}$	1.667	1.994	2.290	2.380	2.647	3.209	3.923	4.571

### Oppgave 1

17 unge kvinner blir lagt inn til behandling mot anorexia. De ble veid ved innleggelsestidspunktet, samt etter en periode med familieterapi. Et utdrag av før- og ettermålingen av vekt i lb (pund), samt differansen mellom de to, er vist under.

Før	Etter	Differanse
83.8	95.2	11.4
83.3	94.3	11.0
86.0	91.5	5.5
.	.	.
.	.	.
.	.	.
89.9	93.8	3.9
86.0	91.7	5.7
87.3	98.0	10.7

La  $D_i$  differansen mellom vekten til kvinne nummer  $i$  etter terapien og før terapien, altså vektökningen fra første til siste måling. Det antas at

(Fortsettes på side 2.)

$D_1, \dots, D_{17} \stackrel{uif}{\sim} N(\mu_D, \sigma_D^2)$ . En ønsker å undersøke om terapien virker ved å lage et konfidensintervall for  $\mu_D$ , samt ved å test hypotesene

$$H_0 : \mu_D \leq 0 \text{ mot } H_a : \mu_D > 0.$$

**a**

- Utled et 95% konfidensintervall for forventet vektøkning  $\mu_D$  etter terapien.
- Beregn intervallet for dataene over, når du får vite at observert snitt og standardavvik for vektdifferensene er  $\bar{d} = (1/17) \sum_{i=1}^{17} d_i = 7.26$  og  $s_d = \sqrt{(1/16) \sum_{i=1}^{17} (d_i - \bar{d})^2} = 7.16$ .

**b**

- Utled en hypotesetest med signifikansnivå 5% for hypotesene over vedrørende effekten av terapien på vektøkningen.
- Hva blir konklusjonen ut fra de observasjonene som ble gjort?

**c**

- Forklar hva en P-verdi generelt betyr, altså definér hva den er.
- Finn et uttrykk for P-verdien til testen i punkt c).
- Bruk tabellen over kritiske verdier for t-fordelingen til å si noe om størrelsesorden for P-verdien.

## Oppgave 2

La  $X_1, \dots, X_n$  være uavhengige stokastiske variabler fra Weibull-fordelingen, dvs. med sannsynlighetstetthet

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, \quad x > 0,$$

med  $0 < \alpha$  og  $0 < \beta$ . I hele oppgaven antas det at  $\alpha$  er kjent og at  $\alpha > 1$ , mens  $\beta$  er ukjent. Videre får du oppgitt at for en Weibull-fordelt variabel er  $E(X^r) = \beta^r \Gamma(1 + \frac{r}{\alpha})$  for en hvilken som helst  $r > 0$  når  $\alpha > 1$ , der  $\Gamma(\cdot)$  er gammafunksjonen.

**a**

Vis at momentestimatoren for  $\beta$  er  $\tilde{\beta} = \frac{\bar{X}}{\Gamma(1 + \frac{1}{\alpha})}$ , der  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

**b**

Bruk sentralgrenseteoremet til å vise at  $\tilde{\beta}$  er tilnærmet  $N(\beta, \sigma_{\tilde{\beta}}^2)$ -fordelt for store  $n$ , med  $\sigma_{\tilde{\beta}}^2 = \frac{\beta^2}{n} \left( \Gamma(1 + \frac{2}{\alpha}) / \Gamma(1 + \frac{1}{\alpha})^2 - 1 \right)$ .

**c**

Sett opp likelihood- og log-likelihood-funksjonen, og finn et uttrykk for maksimum likelihood-estimatoren  $\hat{\beta}$ .

(Fortsettes på side 3.)

**d**

- Begrunn at  $\hat{\beta}$  er tilnærmet  $N(\beta, \sigma_{\hat{\beta}}^2)$ -fordelt for store  $n$ , og finn et uttrykk for  $\sigma_{\hat{\beta}}^2$  når du får vite at Fisher-informasjonen i én observasjon er  $I(\beta) = \frac{\alpha^2}{\beta^2}$ .
- Hvilken av estimatorene  $\tilde{\beta}$  og  $\hat{\beta}$  vil du foretrekke? Begrunn svaret ditt (*Hint:* du kan bruke at  $\alpha^2 \left( \Gamma(1 + \frac{2}{\alpha}) / \Gamma(1 + \frac{1}{\alpha})^2 - 1 \right) > 1$  når  $\alpha > 1$ ).

**Oppgave 3**

Vi går tilbake til vektdataene fra Oppgave 1. Det fulle datasettet inkluderer vektmålinger av 72 unge kvinner på to forskjellige tidspunkter. Av de 72 kvinnene fikk 17 familieterapi (dette er dataene fra Oppgave 1), 29 fikk kognitiv behandling og 26 fikk ingen behandling, og utgjør en kontrollgruppe.

Vi velger først å analysere dataene ved hjelp av enveis variansanalyse, der målingen av interesse er differensen mellom siste og første vektmåling, dvs. vektøkningen fra første til siste måling (disse ble kalt  $D_i$  i Oppgave 1), og grupperingen er gjort med type behandling som faktor. Resultatene fra analysen er vist under.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(behandling)	2	615	307.32	5.422	0.0065 **
Residuals	69	3911	56.68		
<hr/>					
Signif. codes:	0	***	0.001	**	0.01 *
					0.05 .
					0.1 ‘ ’ 1

**a**

- Forklar oppsettet for å utføre en slik analyse på dette datasettet, altså skriv ned modellen og forklar hvilke antakelser som gjøres og hvilke hypoteser som testes.
- Hva kan du si om effekten av behandling ut fra resultatene?

Et alternativ til enveis variansanalyse er en lineær regresjonsmodell, med målt vektøkning som responsvariabel  $Y$ . Videre velger vi i denne omgang å gruppere kvinnene etter om de fikk behandling eller ikke, uten å spesifisere hvilken behandling de eventuelt fikk. Vår ene forklaringsvariabel  $x$  er da gitt ved:

$$x_i = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling} \end{cases}$$

og modellen er:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

der vi antar at  $\varepsilon_i \stackrel{uif}{\sim} N(0, \sigma^2)$ . Resultatet av denne analysen er gitt i R-utskriften nedenfor.

Call:

lm(formula = vekt.diff ~ behandling.gruppert)

Residuals:

(Fortsettes på side 4.)

Min	1Q	Median	3Q	Max
-13.6804	-5.3054	-0.8804	6.1946	16.9196

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.450	1.502	-0.300	0.76535
behandling.gruppert	5.030	1.879	2.677	0.00924 **
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 7.658 on 70 degrees of freedom

Multiple R-squared: 0.09289, Adjusted R-squared: 0.07993

F-statistic: 7.168 on 1 and 70 DF, p-value: 0.009239

## b

- Bruk utskriften fra R til å angi hva estimatene  $\hat{\beta}_0$  og  $\hat{\beta}_1$  er.
- Vis at

$$E(Y_i|x_i = 0) = \beta_0 \quad \text{og} \quad E(Y_i|x_i = 1) = \beta_0 + \beta_1,$$

og bruk dette til å gi en fortolkning av  $\hat{\beta}_0$  og  $\hat{\beta}_1$ .

- Lag også et 95% konfidensintervall for  $\beta_1$  (du trenger ikke å utlede det, men skriv opp formelen du bruker).
- Hva kan du si om effekten av behandling på forventet vektøkning ut fra intervallet?

Vi ønsker nå å undersøke effekten av de to ulike behandlingene. Vi definerer derfor de to forklaringsvariablene  $x_1$  og  $x_2$  ved:

$$x_{1i} = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling 1} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling 1} \end{cases} \quad x_{2i} = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling 2} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling 2} \end{cases}$$

der behandling 1 er familieterapi og behandling 2 er kognitiv behandling.

Vår nye modell er da gitt ved:

$$Y_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \varepsilon_i^*,$$

der vi antar at  $\varepsilon_i^* \stackrel{uif}{\sim} N(0, (\sigma^*)^2)$ . Resultatet av denne analysen er gitt i R-utskriften nedenfor.

Call:

```
lm(formula = vekt.diff ~ factor(behandling), x = TRUE)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.565	-4.543	-1.007	3.846	17.893

(Fortsettes på side 5.)

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.450	1.476	-0.305	0.7614
factor(behandling)1	7.715	2.348	3.285	0.0016 **
factor(behandling)2	3.457	2.033	1.700	0.0936 .
---				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Residual standard error: 7.528 on 69 degrees of freedom

Multiple R-squared: 0.1358, Adjusted R-squared: 0.1108

F-statistic: 5.422 on 2 and 69 DF, p-value: 0.006499

**c**

- Finn  $E(Y_i|x_{1i} = j, x_{2i} = k)$  for de tre mulige kombinasjonene  $(0,0)$ ,  $(1,0)$  og  $(0,1)$  for  $j, k$ , og bruk disse til å gi en fortolkning av  $\hat{\gamma}_0$ ,  $\hat{\gamma}_1$  og  $\hat{\gamma}_2$ .
- Hvorfor er  $\gamma_0 = \beta_0$ , der  $\beta_0$  er konstantleddet fra Oppgave a)?

Vi ønsker å finne ut om hver av de to behandlingsformene har en effekt på forventet vektøkning eller ikke. Vi vil derfor teste hypotesene

$$H_0 : \gamma_j = 0 \text{ mot } H_a : \gamma_j \neq 0,$$

for  $j = 1, 2$ .

**d**

- Forklar hvilke testobservatorer og tilhørende forkastningsområder du kan bruke for å teste hypotesene over med 5% signifikansnivå.
- Utfør så testene for  $j = 1$  og  $2$  basert på R-utskriftene.
- Hva kan du si om effekten av familieterapi (behandling 1) og kognitiv behandling (behandling 2), og hvordan samsvarer disse resultatene med resultatene fra Oppgave a) og b)?

END

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder og dataanalyse

Eksamensdag: Torsdag 26. november 2020

Tid for eksamen: 09.00–13.00

Oppgavesettet er på 6 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Alle hjelpeemidler tillatt

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

Nedenfor er det gitt kritiske verdier  $t_{\alpha,\nu}$  for t-fordelingen med  $\nu$  frihetsgrader for noen verdier av  $\alpha$  og  $\nu$ . Du vil få bruk for tabellen i Oppgave 1 og 3.

$\alpha :$	0.05	0.025	0.01	0.005	0.001	0.0001	0.00001
$t_{\alpha,8}$	1.860	2.306	2.896	3.355	4.501	6.442	8.907
$t_{\alpha,9}$	1.833	2.262	2.821	3.250	4.297	6.010	8.102
$t_{\alpha,10}$	1.812	2.228	2.764	3.169	4.144	5.694	7.527
:			:			:	
$t_{\alpha,26}$	1.706	2.056	2.479	2.779	3.435	4.324	5.197
$t_{\alpha,27}$	1.703	2.052	2.473	2.771	3.421	4.299	5.157
$t_{\alpha,28}$	1.701	2.048	2.467	2.763	3.408	4.275	5.120

### Oppgave 1

En eplebonde har fått målt gjennomsnittlig vekt i gram per eple på 28 av epletrærne sine. Dataene er vist under.

85.3 86.9 96.8 108.5 113.8 87.7 94.5 99.9 92.9 67.3 90.6  
129.8 48.9 117.5 100.8 94.5 94.4 98.9 96.0 99.4 79.1 108.5  
84.6 117.5 70.0 104.4 127.1 135.0

La  $X_i$  være gjennomsnittlig vekt per eple på tre nummer  $i$ . Det antas at  $X_1, \dots, X_{28} \stackrel{uif}{\sim} N(\mu, \sigma^2)$ .

a

Utled et 99% konfidensintervall for forventet gjennomsnittlig vekt  $\mu$  per eple. Beregn intervallet for dataene over, når du får vite at observert snitt og standardavvik er  $\bar{x} = 1/28 \sum_{i=1}^{28} x_i = 97.52$  og  $s = \sqrt{1/27 \sum_{i=1}^{28} (x_i - \bar{x})^2} = 18.95$ .

Ifølge MatPrat veier et gjennomsnittlig eple 115 gram. Eplebonden ønsker å finne ut om epletrærne hans gir epler av normal størrelse, eller om de er litt

(Fortsettes på side 2.)

små. Han vil derfor teste hypotesene

$$H_0 : \mu \geq 115 \text{ mot } H_a : \mu < 115.$$

**b**

Utled en test med signifikansnivå 1%. Hva blir konklusjonen ut fra de observasjonene som ble gjort?

**c**

Finn et uttrykk for P-verdien. Forklar hva den betyr. Bruk tabellen over kritiske verdier for t-fordelingen til å si noe om størrelsesordenen for P-verdien.

## Oppgave 2

La  $X_1, \dots, X_n$  være uavhengige stokastiske variabler med punktsannsynlighet

$$f(x; \theta) = \frac{e^{-\theta}x(\theta x)^{x-1}}{x!}, \quad x = 0, 1, 2, \dots$$

med  $0 < \theta < 1$ . Vi sier at de stokastiske variablene er Borel-fordelt. For Borel-fordelingen har vi  $E(X) = \frac{1}{1-\theta}$  (du skal ikke vise dette).

**a**

Vis at momentestimatoren for  $\theta$  er  $\tilde{\theta} = \frac{\bar{X}-1}{\bar{X}}$ , der  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

**b**

Sett opp likelihood- og log-likelihood-funksjonen, og finn et uttrykk for maksimum likelihood-estimatoren  $\hat{\theta}$ .

**c**

Vis at Fisher-informasjonen i én observasjon er  $I(\theta) = \frac{1}{\theta(1-\theta)}$ .

**d**

Begrunn at  $\hat{\theta}$  er tilnærmet  $N(\theta, \sigma_{\hat{\theta}}^2)$ -fordelt og finn et uttrykk for  $\sigma_{\hat{\theta}}^2$ .

## Oppgave 3

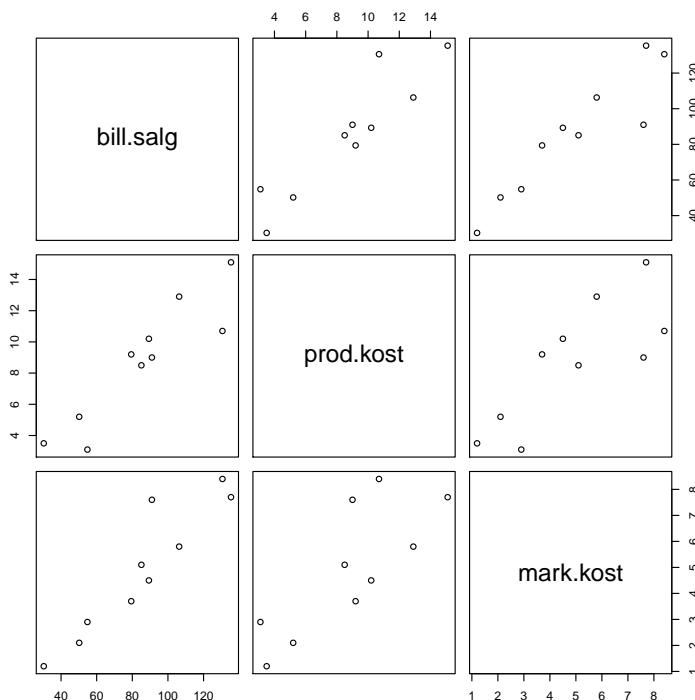
En ønsker å finne ut hvordan billettsalget for en kinofilm avhenger av produksjons- og markedsføringskostnadene. Dataene, som er basert på 10 Hollywood-filmer, består av

- responsvariabel  $y$ : billettsalg
- forklaringsvariabel  $x_1$ : produksjonskostnader
- forklaringsvariabel  $x_2$ : markedsføringskostnader,

(Fortsettes på side 3.)

alle oppgitt i millioner USD, og er vist i tabellen og matrisespredningsplottet under.

	bill.salg	prod.kost	mark.kost
1	85.1	8.5	5.1
2	106.3	12.9	5.8
3	50.2	5.2	2.1
4	130.6	10.7	8.4
5	54.8	3.1	2.9
6	30.3	3.5	1.2
7	79.4	9.2	3.7
8	91.0	9.0	7.6
9	135.4	15.1	7.7
10	89.3	10.2	4.5



Vi gjør først en regresjonsanalyse med produksjonskostnader som eneste forklaringsvariabel. Vi tilpasser da den lineære regresjonsmodellen:

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \epsilon_i, \quad i = 1, \dots, 10,$$

der  $\bar{x}_1 = 1/n \sum_{i=1}^n x_{i1}$  og vi antar at  $\epsilon_1, \dots, \epsilon_{10} \stackrel{uif}{\sim} N(0, \sigma^2)$ . Resultatet av denne analysen er gitt i R -utskriften nedenfor.

Call:

```
lm(formula = bill.salg ~ I(prod.kost - mean(prod.kost)), data = film.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.136	-9.029	-3.689	3.208	29.723

(Fortsettes på side 4.)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	85.240	4.509	18.906	6.34e-08 ***
I(prod.kost - mean(prod.kost))	7.978	1.223	6.522	0.000184 ***
<hr/>				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

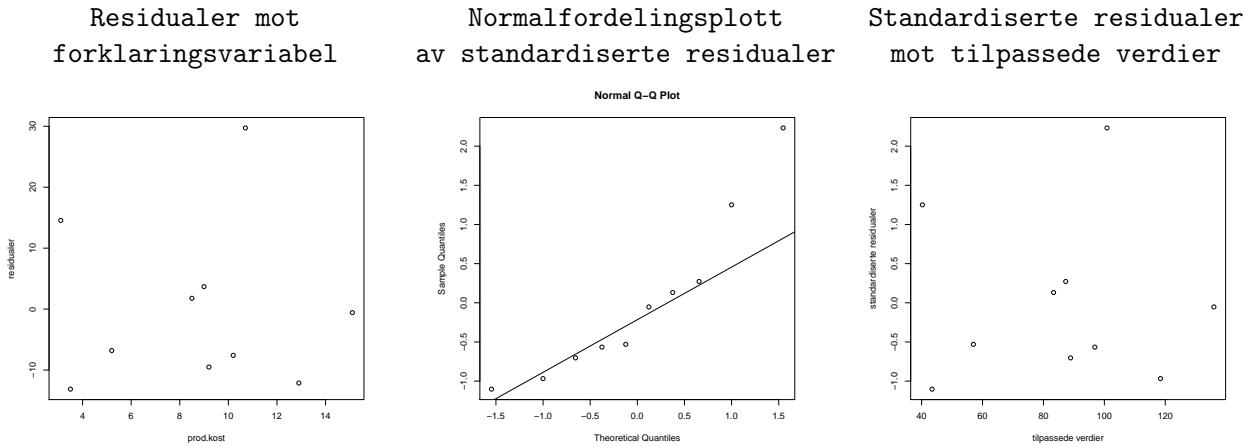
Residual standard error: 14.26 on 8 degrees of freedom

Multiple R-squared: 0.8417, Adjusted R-squared: 0.8219

F-statistic: 42.54 on 1 and 8 DF, p-value: 0.0001838

a

Gi en fortolkning av estimatene  $\hat{\beta}_0$  og  $\hat{\beta}_1$ . Lag så et 95% konfidensintervall for  $\beta_1$ . Et produksjonsselskap lurer på om de kan forvente å få økt billettsalget med 10 millioner USD per ekstra million de investerer. Kan du si noe om dette basert på konfidenintervallet du har laget?



b

Benytt residualplottene over, som er fra den enkle lineære regresjonsmodellen vurdert så langt, til å vurdere gyldigheten av modellantagelsene. Forklar spesielt hvordan avvik fra modellen kan påvises i de forskjellige plottene.

La  $x_1^*$  være en ny verdi av produksjonskostnadene, og la  $Y$  være det tilsvarende billettsalget. Videre la

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(x_1^* - \bar{x})$$

være det predikerte billettsalget. Det kan vises (dette skal du ikke å gjøre) at  $\hat{Y}$  kan skrives som

$$\hat{Y} = \sum_{i=1}^{10} \left( \frac{1}{10} + \frac{(x_1^* - \bar{x})(x_{i1} - \bar{x}_1)}{S_{xx}} \right) Y_i$$

med  $S_{xx} = \sum_{i=1}^{10} (x_{i1} - \bar{x}_1)^2$ .

(Fortsettes på side 5.)

**c**

Bruk resultatene over til å argumentere for at  $\hat{Y} \sim N(\mu_{Y|x^*}, \sigma_{\hat{Y}}^2)$ . Vis at  $\mu_{Y|x^*} = E(Y|x^*)$  og  $\sigma_{\hat{Y}}^2 = \text{Var}(\hat{Y}) = \sigma^2 \left( \frac{1}{10} + \frac{(x_1^* - \bar{x}_1)^2}{S_{xx}} \right)$  (du kan ta for gitt at  $\hat{\beta}_0$  og  $\hat{\beta}_1$  er forventningsrette for  $\beta_0$  og  $\beta_1$ ).

**d**

Bruk resultatene fra c) til å utlede et  $100 \cdot (1 - \alpha)\%$  konfidensintervall for  $\mu_{Y|x^*}$  når du også kan ta for gitt at  $(28 - 2)S^2/\sigma^2 \sim \chi^2_{28-2}$  og  $\hat{Y}$  er uavhengige, der  $S^2$  er den vanlige forventningsrette estimatoren for  $\sigma^2$ .

Den nye James Bond-filmen (som etter planen skulle hatt premiere i november) har kostet utrolige 301 millioner USD.

Hva er det forventede billettsalget på denne filmen basert på modellen over? Lag et 95% konfidensintervall for denne forventningen når du får vite at  $\bar{x}_1 = 8.74$  og  $S_{xx} = \sum_{i=1}^{10} (x_{i1} - \bar{x}_1)^2 = 135.864$ .

Kan du støle på resultatene du har fått? Begrunn svaret ditt.

En ønsker nå å vurdere om også markedsførsingskostnader bør være med i modellen, dvs.:

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2x_{i2} + \epsilon_i, \quad i = 1, \dots, 10, \quad \epsilon_1, \dots, \epsilon_{10} \stackrel{uif}{\sim} N(0, \sigma^2).$$

R-utskriftene under viser resultatene av tilpasningen til den multiple lineære regresjonsmodellen over, samt fra en hypotesetest som sammenligner denne med den enkle lineære regresjonsmodellen brukt tidligere i oppgaven.

**e**

Gi en forklaring av estimatene  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  og  $\hat{\beta}_2$  i denne nye modellen. Sett i lys av matrisespredningsplottet, hvordan forklarer du at  $\hat{\beta}_1$  har endret seg sammenlignet med den enkle lineære regresjonsmodellen?

Formuler så hypotesene som blir testet i hypotesetesten gjengitt i variansanalysetabellen, altså under "Analysis of Variance Table".

Hvilken modell vil du velge? Begrunn svaret ditt.

Call:

```
lm(formula = bill.salg ~ I(prod.kost - mean(prod.kost)) + mark.kost,
  data = film.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.4168	-2.5696	0.8052	2.1200	11.0463

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	48.803	9.227	5.289	0.00114 **
I(prod.kost - mean(prod.kost))	4.228	1.153	3.667	0.00800 **
mark.kost	7.436	1.806	4.117	0.00448 **
---				

(Fortsettes på side 6.)

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.241 on 7 degrees of freedom  
Multiple R-squared: 0.9537, Adjusted R-squared: 0.9405  
F-statistic: 72.14 on 2 and 7 DF, p-value: 2.131e-05

#### Analysis of Variance Table

Model 1: bill.salg ~ I(prod.kost - mean(prod.kost))  
Model 2: bill.salg ~ I(prod.kost - mean(prod.kost)) + mark.kost

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	8	1626.27			
2	7	475.37	1	1150.9	16.947 0.004478 **
---					

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

END

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse

Eksamensdag: Mandag 16. desember 2019.

Tid for eksamen: 14.30 – 18.30.

Oppgavesettet er på 4 sider.

Vedlegg: Tabell over standardnormalfordelingen.

Tillatte hjelpeemidler: Godkjent kalkulator.  
Formelsamling til STK1100 og STK1110.

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

Nedenfor er det gitt kritiske verdier  $t_{\alpha,\nu}$  for t-fordelingen med  $\nu$  frihetsgrader for noen verdier av  $\alpha$  og  $\nu$ . Du vil få bruk for tabellen i Oppgave 3.

$\alpha :$	0.05	0.025	0.01	0.005
$t_{\alpha,17} :$	1.740	2.110	2.567	2.898
$t_{\alpha,18} :$	1.734	2.101	2.552	2.878
$t_{\alpha,19} :$	1.729	2.093	2.539	2.861

### Oppgave 1

Etter forskriftene skal mengden av organisk karbon i drikkevannet ikke overstige 5.0 mg/l. Ingeniørene ved et vannverk tar fem prøver av drikkevannet for å kontrollere vannkvaliteten. La  $X_1, \dots, X_5$  være mengden av organisk karbon i hver av de fem prøvene. Vi vil anta at  $X_i$ -ene er uavhengige og  $N(\mu, \sigma^2)$ -fordelte, der  $\mu$  er den faktiske konsentrasjonen av organisk karbon i drikkevannet. I punktene a-d vil vi anta at  $\sigma = 0.5$  mg/l.

- a) Ingeniørene ved vannverket vil bruke hypotesetesting til å vurdere om drikkevannet inneholder for mye organisk karbon. De diskuterer om de skal teste

$$H_0 : \mu \leq 5.0 \quad \text{mot} \quad H_a : \mu > 5.0,$$

eller om de skal teste

$$H_0 : \mu \geq 5.0 \quad \text{mot} \quad H_a : \mu < 5.0.$$

Diskuter de to formuleringene av hypotesetestingsproblemet. Kommenter spesielt hvilken konklusjon en vil få for hver av de to formuleringene hvis nullhypotesen blir forkastet.

I resten av oppgaven vil vi se på testing av  $H_0 : \mu \geq 5.0$  mot  $H_a : \mu < 5.0$ .

Innholdet av organisk karbon i de fem vannprøvene ingeniørene tok var

5.37, 4.88, 4.32, 4.59, 4.14

(Fortsettes på side 2.)

- b) Utled en test med signifikansnivå 5%.  
 Hva blir konklusjonen ut fra de fem observasjonene som ble gjort?
- c) Bestem P-verdien og forklar hva den betyr.
- d) Hvor mange vannprøver må ingeniørene ta for at sannsynligheten for feil av type II skal være høyest 10% hvis konsentrasjonen av organisk karbon er 4.5 mg/l.

Vi har i punktene a-d antatt at  $\sigma = 0.5$  mg/l. I praksis vil ikke  $\sigma$  være kjent, så den må estimeres fra vannprøvene.

- e) Forklar hvordan du nå vil gå fram for å teste  $H_0 : \mu \geq 5.0$  mot  $H_a : \mu < 5.0$ .  
 (Du skal ikke gjøre noen beregninger her.)

## Oppgave 2

La  $X_1, X_2, \dots, X_n$  være uavhengige stokastiske variabler med sannsynlighetstetthet

$$f(x) = \begin{cases} \frac{1}{\theta^2} x e^{-x^2/(2\theta^2)} & \text{for } x > 0, \\ 0 & \text{ellers.} \end{cases}$$

Vi sier at de stokastiske variablene er Rayleigh fordelt. For Rayleigh fordelingen har vi at  $E(X_i) = \theta\sqrt{\frac{\pi}{2}}$  og  $E(X_i^2) = 2\theta^2$ . (Du skal ikke vise dette.)

- a) Vis at momentestimatoren er  $\tilde{\theta} = \bar{X}\sqrt{2/\pi}$ , der  $\bar{X} = \sum_{i=1}^n X_i/n$ .

Bruk sentralgrensesetningen til å vise at  $\tilde{\theta}$  er tilnærmet  $N(\theta, \sigma_{\tilde{\theta}}^2)$ -fordelt, der

$$\sigma_{\tilde{\theta}}^2 = \frac{4 - \pi}{n\pi}\theta^2.$$

- b) Sett opp likelihooden og log-likelihooden og finn et uttrykk for maksimum likelihood estimatoren  $\hat{\theta}$ .
- c) Vis at Fisher informasjonen i én observasjon er  $I(\theta) = 4/\theta^2$ .
- d) Begrunn at  $\hat{\theta}$  er tilnærmet  $N(\theta, \sigma_{\hat{\theta}}^2)$ -fordelt og finn et uttrykk for  $\sigma_{\hat{\theta}}^2$ .  
 Hvilkene av estimatorene  $\tilde{\theta}$  og  $\hat{\theta}$  vil du foretrekke? Begrunn svaret ditt.

## Oppgave 3

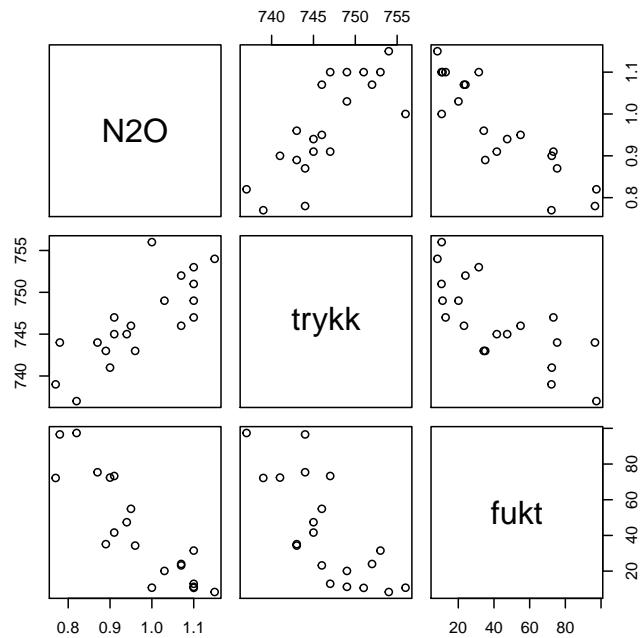
Det har blitt gjort en studie av utsipp av dinitrogenoksid ( $N_2O$ ) fra en liten dieseldrevet lastebil. Formålet med studien var å finne ut hvordan utsippet varierer med luftfuktigheten og lufttrykket. Det ble tilsammen gjort 20 målinger av utsippet under ulike forhold.

Vi vil bruke lineær regresjon til å analysere dataene. Responsen,  $N_2O$ , er utsipp av dinitrogenoksid (ppm), mens forklaringsvariablene er:

**trykk** lufttrykk (mmHg),  
**fukt** luftfuktighet (prosent).

På neste side er det gitt et matrise-spredningsplot som gir en oversikt over dataene.

(Fortsettes på side 3.)



Vi gjør først en regresjonsanalyse med lufttrykk som eneste forklaringsvariabel. Vi tilpasser da den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \quad (1)$$

der responsen  $Y_i$  er utslipp av dinitrogenoksid og forklaringsvariablen  $x_{i1}$  er lufttrykk minus 750 mmHg. Resultatet av denne analysen er gitt i R-utskriften nedenfor. (Vi har fjernet en del av resultatene i utskriften.)

Call:

```
lm(formula = N2O ~ I(trykk - 750), data = utslepp)
```

Coefficients:

	Estimate	Std. Error
(Intercept)	1.033773	0.020147
I(trykk - 750)	0.018195	0.003384
<hr/>		
Residual standard error:	0.07343	
Multiple R-squared:	0.6163	

- a) Gi en forklaring av estimatene for  $\beta_0$  og  $\beta_1$  og bestem et 99% konfidensintervall for  $\beta_1$ .

Vi gjør så en regresjonsanalyse med både lufttrykk og luftfuktighet som forklaringsvariabler. Vi tilpasser da den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad (2)$$

der  $Y_i$  og  $x_{i1}$  er som for modell (1) og  $x_{i2}$  er luftfuktighet minus 50 prosent. Resultatet av denne analysen er gitt i R-utskriften på neste side.

(Fortsettes på side 4.)

Call:

```
lm(formula = N20 ~ I(trykk - 750) + I(fukt - 50), data = utslipp)
```

Coefficients:

	Estimate	Std. Error
(Intercept)	0.9737589	0.0199131
I(trykk - 750)	0.0064514	0.0036226
I(fukt - 50)	-0.0026547	0.0006131
---		
Residual standard error:	0.05211	
Multiple R-squared:	0.8175	

- b) Sammenlign estimatene for  $\beta_1$  i modell (1) og modell (2). Hvordan kan du forklare at estimatene er såpass forskjellige?
- c) Bestem et estimat for forventet utslipp når lufttrykket er 740 mmHg og luftfuktigheten er 80 prosent. Forklar hvorfor du ikke kan finne standardfeilen til dette estimatet ut fra informasjonen i R-utskriften.

Vi innfører nå matrisen

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{20,1} & x_{20,2} \end{pmatrix}.$$

Da har vi at

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.1460 & 0.0212 & 0.00313 \\ 0.0212 & 0.00483 & 0.000612 \\ 0.00313 & 0.000612 & 0.000138 \end{pmatrix}.$$

- d) Bruk resultatet ovenfor og R-utskriften til å bestemme standardfeilen til estimatet i punkt c. Bestem også et 95% konfidensintervall for forventet utslipp når lufttrykket er 740 mmHg og luftfuktigheten er 80 prosent.

**SLUTT**

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 – Statistiske metoder og dataanalyse 1

Eksamensdag: Tirsdag 18. desember 2018

Tid for eksamen: 09.00–13.00

Oppgavesettet er på 5 sider.

Vedlegg: Tabell over normalfordelingen

Tillatte hjelpeemidler: Formelsamling for STK1100/STK1110.  
Godkjent kalkulator.

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

### Oppgave 1

La  $X$  være binomisk fordelt med parametre  $n$  som antall forsøk og  $p$  som suksessanssynlighet,  $X \sim B(n, p)$ . En estimator for  $p$  er da gitt ved  $\hat{p} = \frac{X}{n}$ .

**a**

Vis at  $\hat{p}$  er forventningsrett.

Utled et uttrykk for variansen til  $\hat{p}$ .

**b**

Forklar hva en likelihoodfunksjon er og vis at  $L(p) = \binom{n}{X} p^X (1-p)^{n-X}$  er likelihooden for  $p$  i dette tilfellet.

Vis at  $\hat{p} = \frac{X}{n}$  er maksimum likelihood estimator for  $p$ .

Benytt likelihoodteori for å finne et (tilnærmet) uttrykk for variansen til  $\hat{p}$ . Sammenlign med variansuttrykket fra punkt a.

**c**

Tilnærmet har at

$$Z(p) = \frac{\hat{p} - p}{\sqrt{p(1-p)}} \sqrt{n}$$

er godt standardnormalfordelt når  $np > 10$  og  $n(1-p) > 10$ .

Bruk dette til å konstruere en hypotesetest med tilnærmet 5% nivå for  $H_0 : p = p_0$  mot  $H_a : p \neq p_0$ .

Utfør testen når  $X = 15$ ,  $n = 50$  og  $p_0 = 0.5$ . Finn også tilnærmet P-verdi for testen.

(Fortsettes på side 2.)

**d**

Forklar hvorfor mengden  $\{p : |Z(p)| < 1.96\}$  blir et tilnærmet 95% konfidensintervall for  $p$ .

Vis hvordan dette intervallet kan avleses grafisk ved å plotte  $Z(p)$  mot  $p$ .

Vis også at intervallet kan finnes analytisk ved å løse en 2.gradsulikhet (men du skal ikke utlede de faktiske løsningene, de er oppgitt under).

**e**

Løsningen på 2.gradsulikheten kan uttrykkes som

$$\frac{\hat{p} + 1.96^2/2n}{1 + 1.96^2/n} \pm \frac{1.96}{\sqrt{n}} \frac{\sqrt{\hat{p}(1 - \hat{p}) + 1.96^2/4n}}{1 + 1.96^2/n}$$

(dette skal du ikke vise!)

Når  $n$  er stor kan leddet  $a/n$  ignoreres. Finn forenklingen av intervallet etter denne tilnærmingen.

Sammenlign dette intervallet med det tilnærmede 95% konfidensintervallet du får basert på at

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})}} \sqrt{n}$$

er tilnærmet standardnormalfordelt.

Vis utledningen av konfidensintervallet basert på denne tilnærmingen.

## Oppgave 2

Dataene i denne oppgaven kommer fra et eksperiment vedrørende reduksjon i isoleringsevne for en elektrisk isolator som har blitt utsatt for ulike nivåer av temperatur over varierende lengde av tid. Responsvariabelen er isoleringsevne i kilo-volt og forklaringsvariablene er tid i uker og temperatur målt i Celciusgrader. Totalt antall observasjoner er  $n = 128$  og det er fire målinger av isoleringsevne for hver kombinasjon av temperatur (med nivåer 180, 225, 250 og 275 grader i Celcius) og tid (med 8 nivåer 1, 2, 4, 8, 16, 32, 48 og 64 uker). Vi har således et balansert design. Vi vil benytte forklaringsvariablene kodet numeriske som  $x_{1i}$  = grader Celcius og  $x_{2i} = \log(\text{Tid})$  og respons variablen  $Y_i$  = isoleringsevne i KVolt for observasjon nr.  $i$ .

**a**

I en første modell betraktes en enkel lineær regresjonsmodell  $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$  for  $i = 1, \dots, n$  der  $\varepsilon_i$ -ene antas uavhengige og normalfordelte med forventning null og ukjent varians  $\sigma^2$ .

Bestem og for tolk minste kvadraters estimatene for  $\beta_0, \beta_1$  samt estimatet for  $\sigma$  fra R-output øverst på neste side.

Forklar hvordan minste kvadraters estimator bestemmes (det spørres ikke om en detaljert utledning av formler for estimatene).

(Fortsettes på side 3.)

Forklar hvordan P-verdien for variablen Temperatur beregnes. Spesifiser nullhypotesen og den alternative hypotesen som testes.

Benytt utskriften til å finne empiriske korrelasjoner mellom  $x_{1i}$ -ene og  $Y_i$ -ene.

```
> summary(lm(Strength~Temperature))
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	30.511906	1.767079	17.27	<2e-16
Temperature	-0.082898	0.007515	-11.03	<2e-16
---				
Residual standard error:	2.983	on 126 degrees of freedom		
Multiple R-squared:	0.4913	Adjusted R-squared:	0.4872	
F-statistic:	121.7	on 1 and 126 DF,	p-value:	< 2.2e-16

## b

Under ser du utskrift fra tilpasning av en modell  $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ , altså med forklaringsvariabel  $x_{2i}$  lagt til i modellen og ellers de samme antagelsen for  $\varepsilon_i$ -ene som i punkt a).

Forklar hvorfor estimatet for  $\beta_1$  er det samme som i punkt a).

Gi en diskusjon av hva som har skjedd med estimatet for  $\sigma$  og for den multiple R-squared verdien.

Forklar også effekten på standardfeilen til  $\hat{\beta}_1$ . Gi en forklaring for disse endringene.

Bestem korrelasjonen mellom  $Y_i$ -ene og  $x_{2i}$ -ene.

```
> summary(lm(Strength~Temperature+log(Time)))
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	33.735643	1.347989	25.03	<2e-16
Temperature	-0.082898	0.005573	-14.87	<2e-16
log(Time)	-1.399549	0.137163	-10.20	<2e-16
---				
Residual standard error:	2.212	on 125 degrees of freedom		
Multiple R-squared:	0.7224	Adjusted R-squared:	0.718	
F-statistic:	162.7	on 2 and 125 DF,	p-value:	< 2.2e-16

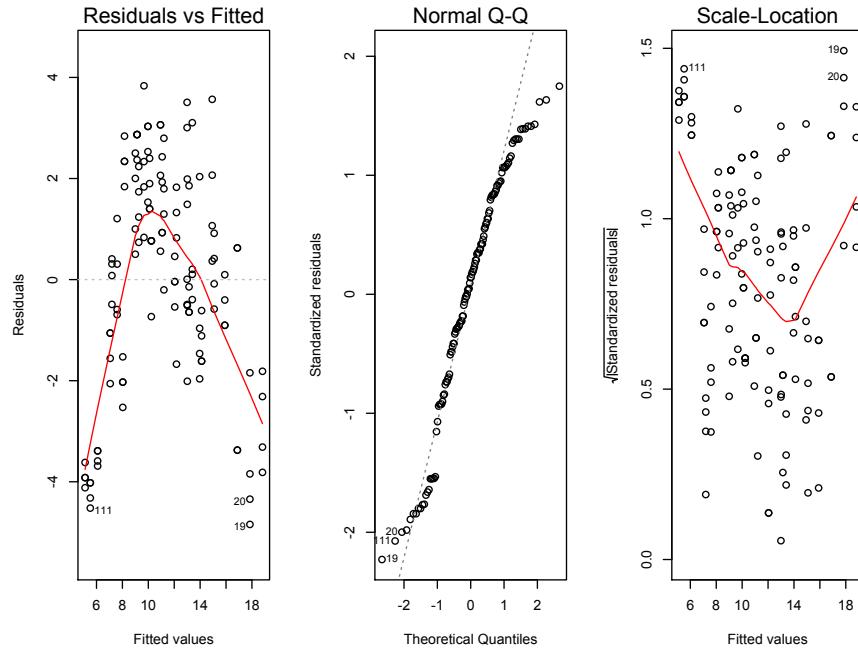
## c

Benytt residualplottene på neste side, som er fra modellen i punkt b), til å vurdere gyldigheten av modellantagelsene.

Forklar spesielt hvilke størrelser som plottes og hvordan avvik fra modellen kan påvises.

Foreslå mulige forbedringer av modellen.

(Fortsettes på side 4.)



### Oppgave 3

Anta en multippel lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, i = 1, \dots, n$$

der  $Y_i$  er respons,  $x_{ji}, j = 1, \dots, k$  forklaringsvariable og  $\varepsilon_i$  feilredd for enhet  $i$  og  $\beta_j, j = 0, \dots, k$  regresjonskoeffisienter. Vi antar at  $\varepsilon_i$  er uavhengige og normalfordelte med forventning 0 og varians  $\sigma^2$ .

La  $\hat{\beta}_j, j = 0, \dots, k$  være minste kvadraters estimatorene for  $\beta_j$ -ene,  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$  og  $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  residual kvadratsum. Da har at  $SSE/\sigma^2 \sim \chi^2_{n-k-1}$ , dvs. er kji-kvadratfordelt med  $n - k - 1$  frihetsgrader. (Dette skal du ikke vise).

**a**

Forklar hvorfor forventningen til  $s^2 = SSE/(n - k - 1)$  er lik  $\sigma^2$ .

Vis at  $V(s^2)$  går mot 0 når  $n - k$  går mot uendelig og argumenter for at dette betyr at  $s^2$  er konsistent for  $\sigma^2$ , dvs. konvergerer i sannsynlighet mot  $\sigma^2$ .

Anta så at  $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ . Forklar hvorfor nå også  $SST/(n - 1)$ , der  $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$ , er konsistent for  $\sigma^2$ .

**b**

Forklart andel av varians er definert som  $R^2 = 1 - SSE/SST$ . Anta at både  $n$  og  $k$  går mot uendelig på en slik måte at  $k/n \rightarrow \rho$  der  $0 < \rho < 1$ . Anta videre  $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ .

(Fortsettes på side 5.)

Vis at da vil  $R^2 \rightarrow \rho$  (som er en verdi større enn 0).

Hint: Du kan bruke at hvis både  $U_n \rightarrow u$  og  $V_n \rightarrow v$  i sannsynlighet der  $V_n$  og  $U_n$  er (sekvenser av) stokastiske variable og  $u$  og  $v$  er skalare verdier så vil  $V_n/U_n \rightarrow v/u$  når  $u \neq 0$ .

Finn tilsvarende grense for justert  $R^2$  definert som  $R_{adj}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$  (under samme forutsetninger).

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 – Statistiske metoder og dataanalyse 1

Eksamensdag: Torsdag 14. desember 2017

Tid for eksamen: 09.00–13.00

Oppgavesettet er på 4 sider.

Vedlegg: Tabeller over normal-, t- og  $\chi^2$ -fordelingene

Tillatte hjelpeemidler: Formelsamling for STK1100/STK1110.  
Godkjent kalkulator.

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

### Oppgave 1

La  $X_1, \dots, X_n$  være uavhengige og normalfordelte stokastiske variable med forventning  $\mu$  og varians  $\sigma^2$ . De observerte verdiene av  $X_i$  angis med  $x_i$ .

**a**

Anta at  $\mu$  og  $\sigma^2$  er ukjente, men estimert med henholdsvis  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 2.72$  og  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 3.14$  når vi har  $n = 25$  observasjoner.

Test nullhypotesen  $H_0 : \mu = 2$  mot  $H_a : \mu \neq 2$  med signifikansnivå  $\alpha = 0.05$  ved en vanlig t-test. Formuler en konklusjon på testen.

Angi spesielt et intervall for testens P-verdi.

**b**

Formelen for et  $(1 - \alpha)100\%$  konfidensintervall for  $\mu$  i denne situasjonen gis ved  $\bar{x} \pm t_{\alpha/2}s/\sqrt{n}$  der  $t_\alpha$  er  $(1 - \alpha)100$ -persentilen i t-fordelingen med  $n - 1$  frihetsgrader og  $s = \sqrt{s^2}$ .

Vis at dette konfidensintervallet er lik mengden av  $\mu_0$  slik at t-testen for  $H_0 : \mu = \mu_0$  mot  $H_a : \mu \neq \mu_0$  med nivå  $\alpha$  ikke leder til forkastning.

Illustrer dette resultatet på dataene fra punkt a med  $n = 25$ ,  $\bar{x} = 2.72$ ,  $s^2 = 3.14$  og  $\alpha = 0.05$ .

**c**

En tests styrkefunksjon er definert ved  $\gamma(\theta) = P(\text{Forkaste nullhypotesen} \mid \theta)$  der  $\theta$  er den faktiske verdien av parameteren  $\theta$ .

Hva er sammenhengen mellom sannsynligheten for Type II feil og styrkefunksjonen?

(Fortsettes på side 2.)

Anta at variansen  $\sigma^2$  til  $X_i$ -ene definert først i oppgaven er kjent. Vis at med en test som gir forkastning hvis  $|\bar{x} - \mu_0| \sqrt{n}/\sigma > z_{\alpha/2}$  der  $z_\alpha$  er  $(1 - \alpha)100$ -percentilen i standardnormalfordelingen får man styrkefunksjon

$$\gamma(\mu) = 1 - \Phi(z_{\alpha/2} - (\mu - \mu_0)\sqrt{n}/\sigma) + \Phi(-z_{\alpha/2} - (\mu - \mu_0)\sqrt{n}/\sigma)$$

der  $\Phi(z)$  er kumulativ fordelingsfunksjon for standardnormalfordelingen.

## Oppgave 2

La  $X$  være antall dødsulykker i trafikken i Norge i 2016. Siden  $X$  er en tellevariabel skal vi modellere denne med en Poissonfordeling med en forventning  $\lambda$ , kort skrives dette  $X \sim Po(\lambda)$ .

**a**

Vis at maximum likelihood estimatoren for  $\lambda$  gis ved  $\hat{\lambda} = X$ .

Hvorfor er  $\hat{\lambda}$  også momentestimator?

**b**

Når  $\lambda$  er stor gjelder tilnærmet at  $Z = (\hat{\lambda} - \lambda)/\sqrt{\hat{\lambda}}$  er standardnormalfordelt. Dette skal du ikke vise.

Forklar hvorfor  $\hat{\lambda} \pm 1.96\sqrt{\hat{\lambda}}$  er et tilnærmet 95% konfidensintervall for  $\lambda$ .

Beregn intervallet basert på at det var  $X = 135$  dødsulykker i trafikken i Norge i 2016.

Er det grunn til å tro at tilnærmingen er god?

**c**

Et alternativt, også tilnærmet, 95% konfidensintervall for  $\lambda$  gis ved uttrykket  $\hat{\lambda} + 1.96^2/2 \pm \sqrt{4 * 1.96^2 \hat{\lambda} + 1.96^4}/2$ . (Med  $\hat{\lambda} = 135$  blir dette intervallet lik (114.07, 159.77)).

Vis hvordan man kan utlede dette konfidensintervallet.

Beskriv også hvordan intervallet kan avleses grafisk.

## Oppgave 3

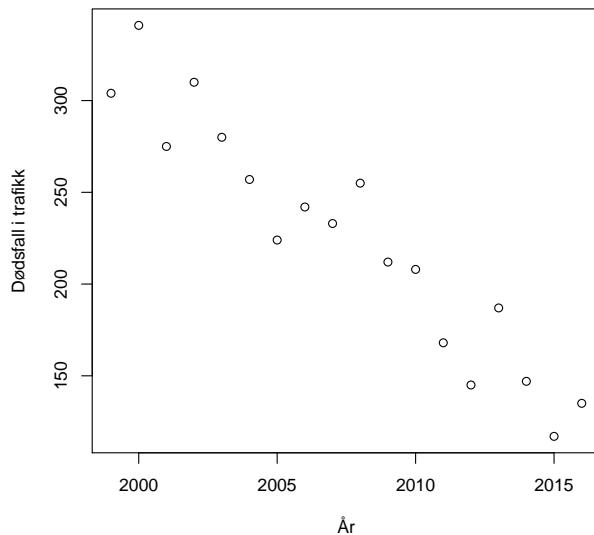
Også i denne oppgaven skal du se på data om dødsulykker i trafikken i Norge, nå i perioden 1999-2016. På neste side er det satt opp en tabell over antall dødsfall for hvert av disse årene samt en figur over utviklingen.

La  $Y_i$  være antall dødsulykker i år nummer  $i$ . Vi skal anta en enkel lineær regresjonsmodell for dataene slik at  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  der  $x_i$  er årstall-2000 (og altså tar verdier -1 til 16) og  $\varepsilon_i$  er uavhengige feilredd med forventning 0 og varians  $\sigma^2$ . I utskriften på neste side er det satt opp et sammendrag av analysen.

(Fortsettes på side 3.)

År	1999	2000	2001	2002	2003	2004	2005	2006	2007
Dødsfall	304	341	275	310	280	257	224	242	233
År	2008	2009	2010	2011	2012	2013	2014	2015	2016
Dødsfall	255	212	208	168	145	187	147	117	135

Trafikkdødsfall i Norge 1999–2016



Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	311.4104	8.6506	36.00	< 2e-16 ***
aar	-11.5955	0.9486	-12.22	1.57e-09 ***
---				

Residual standard error: 20.88 on 16 degrees of freedom  
Multiple R-squared: 0.9033, Adjusted R-squared: 0.8972

**a**

Angi og fortolk parameterestimatene i modellen fra utskriften.

Finn et 95% konfidensintervall for  $\beta_1$  og test hypotesen  $H_0 : \beta_1 = -10$  mot  $H_a : \beta_1 \neq -10$ .

Diskuter graden av samsvar mellom  $x_i$ -er og  $Y_i$ -er.

**b**

Parameterestimatene for  $\beta_0$  og  $\beta_1$  er minste kvadraters estimatorer. Forklar hva dette innebærer.

Angi en betingelse for at de også er maksimum likelihood estimatorer. Vis denne egenskapen.

(Fortsettes på side 4.)

**c**

Minste kvadraters estimatoren  $\hat{\beta}_1$  for  $\beta_1$  kan skrives som følgende uttrykk:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2}$$

hvor summen er over alle  $i$  og  $\bar{Y}$  og  $\bar{x}$  er gjennomsnittet av henholdsvis  $Y_i$ -er og  $x_i$ -er. Dette skal du ikke vise.

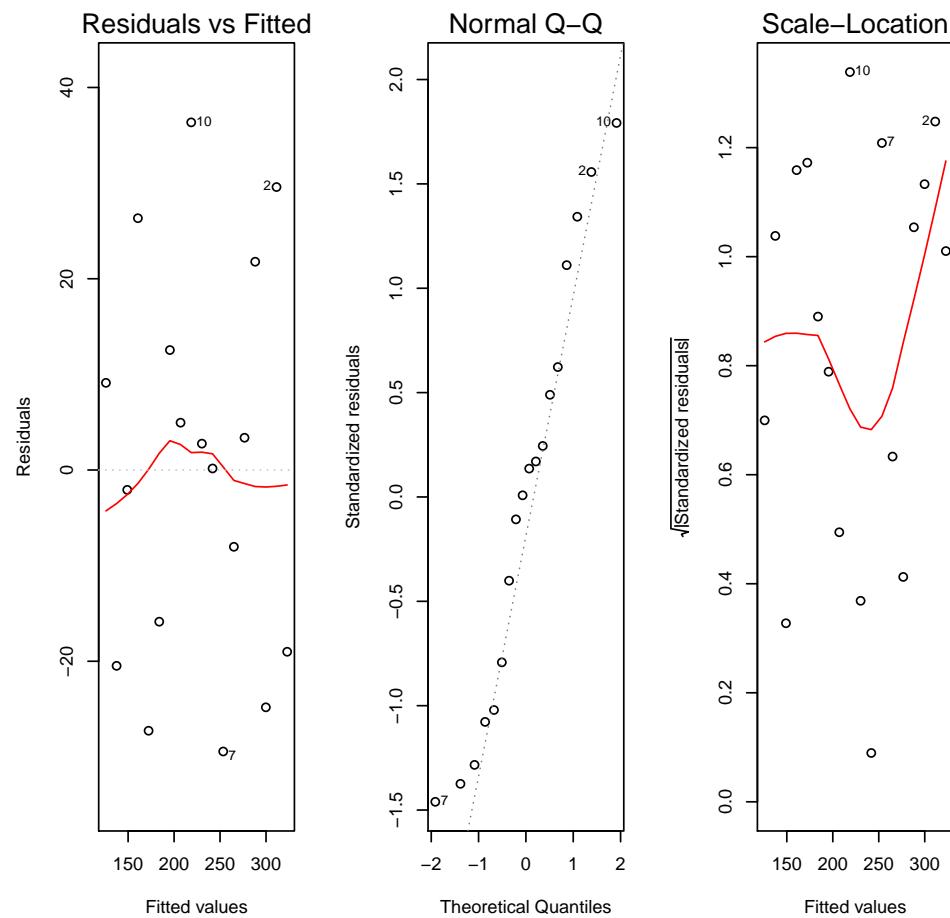
Vis at  $\hat{\beta}_1$  er forventningsrett.

Utled også at variansen til  $\hat{\beta}_1$  kan skrives som  $\sigma^2 / \sum_i (x_i - \bar{x})^2$ .

**d**

Plottene under er residualplottene for regresjonsanalysen. Forklar og fortolk plottene. Finner du betydelig avvik fra modellen?

I Oppgave 2 antok vi at antall dødsfall i trafikken i et år var Poissonfordelt. Passer dette med regresjonsmodellen i denne oppgaven? Kommenter.



SLUTT

# UNIVERSITETET I OSLO

## Matematisk Institutt

EKSAMEN I: **STK 1110 – Statistiske metoder og dataanalyse 1**  
TID FOR EKSAMEN: **Mandag 28. november 2016, kl. 14:30–18:30**  
HJELPEMIDLER: **Formelsamling til STK 1100 og STK 1110,  
godkjent kalkulator**

Dette eksamenssettet inneholder fire oppgaver og er på seks sider (inkludert et kort appendiks på siste side, som kan være til hjelp under løsningen av visse punkter).

### Oppgave 1

I denne oppgaven skal vi se på problemstillinger knyttet til den eksponentielle fordelingen, med først ett og deretter to utvalg.

- (a) Anta at observasjoner  $X_1, \dots, X_n$  er uavhengige og eksponentialfordelte med rate  $\theta$ , altså at tettheten for en enkeltobservasjon er på formen

$$f(x, \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) \quad \text{for } x > 0.$$

Vis at variablene

$$V_i = \frac{2X_i}{\theta}$$

har  $\chi^2$ -fordelingen med 2 frihetsgrader. – Her minner jeg om at  $\chi^2$ -fordelingen med  $\nu$  frihetsgrader har tettheten

$$h(z) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} z^{\nu/2-1} \exp(-\frac{1}{2}z) \quad \text{for } z > 0,$$

og at dennes forventning og varians er  $\nu$  og  $2\nu$ .

- (b) Vis at  $X_i$  har forventning  $\theta$  og varians  $\theta^2$ .  
(c) Vis at  $\hat{\theta} = \bar{X}$ , gjennomsnittet  $(1/n) \sum_{i=1}^n X_i$  av de  $n$  observasjonene, er en forventningsrett estimator for  $\theta$ . Finn dens varians.  
(d) Forklar hva sentralgrenseteoremet (the central limit theorem) medfører for fordelingen for  $\hat{\theta}$ , og bruk dette til å sette opp et konfidensintervall for  $\theta$ , med dekningsgrad tilnærmet lik 95%.  
(e) Anta vi observerer  $X_1, \dots, X_{20}$  fra laboratorium A, antatt uavhengige og eksponentielle med rate  $\theta_1$ , samt  $Y_1, \dots, Y_{20}$  fra laboratorium B, uavhengige og eksponentielle med rate  $\theta_2$ . Vi vil teste hypotesen  $H_0$  at  $\theta_1 = \theta_2$ . Her har  $X_i$ -ene gjennomsnitt  $\bar{X} = 2.222$  og empirisk standardavvik 2.468, mens  $Y_i$ -ene har gjennomsnitt  $\bar{Y} = 4.444$  og empirisk standardavvik 4.987. Lag en testobservator for  $H_0$ , av typen

$$t = \frac{\bar{Y} - \bar{X}}{W},$$

der du skal konstruere  $W$  slik at  $t$  har en tilnærmet  $N(0, 1)$ -fordeling dersom  $H_0$  er korrekt. Beregn  $t$  for denne situasjonen, og kommenter det du finner ut.

- (f) Finn også et 95% konfidensintervall, eksakt eller tilnærmet, for brøkparameteren  $\rho = \theta_2/\theta_1$ , basert på disse dataverdiene. Du kan her få bruk for verdier gitt i tabellen bakerst i oppgavesettet.

## Oppgave 2

En bestemt type operasjon blir jevnlig utført på barn under ett år med en viss type dramatisk hjertelidelse (der barna vil dø om de ikke blir operert). Operasjonen ender av og til med at barnet ikke kan reddes. Tabellen under viser dødsratene, i prosent, for elleve britiske sykehus, over en periode på noen få år. I tillegg til å forstå risikoen for død, samt hvilken type omstendigheter som øker denne risikoen, er man interessert i å finne ut om det er systematiske forskjeller mellom de elleve sykehus, eller om dødsrisikoen for små barn med denne ekstreme hjertelidelsen essensielt er den samme ved hvert av dem.

1	Leicester	13.37
2	Leeds	7.43
3	Oxford	18.85
4	Guys	15.24
5	Liverpool	10.37
6	Southampton	10.04
7	Great Ormond St	11.00
8	Newcastle	13.33
9	Harefield	14.12
10	Birmingham	9.98
11	Brompton	10.30

En mer detaljert analyse er mulig, men i denne oppgaven skal vi for enkelhets skyld se på disse estimatene som resultatet av uavhengige stokastiske variable

$$\hat{\theta}_i \sim N(\theta_i, \sigma^2) \quad \text{for } i = 1, \dots, 11,$$

der standardavviket  $\sigma = 2.112$  er såpass godt estimert at den kan anses som en kjent verdi. Parameterverdiene  $\theta_1, \dots, \theta_{11}$  ses på som de reelle, underliggende rater, som vi altså ikke kan observere eksakt, kun estimere, med de data vi har (som er det vi har gjort over).

- (a) Finn to 95% konfidensintervall, ett for parameteren  $\theta_1$  og ett for  $\theta_2$ , dødsfrekvensen for slike operasjoner utført ved sykehusene i henholdsvis Leicester og Leeds.
- (b) For å analysere i hvilken grad  $\theta_i$ -ene varierer fra sykehus til sykehus, eller om de kanskje er essensielt like, tenker vi oss at  $\theta_1, \dots, \theta_{11}$  selv er uavhengige stokastiske variable, fra fordelingen

$$\theta_i \sim N(\theta_0, \tau^2).$$

Her er  $\theta_0$  og  $\tau$  faste, men ukjente, parametre. Vi kan skrive

$$\theta_i = \theta_0 + \delta_i \quad \text{og} \quad \hat{\theta}_i = \theta_i + \varepsilon_i \quad \text{for } i = 1, \dots, 11,$$

der  $\delta_i \sim N(0, \tau^2)$  og  $\varepsilon_i \sim N(0, \sigma^2)$  er uavhengige. Forklar hvorfor dette leder til

$$\hat{\theta}_i \sim N(\theta_0, \sigma^2 + \tau^2) \quad \text{for } i = 1, \dots, 11.$$

(c) For estimatene over kan vi beregne

$$S^2 = \frac{1}{10} \sum_{i=1}^{11} (\hat{\theta}_i - \bar{\theta})^2 = 10.0607 = 3.1719^2,$$

der  $\bar{\theta}$  er gjennomsnittet  $(1/11) \sum_{i=1}^{11} \hat{\theta}_i = 12.1857$ . Bruk dette til å estimere  $\tau$ .

- For punktene (d) og (e) under kan du anvende at kvantilene 0.025, 0.050, 0.500, 0.950, 0.975 for  $\chi^2_{10}$ , altså  $\chi^2$ -fordelingen med 10 frihetsgrader, er  
 $3.2470 \quad 3.9403 \quad 9.3418 \quad 18.3070 \quad 20.4832$
- (d) Lag en test med nivå 0.05 for hypotesen om at  $\theta_i$ -ene er helt like, mot alternativet at det altså er forskjeller. Utfør testen.
- (e) Lag til sist et 95% konfidensintervall for  $\tau$ .

### Oppgave 3

Anta at  $X$  er binomisk fordelt  $(n, p)$ , altså med de klassiske punktsannsynligheter

$$f(x, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

- (a) For et gitt utfall  $x$ , sett opp log-likelihood-funksjonen  $\ell(p)$ , og vis at denne maksimeres for  $\hat{p} = x/n$ .
- (b) Anta at vi har tre uavhengige binomiske eksperimenter,

$$X \sim \text{bin}(n, p), \quad Y \sim \text{bin}(n, q), \quad Z \sim \text{bin}(n, r).$$

Sett opp et uttrykk for log-likelihood-funksjonen  $\ell(p, q, r)$ . Dersom hypotesen  $H_0$  at de tre sannsynlighetene er like, altså  $p = q = r$ , holder, hva er da maximum-likelihood-estimatet for den felles verdi av denne sannsynligheten?

- (c) Anta at  $n = 50$  og at man i de tre eksperimenter observerer  $X = 17, Y = 22, Z = 14$ . Finn verdier for

$$\ell_{\max}(\text{big}) \quad \text{og} \quad \ell_{\max}(H_0),$$

der  $\ell_{\max}(\text{big})$  er maksimum av log-likelihood-funksjonen under modellen der  $p, q, r$  er frie parametre, og  $\ell_{\max}(H_0)$  tilsvarende maksimum av samme funksjon under  $H_0$ . Utfør en test for  $H_0$  og kommenter det du finner.

### Oppgave 4

Tabellen under viser et helsepolitisk og historisk viktig datasett, fra 1963, idet det ble benyttet for å arbeide frem og etter noen år gjennomføre vedtak, i flere land, om at sigarettpakker skulle utstyres med en advarsel. For hvert av de angitte land vises

$x$  = gjennomsnittlig antall sigaretter pr. person pr. år,

$y$  = antall døde, av hjerte-og-karsykdommer, pr. hundre tusen innbyggere.

Gjennomsnittstallet  $x$  omfatter altså både røykere og ikke-røykere, av alle over 18 år. Populasjonene det ble rapportert om for  $y$ , i disse undersøkelsene, gjelder dem fra 35 til 64 år.

	<i>x</i>	<i>y</i>	
1962	3350	211.6	Canada
1962	3220	238.1	Australia
1962	3220	211.8	New Zealand
1962	2790	194.1	United Kingdom
1962	2780	124.5	Switzerland
1962	2770	187.3	Ireland
1962	2290	110.5	Iceland
1962	2160	233.1	Finland
1963	1890	150.3	West Germany
1962	1810	124.7	Netherlands
1962	1800	41.2	Greece
1962	1770	182.1	Austria
1962	1700	118.1	Belgium
1962	1680	31.9	Mexico
1963	1510	114.3	Italy
1961	1500	144.9	Denmark
1962	1410	59.7	France
1962	1270	126.9	Sweden
1961	1200	43.9	Spain
1962	1090	136.3	Norway
1962	3900		USA

En enkel lineær regresjonsmodell setter

$$y_i = a + bx_i + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

der  $\varepsilon_i$ -ene tenkes uavhengige og  $N(0, \sigma^2)$ . Her tar vi med de  $n = 20$  land i tabellen over, der vi har både  $x$  og  $y$ ; USA er altså ikke med i denne delen av analysen. Den vanlige kommandoen `lm(y ~ x)` i programpakken **R** gir bl.a. følgende:

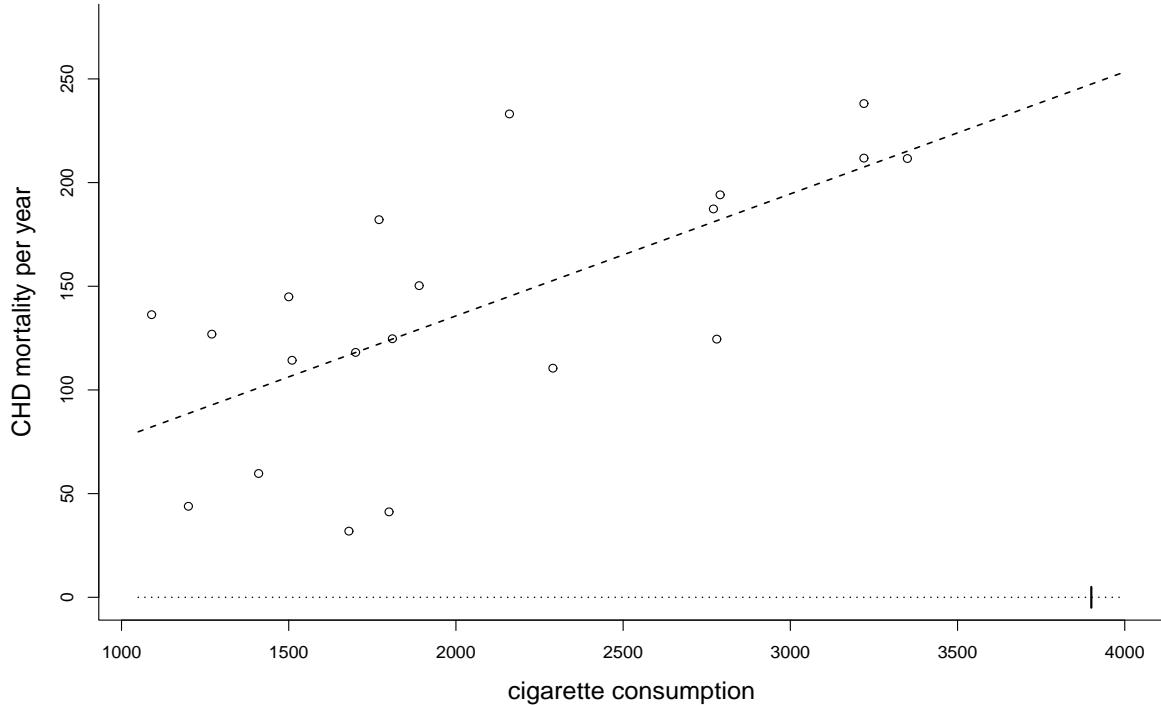
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	18.03462	33.28179	0.542	0.59455
x	0.05884	0.01529	3.848	0.00118 **

Jeg opplyser også om at  $\sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 = 41388.09$ , der  $\hat{a}$  og  $\hat{b}$  er minste-kvadratsums-estimatene. Dataene, med regresjonslinjen, er også vist frem i figuren neste side.

- (a) Gi verdiene for  $\hat{a}$  og  $\hat{b}$ , og vis at det vanlige estimatet for  $\sigma$ , det som tar utgangspunkt i forventningsretthet for  $\hat{\sigma}^2$ , blir  $\hat{\sigma} = 47.951$ .
- (b) Dersom sigarettkonsumet i et land går ned, så meget at gjennomsnittstallet for antall sigareetter går ned med 100 (pr. person pr. år), hvor mange færre døde vil da dette landet senere kunne forvente, av hjerte-og-karsykdommer, i dette alderssegmentet fra 35 til 64 år?

- (c) Sambandsstatene (USA) hadde altså så høyt sigarettkonsum i 1962 som 3900 pr. voksenperson pr. år. La  $Y_0$  være antallet døde av hjerte-og-karsykdommer i 1962, i USA, per hundre tusen, i alderen 35 til 64 år. Gi et estimat for  $Y_0$ .



- (d) I tillegg til estimatet  $\hat{y}_0$  for  $Y_0 = a + bx_0 + \varepsilon_0$  ønsker vi oss et fullt prediksjonsintervall. Vi innfører  $Z_0 = Y_0 - \hat{a} - \hat{b}x_0$ , der  $x_0 = 3900$ . Finn forventning og varians til  $Z_0$ . Her kan du bruke følgende, fra pensum:

$$\hat{a} + \hat{b}x_0 = \bar{Y} + \hat{b}(x_0 - \bar{x}),$$

der  $\bar{x} = (1/n) \sum_{i=1}^n x_i = 2060.500$  og  $\bar{Y} = (1/n) \sum_{i=1}^n Y_i = 139.265$  er gjennomsnitene, og  $\hat{b} = \sum_{i=1}^n (x_i - \bar{x})Y_i/M$ , med  $M = \sum_{i=1}^n (x_i - \bar{x})^2 = 9833895$ .

- (e) Beregn et intervall som med sannsynlighet 90% (eksakt eller tilnærmet) inneholder  $Y_0$ . Kommenter kort forutsetninger som ligger til grunn. Du kan få bruk for at 0.95-kvantilen for  $t$ -fordelingen med 18 frihetsgrader er  $qt(0.95, 18) = 1.734$ .
- (f) Ifølge offisielt tilgjengelige kilder har Ungarn i 2014 en dødsrate på 172.6 pr. hundre tusen, relatert til hjerte-og-karsykdommer. Hva tror du sigarettkonsumet er i Ungarn?

### Appendiks: en liten F-tabell

Her er en tabell over (low, up), 0.025-kvantilen og 0.975-kvantilen i  $F$ -fordelingen (Fisherfordelingen) med frihetsgrader ( $m_1, m_2$ ), der vi i denne situasjonen kun ser på  $m_1 = m_2 = m$ , og for  $m$ -verdier fra 15 til 45.

m1	m2	low	up
15	15	0.349	2.862
16	16	0.362	2.761
17	17	0.374	2.673
18	18	0.385	2.596
19	19	0.396	2.526
20	20	0.406	2.464
21	21	0.415	2.409
22	22	0.424	2.358
23	23	0.433	2.312
24	24	0.441	2.269
25	25	0.448	2.230
26	26	0.456	2.194
27	27	0.463	2.161
28	28	0.470	2.130
29	29	0.476	2.101
30	30	0.482	2.074
31	31	0.488	2.049
32	32	0.494	2.025
33	33	0.499	2.002
34	34	0.505	1.981
35	35	0.510	1.961
36	36	0.515	1.942
37	37	0.520	1.924
38	38	0.524	1.907
39	39	0.529	1.891
40	40	0.533	1.875
41	41	0.538	1.860
42	42	0.542	1.846
43	43	0.546	1.833
44	44	0.550	1.820
45	45	0.553	1.807

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse 1

Eksamensdag: Mandag 30. november 2015.

Tid for eksamen: 14.30 – 18.00.

Oppgavesettet er på 5 sider.

Vedlegg: Ingen

Tillatte hjelpeemidler: Godkjent kalkulator og formelsamling  
for STK1100/STK1110

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

Nedenfor får du oppgitt følgende øvre kvantiler i standard normal fordelingen som kan brukes ulike steder i oppgavesettet.

$\alpha$	0.100	0.050	0.025	0.010	0.001	0.0001	0.00001	0.000001
$z_\alpha$	1.281	1.645	1.960	2.326	3.090	3.719	4.265	4.753

### Oppgave 1

Vi har observert  $n = 20$  datapunkter  $x_1, \dots, x_{20}$ :

$$\begin{array}{cccccccccccc} 13.46 & 13.85 & 5.73 & 0.82 & 4.46 & 17.37 & 2.20 & 10.29 & 4.36 & 9.15 \\ 4.12 & 2.49 & 9.99 & 3.95 & 13.25 & 4.90 & 19.27 & 5.36 & 16.42 & 4.96 \end{array}$$

Du får også oppgitt at  $\sum_{i=1}^n x_i = 166.40$  og  $\sum_{i=1}^n \ln(x_i) = 37.15$ . Vi vil anta observasjonene er uavhengige og identiske fordelte fra en kontinuerlig fordeling med sannsynlighetstetthetsfunksjon

$$f(x; \theta) = \frac{1}{\theta^2} x e^{-x/\theta}, \quad 0 < x < \infty; 0 < \theta < \infty$$

En kan vise at  $E[X] = 2\theta$  og  $V(X) = 2\theta^2$  for denne fordelingen (dette behøver du ikke å vise).

(Fortsettes på side 2.)

- (a) Sett opp likelihoodfunksjonen og vis at log-likelihoodfunksjonen kan skrives som

$$l(\theta) = -2n \ln \theta + \sum_{i=1}^n \ln(x_i) - \frac{1}{\theta} \sum_{i=1}^n x_i.$$

- (b) Forklar prinsippet bak maksimum likelihood (ML) estimering og vis at ML estimatoren for  $\theta$  er

$$\hat{\theta} = \frac{1}{2n} \sum_{i=1}^n x_i.$$

- (c) Vis at  $\hat{\theta}$  også er momentestimatoren for  $\theta$ .

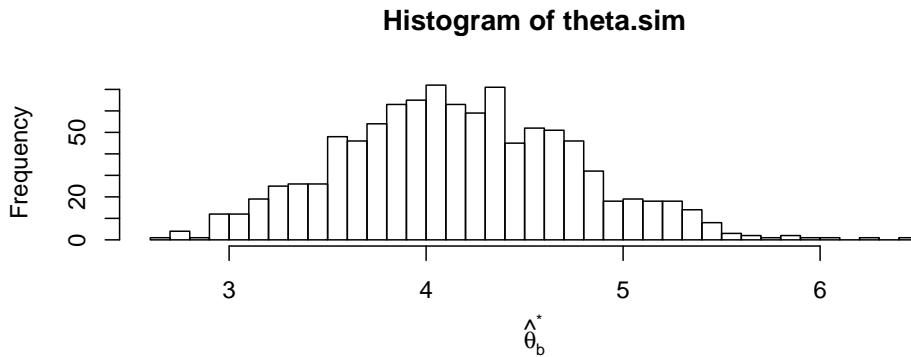
Er  $\hat{\theta}$  forventningsrett?

- (d) Regn ut  $\hat{\theta}$  for de gitte data. Hva blir standardfeilen for  $\hat{\theta}$  og estimatelet på denne i dette tilfellet?

Generelle resultater om ML estimatorer tilsier at  $\hat{\theta}$  er tilnærmet normalfordelt for  $n$  stor (dette behøver du ikke å vise).

- (e) Utled et (tilnærmet) 95% konfidensintervall for  $\theta$ . Hva blir dette intervallet for de gitte data?

Nedenfor er vist et histogram av bootstrapsimuleringer av  $\hat{\theta}^*$  basert på ikke-parametrisk bootstrapping.



Tabellen nedenfor viser også ulike empiriske kvantiler av de simulerte  $\hat{\theta}^*$ :

Kvantil	0.01	0.025	0.05	0.95	0.975	0.99
Verdil	2.83	2.98	3.14	5.21	5.40	5.59

- (f) Forklar hva vi mener med ikke-parametrisk bootstrapping. Bruk bootstrapsimuleringene til å lage et 95% konfidensintervall.

Diskuter likheter/forskjeller i forhold til intervallet du fant i (e).

(Fortsettes på side 3.)

## Oppgave 2

Data som vi vil bruke i denne oppgaven er knyttet til prostatakreft og er hentet fra Stamey et al. (1989). Responsvariabelen angir nivå av en prostata-spesifikk antigen `lpsa` mens en rekke mulige forklaringsvariable (ulike kliniske målinger) samtidig er samlet inn. Vi vil her konsentrere oss om 3 av disse, prostata volum på log-skala (`lcavol`), vekt av prostata på log-skala (`lweight`) og alder (`age`).

Tabellen nedenfor viser de 6 første av totalt 97 målinger:

	<code>lpsa</code>	<code>lcavol</code>	<code>lweight</code>	<code>age</code>
1	-0.4307829	-0.5798185	2.769459	50
2	-0.1625189	-0.9942523	3.319626	58
3	-0.1625189	-0.5108256	2.691243	74
4	-0.1625189	-1.2039728	3.282789	58
5	0.3715636	0.7514161	3.432373	62
6	0.7654678	-1.0498221	3.228826	50

Vi vil i første omgang se på en enkel regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, i = 1, \dots, n \quad (*)$$

der  $x_{i1}$  er `lcavol` mens  $y_i$  er `lpsa`. Her følger  $\varepsilon_i$ -ene de vanlige antagelsene, dvs uavhengige og  $N(0, \sigma^2)$  fordelte. En utskrift fra tilpasning av en slik modell til de gitte data er gitt nedenfor.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.50730	0.12194	12.36	<2e-16
<code>lcavol</code>	0.71932	0.06819	10.55	<2e-16

Residual standard error: 0.7875 on 95 degrees of freedom

Multiple R-squared: 0.5394, Adjusted R-squared: 0.5346

F-statistic: 111.3 on 1 and 95 DF, p-value: < 2.2e-16

- (a) Vis at maksimum likelihood estimatorene for  $\beta_0$  og  $\beta_1$  svarer til minste kvadraters estimatorene.

(Du behøver ikke å utlede selve formlene, kun vise at det svarer til samme prinsipp).

- (b) Vi ønsker å utføre en test på  $H_0 : \beta_1 = 0$  mot  $H_a : \beta_1 \neq 0$ . Spesifiser hva slags test-observator du kan bruke til dette og utled fordelingen til denne.

Utfør testen basert på utskriften ovenfor og konkluder.

(Fortsettes på side 4.)

Vi vil nå utvide modellen ovenfor til

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, i = 1, \dots, n \quad (**)$$

der  $x_{i2}$  er `lweight` og  $x_{i3}$  er `age`. En tilpasning av denne modellen til de gitte data ga følgende resultat:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.146941	0.772372	0.190	0.84953
lcavol	0.687819	0.067418	10.202	< 2e-16
lweight	0.549937	0.163838	3.357	0.00114
age	-0.009486	0.011003	-0.862	0.39081

Residual standard error: 0.7517 on 93 degrees of freedom

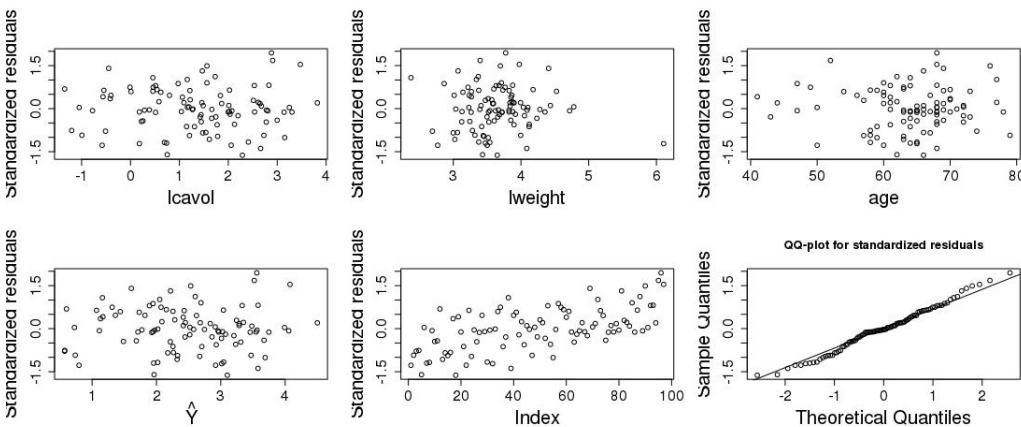
Multiple R-squared: 0.5892, Adjusted R-squared: 0.576

F-statistic: 44.47 on 3 and 93 DF, p-value: < 2.2e-16

- (c) Basert på resultatene, vil du si at modell (\*\*) ovenfor er en bedre modell enn modell (\*) på forrige side? Begrunn svaret.
- (d) For å sjekke en modell, bruker vi ofte residualene, som på vektorform kan defineres ved  $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$  der  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ .

Vis at  $\mathbf{E} = [\mathbf{I} - \mathbf{H}] \mathbf{Y}$  (der  $\mathbf{I}$  er identitetsmatrisen mens du selv må finne ut hva  $\mathbf{H}$  er) og bruk dette til å vise at residualene er normalfordelte med forventning 0 og varians  $\sigma^2(1 - h_{ii})$  for den  $i$ te residual. Her er  $h_{ii}$  diagonalelement nr  $i$  av  $\mathbf{H}$ .

Plottet nedenfor viser ulike residualplott (basert på standardiserte residualer).



- (e) Forklar hva standardiserte residualer er og hvorfor det er mer hensiktsmessig å bruke disse enn residualene selv.

Basert på disse plottene, diskuter om antagelsene som ligger til grunn for modell (\*\*) er rimelige.

(Fortsettes på side 5.)

## Oppgave 3

Tabellen nedenfor viser antall personer involvert i alvorlige sykkelykninger (fordelt på kjønn) samt hvor mange av de som ble testet positiv for alkohol.

Kjønn	$n$	$Y$ (testet positiv)	$\hat{p}$ (andel testet positiv)
Menn	1520	515	0.339
Kvinner	191	27	0.141

Vi vil anta at alle disse ulykkene er skjedd uavhengige av hverandre.

La  $p_M$  være sannsynligheten for at menn som er involvert i alvorlige sykkelykninger tester positiv på alkohol og  $p_K$  tilsvarende for kvinner. Vår interesse vil være i å sammenlikne  $p_K$  med  $p_M$ .

(a) La

$$Z = \frac{\hat{p}_K - p_K}{\sqrt{p_K(1 - p_K)/n_K}}.$$

Her er  $\hat{p}_K = Y_K/n_K$  der  $Y_K$  er  $Y$  tilhørende gruppen av kvinner og tilsvarende for  $n_K$ .

Hva blir forventning og varians til  $Z$ ? Hva slags fordeling har  $Z$  tilnærmet når  $n_K$  er stor?

(b) Utfør en test av

$$H_0 : p_K = 0.339 \text{ mot } H_a : p_K \neq 0.339.$$

Hva blir konklusjonen hvis du velger signifikansnivå  $\alpha = 0.05$ ?

(c) Test så istedet

$$H_0 : p_K = p_M \text{ mot } H_a : p_K \neq p_M$$

Hva blir konklusjonen i dette tilfellet? Angi i dette tilfellet en øvre grense for P-verdien til testen.

(d) Argumenter for hvorfor testen i (c) er mer fornuftig å bruke enn testen i (b).

Hvorfor får vi ikke så veldig forskjellige svar i (b) og (c) i dette tilfellet?

(e) Forklar hvordan du kunne brukt logistisk regresjon for å teste

$$H_0 : p_K = p_M \text{ mot } H_a : p_K \neq p_M.$$

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: Mandag 1. desember 2014.

Tid for eksamen: 14.30–18.30.

Oppgavesettet er på 5 sider.

Vedlegg: Tabell over normal-,  $t$ -, og  $\chi^2$ -fordeling.

Tillatte hjelpeemidler: Godkjent kalkulator og formelsamling for STK1100/STK1110.

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

**Oppgave 1.** Vi skal undersøke hyppigheten av ulykker i forbindelse med dykkerarbeid i Nordsjøen. La  $\lambda$  være en parameter som angir forventet antall ulykker per timeverk. Vi har samlet inn data fra  $n$  forskjellige offshore-anlegg et bestemt år. La  $X_i$  betegne antall ulykker i løpet av et år ved  $i$ -te anlegg,  $i = 1, 2, \dots, n$ . Hver ulykke antas å opptre uavhengig av alle andre ulykker. Det er da rimelig å anta at  $X_i$  er Poisson-fordelt med parameter  $\lambda t_i$ , der  $t_i$  er totalt antall timeverk per år i  $i$ -te anlegg. Dvs.

$$P(X_i = x_i) = \frac{(\lambda t_i)^{x_i}}{x_i!} \exp(-\lambda t_i), \quad x_i = 0, 1, 2, \dots$$

Videre er  $X_1, X_2, \dots, X_n$  uavhengige tilfeldige variable.

a) En mulig estimator for  $\lambda$  er

$$\hat{\lambda} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n t_i}.$$

Vis at  $\hat{\lambda}$  er en forventningsrett estimator for  $\lambda$ . Finn variansen til  $\hat{\lambda}$ . Hva blir estimatet av  $\lambda$  med data for  $n = 5$  anlegg gitt til slutt i oppgaven ( neste side)?

b) Vis at  $\hat{\lambda}$  er sannsynlighetsmaksimerings-estimatoren (maximum likelihood estimator) for  $\lambda$ .

c) En annen mulig estimator for  $\lambda$  er

$$\hat{\lambda} = \frac{\sum_{i=1}^n X_i t_i}{\sum_{i=1}^n t_i^2}.$$

Hvilken av de to estimatorene vil du foretrekke? Begrunn svaret ditt.

(Fortsettes på side 2.)

Anlegg	1	2	3	4	5
$t_i$	$3.00 \cdot 10^4$	$2.00 \cdot 10^4$	$0.87 \cdot 10^4$	$2.50 \cdot 10^4$	$0.63 \cdot 10^4$
$x_i$	9	4	4	7	3

**Oppgave 2.** De siste ukene, spesielt etter fremlegging av statsbudsjettet, har opposisjonspartiet Ap opplevd rekordhøy oppslutning i flere meningsmålinger. Siste partibarometer fra 12. november 2014 viste at 42.2 % av de spurte ville ha stemt Ap hvis det hadde vært valg denne dagen. Vi kan lese på nettsidene til Opinion Perduco, som har gjennomført målingen, at de har spurt 960 personer.

- a) Kan vi med dette si at det er nesten sikkert at Ap ville ha fått minst 40 % av stemmene i et eventuelt stortingsvalg den dagen? Besvar spørsmålet ved å formulere og gjennomføre en hypotesetest.

**Oppgave 3.** Forurensninger i vannet på to strender A og B i Oslofjorden skal undersøkes. Det mistenkes at strand A er mer forurenset enn strand B. Det blir tatt 9 prøver, 4 fra strand A og 5 fra strand B. La  $X_1, X_2, \dots, X_4$  betegne målingene fra strand A, som kan antas uavhengige identisk fordelte variable fra normalfordelingen  $N(\mu_A, \sigma^2)$ . La  $Y_1, Y_2, \dots, Y_5$  være målingene fra strand B, som kan antas uavhengige identisk fordelte fra normalfordelingen  $N(\mu_B, \sigma^2)$ . Målingene fra strand A og strand B er også uavhengige av hverandre. Vi skal først anta at variansen  $\sigma^2$  er kjent lik 4.

Observasjonene var

Forurensning strand A ( $x_i$ ) (i ppm): 16, 12, 14, 10

Forurensning strand B ( $y_i$ ) (i ppm): 11, 8, 14, 7, 10

- a) Gir målingene grunnlag for å konkludere med at strand A er mer forurenset enn strand B? Besvar spørsmålet ved å gjennomføre en hypotesetest. Husk å spesifisere hypotesene, beregn forkastningsområdet og skriv en skikkelig konklusjon. Bruk signifikansnivå  $\alpha = 0.01$ .
- b) Hva er en P-verdi? Finn P-verdien for testen ovenfor.
- c) Hvor mange prøver må man ta hvis man skal ha like mange prøver fra A som fra B, og man i testen i a) foruten signifikansnivået på 1% forlangte at sannsynligheten for å forkaste  $H_0$  når  $\mu_A - \mu_B = 4$  skal være minst 95%?
- d) Variansen  $\sigma^2$  er i virkeligheten ukjent og må estimeres. Finn et estimat for variansen basert på de tilsammen 9 målingene fra de to strendene. Du kan bruke at de empiriske variansene for A og B er hhv.  $s_A^2 = 6.67$  og  $s_B^2 = 7.50$ . Gjennomfør testen i a) når variansen er ukjent.

(Fortsettes på side 3.)

- e) Utled et 99% konfidensintervall for forskjellen  $\mu_A - \mu_B$ , fremdeles med ukjent varians som må estimeres. Forklar hvordan et slikt intervall skal tolkes.

**Oppgave 4.** Når et fast stoff skal løses i vann og løsningsprosessen er forbundet med forbruk av varme, stiger løseligheten (målt i gram oppløst stoff per 100 gram vann) med stigende temperatur. Løselighetskurven for et salt S skal bestemmes ved følgende forsøk: 100 g vann varmes opp til  $x$  grader Celsius, og man bestemmer antall gram ( $Y$ ) av saltet som løses i vannet ved denne temperaturen.

Forsøket utføres 18 ganger. Vi antar at  $Y_1, Y_2, \dots, Y_{18}$  er uavhengige og normalfordelte med samme ukjente varians  $\sigma^2$  og forventning gitt ved

$$E(Y_i) = \beta_0 + \beta_1 x_i,$$

der  $x_i$  er temperaturen i forsøk  $i$ . I utskriften fra R nedenfor finner du observasjonene og flere analyser av datasettet.

- a) Hvilke tall i R-utskriften er minste kvadraters estimater for  $\beta_0$ ,  $\beta_1$  og  $\sigma$ ? Diskuter hvor godt modellen passer til data på bakgrunn av utskrift og figurer.
- b) Beregn et 95% konfidensintervall for  $\beta_1$ . Du kan hente ut størrelser du trenger fra R-utskriften.
- c) Forklar hvordan konfidensintervallet ovenfor henger sammen med å teste hypotesen

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_a : \beta_1 \neq 0$$

på signifikansnivå  $\alpha = 0.05$ .

- d) Du vil anslå løseligheten ved 35 grader Celsius. Forklar hva som er forskjellen på et konfidensintervall for forventet løselighet og et prediksjonsintervall for en ny observasjon ved denne temperaturen. Finn begge ved hjelp av utskriften.
- e) Fra teorien burde løselighetskurven for salt S være et polynom heller enn en linje. Skriv opp en regresjonsmodell der du legger til et annengradsledd. En slik analyse er også gjennomført i R, se vedlagte utskrift. Vurder om annengradsmodellen er mer passende enn den lineære.

```
>
> x=c(10, 10, 10, 20, 20, 20, 30, 30, 30, 40, 40, 40, 40, 50, 50, 50, 60, 60, 60)
> y=c(56.3, 59.9, 58.6, 63.7, 63.6, 65.4, 68.3, 70.6, 69.6, 74.1, 76.3, 74.6,
     80.5, 79.7, 79.1, 84.7, 85.9, 84.9)
> plot(x,y)
>
>
>
```

(Fortsettes på side 4.)

```

> fit=lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3603 -0.5203 -0.1732  0.5654  1.6454 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 53.32889   0.55615   95.89 <2e-16 ***
x           0.53314   0.01428   37.33 <2e-16 ***  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 1.035 on 16 degrees of freedom
Multiple R-squared:  0.9887,    Adjusted R-squared:  0.9879 
F-statistic: 1394 on 1 and 16 DF,  p-value: < 2.2e-16

>
> abline(fit)
> plot(fit$res)
> qqnorm(fit$res)
>
> new=data.frame(x=35)
> predict(fit,new,interval="confidence")
    fit      lwr      upr
1 71.98889 71.47187 72.50591
> predict(fit,new,interval="prediction")
    fit      lwr      upr
1 71.98889 69.73523 74.24254

> fit2=lm(y~x+I(x^2))
> summary(fit2)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0726 -0.4243 -0.2073  0.7668  1.5274 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 52.5233333  1.0765078  48.790 < 2e-16 ***
x           0.5935595  0.0704281   8.428 4.51e-07 ***
I(x^2)     -0.0008631  0.0009849  -0.876   0.395  

```

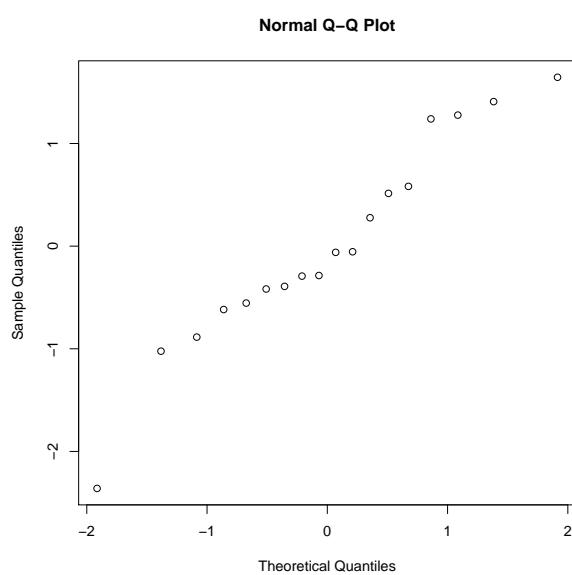
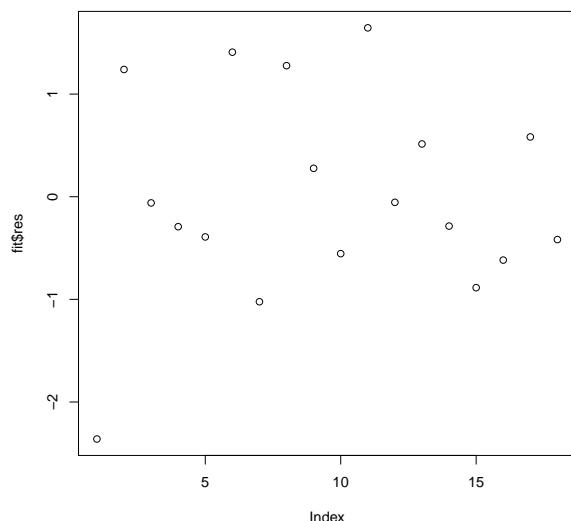
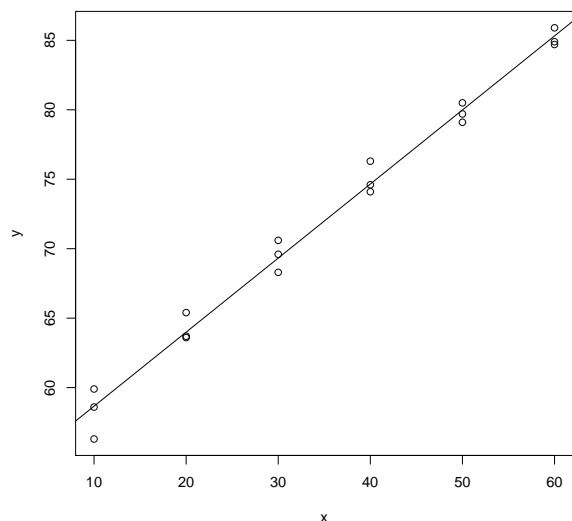
(Fortsettes på side 5.)

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

Residual standard error: 1.042 on 15 degrees of freedom  
 Multiple R-squared: 0.9892, Adjusted R-squared: 0.9878  
 F-statistic: 687.2 on 2 and 15 DF, p-value: 1.777e-15

>



SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: Torsdag 12. desember 2013.

Tid for eksamen: 14.30–18.30.

Oppgavesettet er på 4 sider.

Vedlegg: Tabell over normal-,  $t$ -, og  $\chi^2$ -fordeling.

Tillatte hjelpeemidler: Godkjent kalkulator og formelsamling for STK1100/STK1110.

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

### Oppgave 1.

Samme størrelse måles gjentatte ganger med to forskjellige måleinstrumenter med ulik presisjon. La  $X_1, X_2, \dots, X_m$  betegne  $m$  målinger med instrument A, som kan antas uavhengige identisk fordelte variable fra normalfordelingen  $N(\mu, \sigma_1^2)$ . La  $Y_1, Y_2, \dots, Y_n$  være  $n$  målinger med instrument B, som kan antas uavhengige identisk fordelte fra normalfordelingen  $N(\mu, \sigma_2^2)$ . Målingene fra instrument A og B er også uavhengige av hverandre. De to variansene  $\sigma_1^2$  og  $\sigma_2^2$  er kjent for de to instrumentene.

Vi skal først studere resultatene fra instrument A.

- Utled et 99% konfidensintervall for  $\mu$  basert på de  $m$  observasjonene fra instrument A. Forklar hvordan et slikt intervall skal tolkes.
- Hvor stor må  $m$  være for at lengden på konfidensintervallet for  $\mu$  ikke skal være lenger enn  $w$ ? Hva blir  $m$  dersom  $w = \sigma_1$ ?

Vi er i det videre interessert i å estimere forventningen  $\mu$  fra alle  $n + m$  observasjoner.

- Sett opp uttrykket for likelihood og log likelihood og finn sannsynlighetsmaksimerings-estimatoren (maximum likelihood estimator) for  $\mu$  basert på alle observasjonene.
- Argumenter kort for at formen på estimatoren i c) er rimelig i forhold til hvordan antall observasjoner og presisjon i instrumentene bør påvirke estimeringen av  $\mu$ . Sjekk om estimatoren er forventningsrett, og finn dens varians.

(Fortsettes på side 2.)

**Oppgave 2.** Mange kvinner får problemer med osteoporose (lav bentetthet) etter menopausen. Forskere ved Rikshospitalet studerer hvorledes genetiske faktorer bidrar til denne tilstanden. Vi skal studere et datasett med målt bentetthet for 84 kvinner i den aktuelle alderen, samt genekspresjon (et mål på hvor aktivt et gen er) for 20.000 gener i de samme kvinnenes benceller. Vi skal ikke analysere hele dette høydimensjonale datasettet her, men studere hvordan genekspresjonen til fire utvalgte gener er assosiert med bentetthet hos kvinner etter menopause.

I utskriften nedenfor er variabelen bentetthet kalt 'bone' og de fire genene respektivt 'gene1', 'gene2', 'gene3' og 'gene4'.

- Vi gjennomfører først en univariat analyse med gen 1 alene som kovariat og bentetthet som responsvariabel. Sett opp en enkel lineær regresjonsmodell med vanlige antakelser. Sett opp nullhypotese og alternativ og test om gen 1 er en signifikant forklaringsvariabel for bentetthet. Du skal spesifisere en testobservator, bruke nivå  $\alpha = 0.05$  og beregne forkastningsområde. Sett inn tall fra R-utskrift (I) og konkluder.
- Hva er en P-verdi? Finn en øvre grense for P-verdien for testen ovenfor.
- Beregn et 95% konfidensintervall for stigningstallet i modellen i a). Gjør kort rede for sammenhengen mellom hypotesetesting med tosidig alternativ og konfidensintervall.
- Vi inkluderer deretter gen 2 i en multippel regresjonsanalyse og får resultatene (II) på neste side. Forklar kort hva som skjer med effekten av gen 1 når også gen 2 inkluderes i modellen, og hvorfor dette kan skje. Du kan benytte tilleggsinformasjon som du finner i utskriften etter resultatene (II).
- Til slutt inkluderer vi også gen 3 og gen 4 i modellen. Tilpasning til to alternative modeller finnes i R-utskriften på neste side (III). Velg beste modell, og begrunn valget ditt godt. Skriv en tolkning av de estimerte effektene av genekspresjonene på bentettheten i din utvalgte modell.

```
(I)
Call:
lm(formula = bone ~ gene1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.9309    1.5477   2.54  0.01298  
gene1       -1.1297    0.3609   -3.14 0.00085 ** 
(Edited output)
```

(Fortsettes på side 3.)

(II)

Call:

```
lm(formula = bone ~ gene1 + gene2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.8276	2.0564	3.806	0.000273
gene1	-0.1102	0.5082	-0.217	0.828876
gene2	-1.9110	0.6954	-2.748	0.007387

(edited output)

```
> cor(gene1, gene2)
[1] 0.7300616
```

(III)

Call:

```
lm(formula = bone ~ gene1 + gene2 + gene3 + gene4)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-8.1091	2.9757	-2.725	0.007912
gene1	-0.2141	0.3848	-0.556	0.579455
gene2	-1.8064	0.5182	-3.486	0.000804
gene3	2.4974	0.6097	4.096	0.000101
gene4	1.7865	0.2731	6.541	5.56e-09

Residual standard error: 1.088 on 79 degrees of freedom  
 Multiple R-squared: 0.5588, Adjusted R-squared: 0.5365  
 F-statistic: 25.02 on 4 and 79 DF, p-value: 2.114e-13

Call:

```
lm(formula = bone ~ gene2 + gene3 + gene4)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-8.2321	2.9546	-2.786	0.00666
gene2	-2.0106	0.3643	-5.519	4.09e-07
gene3	2.5455	0.6009	4.236	6.05e-05
gene4	1.7658	0.2694	6.554	5.03e-09

Residual standard error: 1.083 on 80 degrees of freedom  
 Multiple R-squared: 0.5571, Adjusted R-squared: 0.5405  
 F-statistic: 33.54 on 3 and 80 DF, p-value: 3.863e-14

(Fortsettes på side 4.)

**Oppgave 3.** En undersøkelse presentert i siste nummer av Universitas viser at 28% av studentene i Oslo og Akershus føler seg ensomme. I artikkelen får vi vite at undersøkelsen er basert på et representativt utvalg på 1016 studenter. Det står også i samme artikkel at en av fire hos den øvrige befolkningen føler seg ensomme.

a) Er studenter i Oslo og Akershus signifikant mer ensomme enn resten av befolkningen? Kall andelen ensomme studenter for  $p$ , og test  $H_0 : p = 0.25$  mot passende alternativ. Du kan teste ved å beregne en P-verdi. Konkluder og kommenter kort.

**Oppgave 4.**

La  $X$  være inntekten til en tilfeldig valgt lønnsmottaker i en bestemt befolkningsgruppe. Det er vanlig å anta at  $X$  følger en pareto-fordeling, dvs. at  $X$  har sannsynlighetstetthet

$$f(x) = \frac{\theta c^\theta}{x^{\theta+1}}$$

når  $c < x < \infty$ , og  $f(x) = 0$  ellers. Her er  $c$  minsteinntekten i befolkningsgruppen, mens  $\theta > 0$  er en parameter som beskriver lønnsforskjellene i befolkningsgruppen. I denne oppgaven forutsetter vi at minsteinntekten  $c$  er kjent, mens  $\theta$  skal estimeres fra uavhengige observasjoner  $X_1, X_2, \dots, X_n$  av inntekter i den aktuelle befolkningsgruppen.

Det kan vises at sannsynslighetsmaksimerings-estimatoren (maximum likelihood estimatoren) for  $\theta$  er

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \log X_i - n \log c}.$$

Det kan også vises at  $2n\theta/\hat{\theta}$  er  $\chi^2$ -fordelt med  $2n$  frihetsgrader.

Vi skal benytte opplysningene over til å konstruere en test for å undersøke om det er grunnlag for å påstå at  $\theta > \theta_0$ , der  $\theta_0$  er et gitt tall, dvs. teste hypotesen

$$H_0 : \theta = \theta_0 \quad \text{mot} \quad H_a : \theta > \theta_0.$$

a) Vis at med signifikansnivå  $\alpha$  vil testen bli 'Forkast  $H_0$  dersom  $\hat{\theta} \geq 2n\theta_0/\chi^2_{1-\alpha, 2n}$ '.

b) La  $\alpha = 0.01$  og  $n = 20$ . De 20 observerte inntektene gir  $\sum_{i=1}^{20} \log x_i = 235$ , mens minimumsinntekten i befolkningsgruppen er 100.000 kr. Gir dette grunnlag for å påstå at  $\theta > 2$ ?

c) Fremdeles for  $\alpha = 0.01$ ,  $n = 20$  og  $\theta_0 = 2$ , finn et uttrykk  $\beta(\theta')$  for sannsynligheten for type-II-feil når forskjellsparameteren i realiteten er  $\theta'$  (der  $\theta' > 2$ ). Hvor stor må  $\theta'$  være for at sannsynligheten for type-II-feil blir så liten som 0.05?

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: Tirsdag 11. desember 2012.

Tid for eksamen: 14.30 – 18.30.

Oppgavesettet er på 5 sider.

Vedlegg: Tabell over Poisson-, normal-,  $t$ -, og F-fordeling.

Tillatte hjelpeemidler: Godkjent kalkulator og formelsamling for STK1100/STK1110.

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

### Oppgave 1.

Lånekassen definerer en *fulltidsstudent* som en som sørger minst 30 studiepoeng per semester. Vi skal ved hjelp av en spørreundersøkelse finne ut om andelen fulltidsstudenter er forskjellig ved NTNU i Trondheim og ved UiO i Oslo. Kall andelen fulltidsstudenter ved UiO for  $p_1$  og andelen fulltidsstudenter ved NTNU for  $p_2$ . Vi spør et tilfeldig utvalg på størrelse  $n$  fra UiO og et tilfeldig utvalg på størrelse  $m$  fra NTNU om de oppfyller dette kravet inneværende semester.

- a) Utled en test på nivå  $\alpha = 0.05$  for å teste

$$H_0 : p_1 - p_2 = 0$$

mot en tosidig alternativhypotese. Anta at  $n$  og  $m$  er store nok til at du kan bruke tilnærming til normalfordeling.

- b) Konstruer et 95% konfidensintervall for forskjellen  $p_1 - p_2$ . Forklar hvorfor uttrykket for variansen blir forskjellig fra det du brukte i a).
- c) La  $n = 200$  og  $m = 400$ . 102 av de spurte studentene ved UiO og 248 av de spurte ved NTNU kvalifiserte som fulltidsstudenter. Sett inn både i a) og b) og kommenter.

### Oppgave 2.

Et advokatfirma i USA blir kontaktet av en gruppe kvinnelige ansatte i et elektronikkfirma, som mener at firmaet de jobber i gir kvinnene dårligere lønnsvilkår enn mennene. Advokatfirmaet plukker ut tilfeldige ansatte i elektronikkfirmaet, 10 kvinner og 9 menn, og gjennomgår deres lønnsopplysninger. Spesielt samler de tall for utbytte ved siste

(Fortsettes på side 2.)

personlige lønnsforhandling (i \\$) og en vurdering (0-100 poeng) for hvor tilfreds nærmeste overordnet er med medarbeiterens arbeid det siste året. Det sies fra firmaet at utbyttet ved lønnsforhandling blant annet baseres på denne vurderingen (variabelen 'score'). Data er lagt ved til slutt i oppgavesettet, sammen med en utskrift fra R og et plott. Lønnsøkning fra lønnsforhandlingene heter der 'salary\_incr'. Variabelen 'sex' er medarbeiterens kjønn, kodet som 1 for kvinner og 0 for menn.

Det er opplagt at kvinnene i utvalget i gjennomsnitt har fått langt lavere lønnstillegg enn mennene. Firmaet svarer advokatene at det må være fordi kvinnene generelt har fått lavere poengsum for arbeidsinnsatsen. Du skal først undersøke disse poengene for arbeidsinnsats (score). Gjennomsnittlig score i utvalget for kvinner var 43.5, og for menn var den 55.56. Vi oppgir også at estimerte varianser var 922.50 for score for kvinner og 552.78 for score for menn.

- a) Vi antar at variabelen score er normalfordelt både for kvinner og for menn, med forventning henholdsvis  $\mu_1$  og  $\mu_2$  og samme varians  $\sigma^2$ . Gjennomfør en t-test for å avgjøre om det er grunnlag i dataene for å påstå at kvinnene gjør en dårligere jobb (målt ved score fra overordnet) enn mennene. Bruk nivå  $\alpha = 0.05$ . Du må skrive opp hypotesene og uttrykket for testobservator, og spesifisere hvilken fordeling du har brukt. Skriv en konklusjon. Du kan støtte deg på utskriften fra R.
- b) I a) har vi antatt lik varians i fordelingene til score for kvinner og for menn. Er dette en rimelig antakelse? Svar ved hjelp av en hypotesetest. Sett opp hypoteser, testobservator og fordeling, og gjennomfør testen på nivå  $\alpha = 0.1$ . Hva blir din konklusjon? Du kan bruke informasjon fra R-utskriften.

Heretter skal vi betrakte score som en fast forklaringsvariabel (ikke stokastisk). For at advokatene skal kunne bruke tallmaterialet i saken mot elektronikkfirmaet, bestemmer firmaets statistikk-konsulent at man skal utføre en multipel regresjon, der utbytte i lønnsforhandling er responsen som forsøkes forklart av score og kjønn. La responsen være  $y$ , score  $x_1$  og kjønn  $x_2$ . Modellen vi bruker har med et interaksjonsledd, og ser slik ut:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} + \epsilon_i$$

der  $\epsilon_i$  er uavhengige og  $N(0, \sigma^2)$ ,  $i = 1, \dots, n$ .

- c) Spesifiser hvordan modellen over blir for hhv. kvinner og menn, og forklar hvordan parameterne kan brukes til å vurdere to typer lønnsdiskriminering: mulig forskjell i grunnleggende lønnstillegg, og mulig forskjell i uttelling for hvert poeng i arbeidsvurderingen (score).
- d) Du finner den tilpassede regresjonsmodellen, basert på de  $n = 19$  observasjonene, i utskriften fra R. Forklar på bakgrunn av de estimerte parameterverdiene og tilhørende P-verdier hvorfor kvinnene og advokatene deres har en god sak.

(Fortsettes på side 3.)

- e) Konstruer et 99% konfidensintervall for interaksjonsparameteren  $\beta_3$ . Du kan hente noen av tallene du trenger fra utskriften. Forklar hvordan et slikt konfidensintervall skal tolkes.
- f) Hvordan er 'Multiple R-squared' beregnet, og hvordan skal den tolkes? Hvordan vurderer du den tilpassede modellen i forhold til vedlagte plott av dataene? Hvilke andre plott ville du ha konstruert for å vurdere modellens egnethet?

### Oppgave 3.

Du skal i denne oppgaven studere en enkel lineær regresjonsmodell som vi tvinger til å gå gjennom origo, dvs. at skjæringspunktet med  $y$ -aksen er satt lik null. La responsen være  $y$  og forklaringsvariabelen  $x$ . Vi har  $n$  observasjonspar  $(x_i, y_i)$ . Modellen er da altså

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

der  $\epsilon_i$ -ene er uavhengige og normalfordelte med forventning 0 og varians  $\sigma^2$ .

- a) Finn minste kvadraters estimator (least squares estimator)  $\hat{\beta}_{MK}$  for  $\beta$ . Vis at denne er forventningsrett og finn et uttrykk for estimatorens varians.
- b) En alternativ estimator for  $\beta$  er  $\hat{\beta}_A = \sum_i^n y_i / \sum_i^n x_i$ . Vis at også denne er forventningsrett. Finn variansen til  $\hat{\beta}_A$  og sammenlign med variansen til  $\hat{\beta}_{MK}$ . Hvilken estimator vil du bruke? Begrunn svaret.

### Oppgave 4.

Antall tilfeller  $X$  av en sjeldent, medfødt sykdom i Norge per år kan antas å følge en Poisson-fordeling med parameter  $\lambda$ , dvs.  $X \sim \text{Poisson}(\lambda)$ , slik at  $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$  for  $k = 0, 1, 2, \dots$

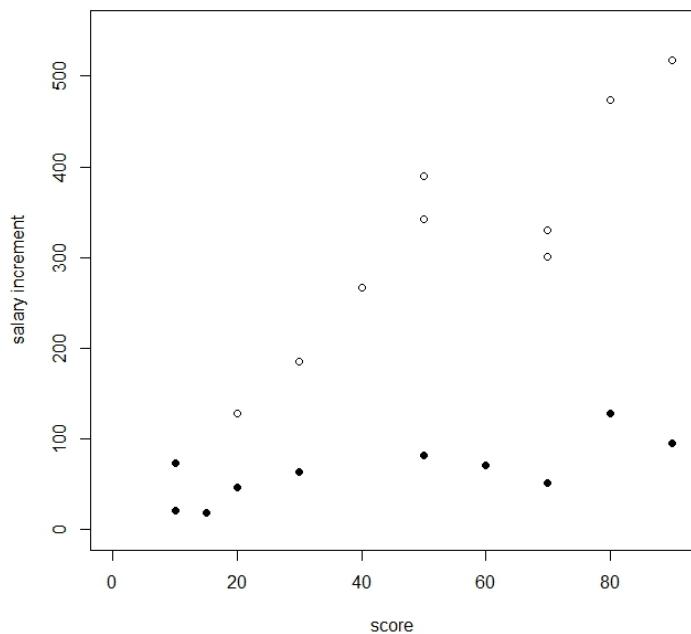
- a) Forventet antall barn født med denne sykdommen har vært ett per år. Hvor mange tilfeller må man minst observere et gitt år for å kunne forkaste  $H_0 : \lambda = 1$  til fordel for  $H_a : \lambda > 1$  på nivå  $\alpha = 0.05$ ?
- b) Hvor sannsynlig er det at man med testen ovenfor oppdager at  $\lambda$  faktisk er økt, dersom den i virkeligheten er doblet i forhold til tidligere (dvs.  $\lambda = 2$ )? Hva er sannsynligheten for type-II-feil i denne situasjonen?
- c) Forskere vil studere forekomsten av sykdommen nærmere. For å estimere sannsynligheten  $p$  for å få et barn med sykdommen, observerer man antall fødte barn  $n_i$  og antall tilfeller  $X_i$  for hvert år  $i$ , der  $i = 1, \dots, m$ . Vi antar at sannsynligheten  $p$  er konstant i disse  $m$  årene. Dessuten er  $n_i$ -ene store og  $p$  liten, slik at vi kan anta at  $X_i \sim \text{Poisson}(n_i p)$ ,  $i = 1, \dots, m$ . Finn sannsynlighetsmaksimerings-estimatoren (Maximum Likelihood Estimator) for  $p$  i denne situasjonen. Finn estimatorens forventning og varians.

(Fortsettes på side 4.)

## Datasett oppgave 1:

	sex	score	salary_incr
1	1	10	21
2	1	90	96
3	1	20	47
4	1	80	128
5	1	30	64
6	1	70	52
7	1	10	73
8	1	15	19
9	1	50	82
10	1	60	71
11	0	20	128
12	0	80	474
13	0	50	342
14	0	70	330
15	0	30	185
16	0	70	301
17	0	40	267
18	0	90	517
19	0	50	390

Plott av data, sort punkt kvinner, hvitt punkt menn:



(Fortsettes på side 5.)

Utskrift fra R, oppgave 1:

```
> t.test(score[1:10],score[11:19],alternative="less", var.equal=T)

Two Sample t-test

data: score[1:10] and score[11:19]
t = -0.959, df = 17, p-value = 0.1755
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf 9.812306
sample estimates:
mean of x mean of y
43.50000 55.55556

> var.test(score[1:10],score[11:19])

F test to compare two variances

data: score[1:10] and score[11:19]
F = 1.6688, num df = 9, denom df = 8, p-value = 0.4822
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.3830055 6.8455251
sample estimates:
ratio of variances
1.66884

> fit=lm(salary_incr~score+sex+score*sex)
> summary(fit)

Call:
lm(formula = salary_incr ~ score + sex + score * sex)

Residuals:
    Min      1Q  Median      3Q     Max 
-93.626 -21.684   0.266  29.609  90.394 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 62.0553    40.7630   1.522 0.148723  
score        4.7510     0.6815   6.972 4.49e-06 *** 
sex         -31.1230    48.3228  -0.644 0.529259  
score:sex   -3.9609     0.8437  -4.695 0.000288 *** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 45.32 on 15 degrees of freedom
Multiple R-squared:  0.9327,    Adjusted R-squared:  0.9192 
F-statistic: 69.29 on 3 and 15 DF,  p-value: 5.101e-09
```

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder og dataanalyse 1

Eksamensdag: Fredag 9. desember 2011

Tid for eksamen: 14.30–18.30

Oppgavesettet er på 4 sider.

Vedlegg: Tabell over normalfordeling,  
tabell over t-fordeling og  
tabell over  $\chi^2$ -fordeling

Tillatte hjelpeemidler: Formelsamling STK1100/STK1110 og  
godkjent kalkulator

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

### Oppgave 1

Utgangspunktet for denne oppgaven er en biologisk undersøkelse der vekten i gram til 9 fisk er bestemt sammen med fiskenes lengde i mm og alder i år. Hensikten er å undersøke sammenhengen mellom vekten på den ene siden og lengde og alder på den andre. Vekt er derfor respons og lengde og alder de uavhengige variablene. Nedenfor er utsnitt av en R-utskrift fra tilpasning av regresjonsmodellen med forventet respons

$$E(Y_i) = \beta_0 + \beta_1 \text{lengde}_i + \beta_2 \text{alder}_i, \quad i = 1, \dots, 9$$

Call:

```
lm(formula = vekt ~ lengde + alder, data = fisk)
```

Coefficients:

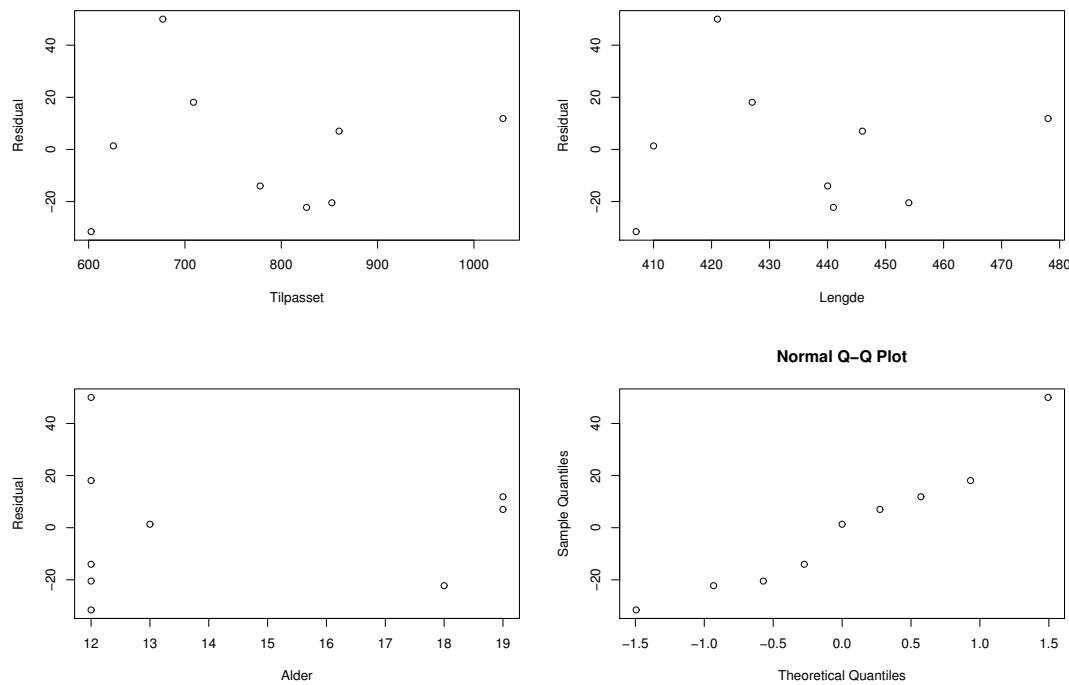
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1647.3967	224.8228	-7.328	0.000330 ***
lengde	5.3173	0.5872	9.055	0.000102 ***
alder	7.1498	4.0309	1.774	0.126468
<hr/>				
Signif. codes:	*** 0.001 ** 0.01 * 0.05 0.1 1			

Residual standard error: 29.23 on 6 degrees of freedom  
Multiple R-squared: 0.9663, Adjusted R-squared: 0.9551  
F-statistic: 86 on 2 and 6 DF, p-value: 3.83e-05

I figuren på neste side er det gjengitt fire plott, som kan brukes for å vurdere hvor god tilpasningen eller føydningen er.

I denne oppgaven antar vi at feileddene  $\epsilon_i$ ,  $i = 1, \dots, 9$  er uavhengige  $N(0, \sigma^2)$  fordelte.

(Fortsettes på side 2.)



Figur 1: Plott for å vurdere føyningen

- Forklar hvordan føyningen, eller tilpasningen, av modellen kan vurderes ut fra den vedlagte utskriften og plottene.
- Bruk resultatene i utskriften til å beregne et 95% konfidensintervall for koeffisienten til alder, det vil si  $\beta_2$ .
- Hva er konklusjonen av en ensidig test med nivå 0.05 av testen for nullhypotesen  $H_0 : \beta_2 = 0$  mot alternativet  $H_a : \beta_2 > 0$ ? Hva er P-verdien?

## Oppgave 2

I en Bernoulli forsøksrekke er en vanlig test å bruke den observerte andelen suksesser for å teste nullhypoteser om andelen  $p$ . Et alternativ er å ta utgangspunkt i antall forsøk som er nødvendige for å observere en suksess. Denne variabelen er geometrisk fordelt med punktsannsynlighet

$$P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \dots$$

- Begrunn at en rimelig test for nullhypotesen  $H_0 : p = \frac{1}{2}$  mot den alternative hypotesen  $H_a : p < \frac{1}{2}$  er å forkaste hvis antallet slike forsøk er stort.
- Bestem forkastningsområdet for testen i punkt a) når nivået er 10%.

(Fortsettes på side 3.)

- c) Forklar hva feil av type II er, og finn sannsynligheten for denne feilen i testen i punkt b) når  $p = 1/4$ . Begrunn hvorfor testen du utledet i punkt b) også har nivå 10% for å teste nullhypotesen  $H_0 : p \geq \frac{1}{2}$  mot den alternative hypotesen  $H_a : p < \frac{1}{2}$ .

### Oppgave 3

Nedenfor er et utsnitt av utskrift av tilpasning av en enkel lineær regresjon til samme data som i oppgave 1 med vekt som respons og lengde som uavhengig variabel, det vil si modellen med forventet respons

$$E(Y_i) = \beta_0 + \beta_1 \text{lengde}_i, \quad i = 1, \dots, 9.$$

Call:

```
lm(formula = vekt ~ lengde, data = fisk)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	-1828.0294	229.1108	-7.979	9.27e-05 ***		
lengde	5.9667	0.5249	11.368	9.13e-06 ***		
---						
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Residual standard error: 33.41 on 7 degrees of freedom  
 Multiple R-squared: 0.9486, Adjusted R-squared: 0.9413  
 F-statistic: 129.2 on 1 and 7 DF, p-value: 9.134e-06

Hvis  $y_1, \dots, y_9$  er vektene og  $x_1, \dots, x_9$  er lengdene til fiskene er  $\sum_{i=1}^9 y_i = 6961$ ,  $\sum_{i=1}^9 x_i = 3924$ ,  $\sum_{i=1}^9 y_i x_i = 3059173$  og  $\sum_{i=1}^9 x_i^2 = 1714916$ .

I denne oppgaven antar vi fortsatt at feleddene  $\epsilon_i$ ,  $i = 1, \dots, 9$  er uavhengige  $N(0, \sigma^2)$  fordelte.

- Hvorfor er minste kvadraters estimatorene for  $\beta_0$  og  $\beta_1$  i denne modellen også sannsynlighetsmaksimeringsestimatorene? Hva er sannsynlighetsmaksimeringestimatoren til  $\sigma^2$ ? Er denne forveningsrett?
- Betrakt nullhypotesen  $H_0 : \beta_1 = 5$  og den alternative hypotesen  $H_a : \beta_1 \neq 5$ . Hva er en rimelig testobservator i denne situasjonen? Forklar hva fordelingen under nullhypotesen er, og hvordan testen kan utføres ved hjelp av resultatene i utskriften ovenfor.
- Gjennomfør beregningene du forklarte i punkt b) ved hjelp av resultatene i utskriften når testen har nivå  $\alpha = 0.05$ . Bruk de vedlagte tabellene til å angi en øvre og nedre grense for P-verdien.
- Finn et 95% konfidensintervall for forventet vekt av fisker med lengde 50cm.
- Finn et 95% prediksjonsintervall for vekten til en fisk som har denne lengden, det vil si 50cm. Hva er fortolkningen av et slikt intervall?

(Fortsettes på side 4.)

- f) Forklar hva "Multiple R-squared: 0.9486" i utskriften ovenfor er og hva tolkningen er. Det er denne størrelsen som betegnes med  $r^2$  i læreboka. Hvis  $(x_1, y_1), \dots, (x_9, y_9)$  er de observerte vektene og lengdene, vis sammenhengen mellom  $r^2$  og den empiriske, eller utvalgs, korrelasjonskoeffisienten

$$\frac{\sum_{i=1}^9 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^9 (x_i - \bar{x})^2 \sum_{i=1}^9 (y_i - \bar{y})^2}}$$

der  $\bar{x} = \frac{1}{9} \sum_{i=1}^9 x_i$  og  $\bar{y} = \frac{1}{9} \sum_{i=1}^9 y_i$  ?

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: 10. desember 2010.

Tid for eksamen: 14.30 – 18.30.

Oppgavesettet er på 4 sider.

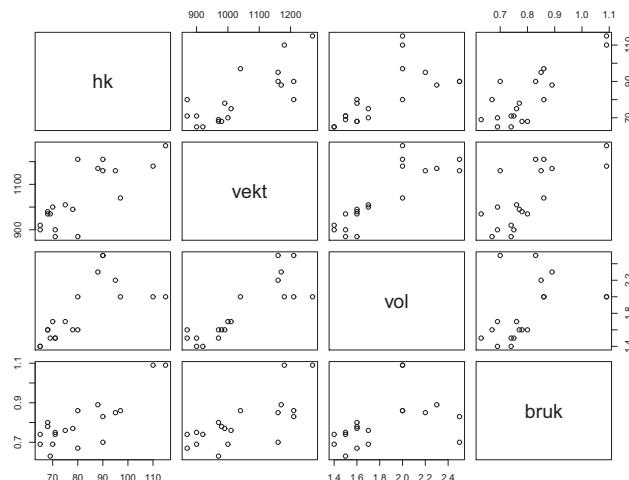
- Vedlegg:
- 1) Data over bensinforbruk for 19 personbiler.
  - 2) Tabell for standardnormalfordelingen.
  - 3) Tabell for  $t$ -fordelingene.
  - 4) Tabell for kji-kvadrat fordelingene.

Tillatte hjelpeemidler: Godkjent lommeregner.  
Formelsamling for STK1100 og STK1110.

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

### Oppgave 1

I denne oppgaven skal vi se nærmere på hvordan bensinforbruks til en bil avhenger av ulike tekniske data for bilen. I vedlegg 1 er det gitt tekniske data og informasjon om bensinforbruks (i liter per mil) for 19 ulike personbiler. (Dataene er tatt med for fullstendighetens skyld, og det er ikke nødvendig å se på dataene for å løse oppgaven.) De tekniske dataene er antall hestekrefter for motoren, vekten til bilen (i kg) og slagvolumet til motoren (i liter). I R-utskriftene står **hk** for antall hestekrefter, **vekt** for bilens vekt, **vol** for slagvolum og **bruk** for bensinforbruk. Plottet nedenfor gir en oversikt over dataene.



(Fortsettes på side 2.)

Resultatet av en multipel lineær regresjon med bensinforbruk som respons og antall hestekrefter, vekt og slagvolum som forklaringsvariabler er gitt nedenfor. Merk at R-utskriften er redigert (ved at noe av informasjonen er fjernet fra utskriften).

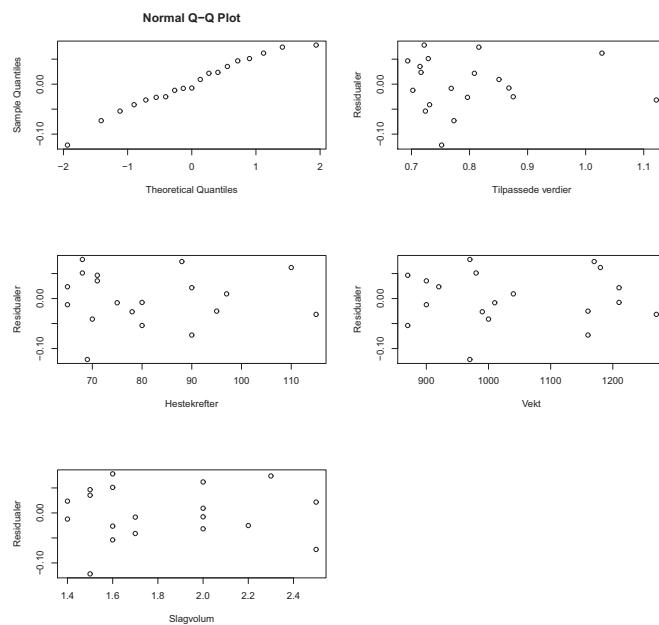
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.01033	0.11966	0.086	0.9323
hk	0.00604	0.00148	4.076	0.0010
vekt	0.00071	0.00023	3.005	0.0089
vol	-0.23979	0.07335		

Residual standard error: 0.0577

Multiple R-squared: 0.8212

- Forklar kort hva de ulike delene av R-utskriften forteller deg.
- Bruk en hypotesetest til å avgjøre om slagvolumet til motoren har signifikant betydning for bensinforbruket. Hvordan kan du forklare at den estimerte effekten av slagvolum er negativ selv om sammenhengen ser ut til å være positiv i plottet først i oppgaven?
- La  $\theta$  være forventet endring i bensinforbruket når styrken til motoren økes med 10 hestekrefter (og vekt og slagvolum ikke endres). Bestem et 95% konfidensintervall for  $\theta$  og forklar hva intervallet sier deg.

Nedenfor er det gitt ulike plott av residualene for den tilpassede modellen.



- Beskriv de forutsetningene regresjonsanalysen bygger på og bruk residualplottene til å vurdere om disse forutsetningene er rimelig godt oppfylt.

(Fortsettes på side 3.)

## Oppgave 2

Menneskene har hatt terninger i flere tusen år. De eldste terningene var skåret ut av et bein fra sauefoten, eller de var lagd av stein, metall eller keramikk. Vi regner i dag med at sannsynligheten er  $1/6$  for å få sekser når vi kaster en terning. Men var det tilfellet også for de terningene som ble brukt i tidligere tider?

På British Museum i London fins det flere terninger fra romertiden, blant annet en terning av marmor og en terning av jern. Det har blitt gjort forsøk med å kaste disse terningene. Da marmorterningen ble kastet 204 ganger, fikk en sekser i 54 av kastene.

- La  $p$  være sannsynligheten for at et kast med marmorterningen vil gi sekser. Bestem en test med signifikansnivå 5% for testing av nullhypotesen  $H_0 : p = \frac{1}{6}$  mot den alternative hypotesen  $H_a : p > \frac{1}{6}$ . Hva blir konklusjonen din?
- Forklar hva  $P$ -verdien til en test er. Bestem  $P$ -verdien for situasjonen i punkt a.

Terningen av jern har også blitt kastet 204 ganger. Da fikk en sekser i 42 av kastene.

- La nå  $p$  være sannsynligheten for at et kast med terningen av jern vil gi sekser. Test nullhypotesen  $H_0 : p = \frac{1}{6}$  mot den alternative hypotesen  $H_a : p > \frac{1}{6}$ . Bruk signifikansnivå 5%. Hva blir konklusjonen din?
- Forklar hva vi mener med feil av type II. Bestem sannsynligheten for feil av type II hvis sannsynligheten for å få sekser med terningen av jern er  $p = 0.20$ .
- Hvor mange kast må en gjøre med terningen av jern for at sannsynligheten for feil av type II skal være 20% hvis  $p = 0.20$ ?

## Oppgave 3

Vi ser på den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i; \quad i = 1, 2, \dots, n; \quad (1)$$

der  $x_i$ -ene er gitte størrelser,  $\epsilon_i$ -ene er uavhengige og  $N(0, \sigma^2)$ -fordelte og  $\beta_0$ ,  $\beta_1$  og  $\sigma^2$  er ukjente parametere. Merk at vi i (1) har trukket fra gjennomsnittet av  $x_i$ -ene. Vi sier at vi har sentrert forklaringsvariabelen.

- Vis at minste kvadraters estimatorer for  $\beta_0$  og  $\beta_1$  kan gis på formen

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{og} \quad \hat{\beta}_1 = \sum_{i=1}^n \frac{x_i - \bar{x}}{S_{xx}} Y_i$$

der  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

(Fortsettes på side 4.)

- b) Vis at  $\hat{\beta}_0$  og  $\hat{\beta}_1$  er forventningsrette.  
c) Bestem  $V(\hat{\beta}_0)$  og  $V(\hat{\beta}_1)$  og vis at  $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$ .

*Hint:* Hvis  $Z_1, \dots, Z_n$  er uavhengige stokastiske variabler og  $a_1, \dots, a_n$  og  $b_1, \dots, b_n$  er konstanter, så er  $\text{Cov}(\sum_{i=1}^n a_i Z_i, \sum_{i=1}^n b_i Z_i) = \sum_{i=1}^n a_i b_i V(Z_i)$ .

Vi er interessert i å estimere  $\mu_{Y|x^*} = \beta_0 + \beta_1(x^* - \bar{x})$ , dvs. forventet respons når  $x = x^*$ .

- d) Vis at  $\hat{\mu}_{Y|x^*} = \hat{\beta}_0 + \hat{\beta}_1(x^* - \bar{x})$  er en forventningsrett estimator for  $\mu_{Y|x^*}$  og bestem variansen til estimatoren.

En estimator for  $\sigma^2$  er

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \bar{x}))^2.$$

Det er kjent at  $S^2$  er uavhengig av  $\hat{\beta}_0$  og  $\hat{\beta}_1$  og at  $(n-2)S^2/\sigma^2$  er kji-kvadrat fordelt med  $n-2$  frihetsgrader. (Du skal ikke vise det.)

- e) Forklar at

$$\frac{\hat{\mu}_{Y|x^*} - \mu_{Y|x^*}}{S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

er  $t$ -fordelt med  $n-2$  frihetsgrader og bruk det til å bestemme et  $100(1-\alpha)\%$  konfidensintervall for  $\mu_{Y|x^*}$ .

SLUTT

## Vedlegg 1: Data for personbiler

Nedenfor er det gitt ulike tekniske data og data om bensinforbruk for 19 ulike merker av personbiler. Hver linje i tabellen gir informasjon for et bilmerke, og kolonnene i tabellen gir antall hestekrefter, vekt (i kg), slagvolumet til motoren (i liter) og bensinforbruks (i liter per mil).

hk	vekt	vol	bruk
68	980	1.6	0.78
95	1160	2.2	0.85
97	1040	2.0	0.86
75	1010	1.7	0.76
115	1270	2.0	1.09
88	1170	2.3	0.89
65	900	1.4	0.69
80	870	1.6	0.67
80	1210	2.0	0.86
71	900	1.5	0.75
68	970	1.6	0.80
90	1210	2.5	0.83
90	1160	2.5	0.70
70	1000	1.7	0.69
65	920	1.4	0.74
69	970	1.5	0.63
78	990	1.6	0.77
110	1180	2.0	1.09
71	870	1.5	0.74

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i STK1110 — Statistiske metoder og dataanalyse 1

Eksamensdag: Onsdag, 2. desember 2009

Tid for eksamen: 9.00 – 12.00

Oppgavesettet er på 4 sider.

Vedlegg: Tabell for  $\chi^2$ -fordeling, tabell for  $t$ -fordeling

Tillatte hjelpeemidler: Godkjent lommeregner og  
Formelsamling for STK1100 og STK1110

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

### Oppgave 1

En investor vil saksøke aksjemegler A fordi han mener avkastningen på porteføljen megleren forvalter for ham, er for lav. Vi har registrert månedlig avkastning (i prosent) for de 36 månedene megleren har tatt hånd om porteføljen. Vi ønsker å sammenligne med månedlig avkastning for de samme 36 månedene for en tilsvarende portefølje forvaltet av en annen megler B. Vi parrer observasjonene ( $\text{diff} = \text{megler A} - \text{megler B}$ ) hver måned, og antar at differansene i avkastning for de 36 månedene kan betraktes som uavhengige av hverandre. Gjennomsnittet av de 36 månedlige avkastningene for megler A var -1.10 prosent, tilsvarende for megler B var 0.95 prosent.

Empirisk standardavvik for de 36 differansene var 5.89 prosent. Et normalfordelingsplott av de 36 differansene viser ingen ekstreme verdier eller skjevheter i fordelingen.

- a) Anta at fordelingen til differansene har forventning  $\mu$  og standardavvik  $\sigma$ . Forklar hvorfor det er naturlig å teste hypotesene

$$H_0 : \mu = 0 \quad \text{mot} \quad H_a : \mu < 0.$$

Finn en passende testobservator, og spesifiser dennes fordeling når nullhypotesen er sann. Gjør rede for hvilke antakelser du legger til grunn.

- b) Beregn forkastningsområdet svarende til at testen skal ha nivå 0.01, og

(Fortsettes på side 2.)

konkluder på bakgrunn av tallene i oppgaveteksten. Beregn også tilhørende P-verdi (så godt du kan fra vedlagte tabell). Bør investoren gå til søksmål?

- c) Beregn et 95% konfidensintervall for standardavviket  $\sigma$ .
- d) Det ble også vurdert å bruke en to-utvalgstest istedenfor testen basert på de parrede observasjonene. Hvorfor ville to-utvalgstesten være et bedre alternativ under visse forutsetninger, og hvorfor tror du den parrede testen ble brukt i dette tilfellet?

## Oppgave 2

En ny antikoagulant (legemiddel mot blodpropp) er under utvikling. Som et ledd i utprøvingen fikk 12 friske menn og 8 friske kvinner antikoagulanten i ulike doser (i milligram), og prothrombin-tid ble målt for hver av dem (i sekunder). Prothrombin-tid er et mål som sier noe om hvor raskt blodet koagulerer. Man ønsker å bruke datamaterialet til å belyse sammenhengen mellom dose av antikoagulanten, kjønn og prothrombin-tid ved å utføre en regresjonsanalyse med dose  $x_1$  og kjønn  $x_2$  som forklaringsvariable og prothrombin-tid som respons  $Y$ .

I figuren på neste side er prothrombin-tid plottet mot dose, med forskjellige symboler for kvinner og menn. Regresjonsmodellen vi vil bruke er

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad (1)$$

der  $x_{i1}$  er dose for person  $i$  og  $x_{i2} = 0$  hvis person  $i$  er en mann og  $x_{i2} = 1$  hvis person  $i$  er en kvinne.  $\epsilon_i$  er uavhengige og normalfordelte  $N(0, \sigma^2)$ ,  $i = 1, \dots, 20$ .

- a) Gi en tolkning av parametrerne  $\beta_1$  og  $\beta_2$ . Virker modellen rimelig i forhold til plottet av dataene? Begrunn svaret.

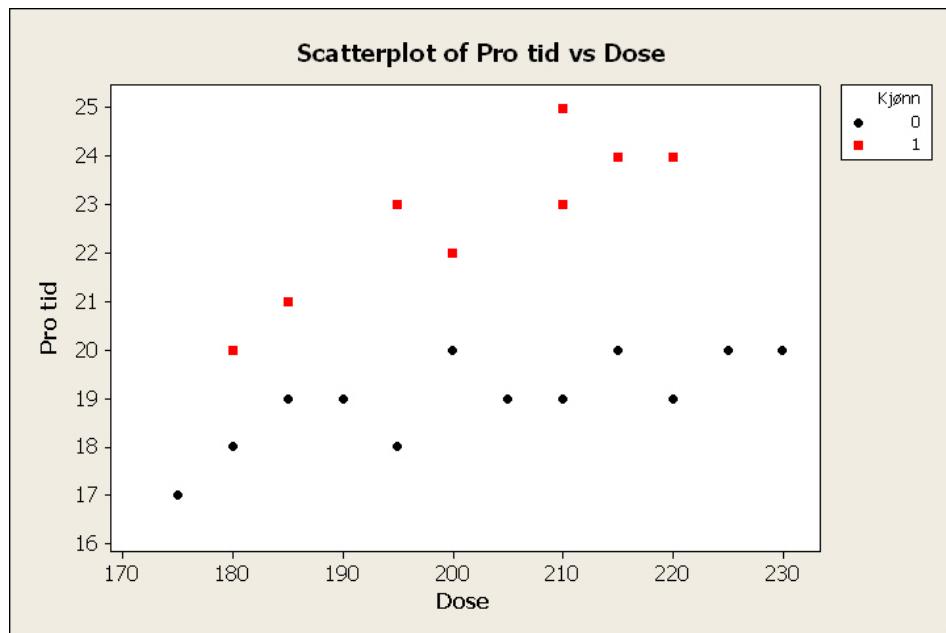
Vi skal uansett gå videre med modell (1) i punkt b) og c). En analyse av datasettet gir følgende estimatorer ved minste kvadraters metode:  $\hat{\beta}_0 = 7.155$ ,  $\hat{\beta}_1 = 0.05849$  og  $\hat{\beta}_2 = 3.7866$ . Estimert standardfeil er beregnet til hhv.  $s_{\hat{\beta}_0} = 2.445$ ,  $s_{\hat{\beta}_1} = 0.01201$  og  $s_{\hat{\beta}_2} = 0.3886$ .

- b) Beregn et 95% konfidensintervall for forventet forskjell i prothrombin-tid mellom kvinner og menn ved lik dose. Bruk intervallet til å teste om denne forskjellen er signifikant forskjellig fra 0. Hva blir signifikansnivået for denne testen?
- c) Hvordan beregnes residualene i denne modellen (du skal ikke beregne dem!)? Beskriv kort forskjellige plott som kaster lys over hvor godt modellen

(Fortsettes på side 3.)

passer til data. Multippel  $R^2$  er beregnet til 87.3 %. Hvordan skal dette tallet tolkes?

- d) Finn et uttrykk for endringen i forventet prothrombin-tid når dosen økes med  $d$  milligram. Vis at denne endringen er den samme for kvinner og menn. Hvordan vil du modifisere regresjonsmodellen dersom forventet endring i prothrombin-tid for en gitt endring i dose antas forskjellig for kvinner og menn? Se igjen på figuren og diskuter kort om dette er en rimelig utvidelse av modellen.



### Oppgave 3

I frykt for en pandemisk influensa planlegger Folkehelseinstituttet å massevaksinere befolkningen. Det er ønskelig at mer enn 60% av befolkningen vaksineres. På et gitt tidspunkt utføres en spørreundersøkelse for å finne ut hvor stor vaksinasjonsviljen er i befolkningen. Anta at  $p$  er andelen av befolkningen som sier ja til vaksinen. Et tilfeldig utvalg på  $n$  personer trekkes ut, og disse må svare på om de vil vaksineres eller ikke. La  $X$  være den tilfeldige variabelen som beskriver antallet som sier ja blant de  $n$ . Vi er interessert i å estimere  $p$ .

- Vis at den intuitive estimatoren for  $p$ ,  $\hat{p} = X/n$ , faktisk er sannsynlighetsmaksimerings-estimatoren (maximum likelihood estimator) for  $p$ . Er den også en momentestimator for  $p$ ?
- Finn forventning  $E(\hat{p})$  og varians  $V(\hat{p})$  for  $\hat{p}$  og vis at estimatoren er (Fortsettes på side 4.)

konsistent, altså at  $\hat{p}$  konvergerer i sannsynlighet mot  $p$  når  $n$  vokser. (Tips: Bruk Chebyshevs ulikhet fra formelsamlingen.)

- c) Anta at utvalgstørrelsen  $n$  er så stor at fordelingen til  $\hat{p}$  kan tilnærmes med en normalfordeling. Utled og gjennomfør en hypotesetest for å undersøke om innsamlede data viser at  $p$  er så stor som Folkehelseinstituttet ønsker. Du kan bruke at  $n = 200$  og at 122 av disse svarte ja.
- d) Finn sannsynlighetsmaksimerings-estimatorer for  $E(X)$  og  $V(X)$ . Vis at den resulterende estimatoren for  $V(X)$  ikke er forventningsrett, og foreslå en modifisert estimator som er det.

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdag: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: Onsdag 3. desember 2008.

Tid for eksamen: 9.00 – 12.00.

Oppgavesettet er på 4 sider.

Vedlegg: Tabell for normal,  $\chi^2$ - og  $t$ -fordelingen.

Tillatte hjelpeemidler: Godkjent lommeregner og Formelsamling for STK1100 og STK1110.

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

### Oppgave 1.

For å bestemme usikkerheten til et måleinstrument for hastighet måler man flere ganger hastigheten til en gjenstand som beveges med en kjent hastighet. Målingene nedenfor angir resultatet fra 5 slike forsøk der gjenstanden hadde en hastighet på 100 km per time

102.93, 99.27, 102.67, 99.82, 98.24.

En rimelig modell er her at målingene er realisasjoner av uavhengige normalfordelte variable,  $X_1, \dots, X_5$ , som alle er normalfordelt  $N(\mu_0, \sigma^2)$ , der  $\mu_0 = 100$  er kjent.

- Vis at fordelingen til  $\frac{1}{\sigma^2} \sum_{i=1}^5 (X_i - \mu_0)^2$  er  $\chi^2$ -fordelt med 5 frihetsgrader. Finn en forventningsrett estimator for  $\sigma^2$ . Hva er variansen?
- Bruk resultater fra punkt a) til å utlede en test med nivå  $\alpha$  for nullhypotesen

$$H_0 : \sigma^2 = \sigma_0^2 \text{ mot alternativet } H_A : \sigma^2 > \sigma_0^2$$

der  $\sigma_0^2$  er kjent.

(Fortsettes side 2.)

- c) Angi en øvre og nedre grense for p-verdien til testen basert på observasjonene i begynnelsen av oppgaven når  $\sigma_0^2 = 2$ . Forklar omhyggelig hva p-verdier er og hvordan resultatet du fant skal tolkes.

## Oppgave 2.

Følgende tabell viser et utsnitt av vektene,  $y_i$ , til 12 barn med en spesiell type spiseforstyrrelse samt deres høyde,  $x_i$ , og alder,  $z_i$ .

Table 1: Vekt i kg,  $y_i$ , høyden i meter,  $x_i$ , og alder i år,  $z_i$ , for 12 barn med spiseforstyrrelser.

Vekt, $y_i$	Høyde, $x_i$	Alder, $z_i$
23.9	1.44	8
26.5	1.50	10
19.8	1.24	6
:	:	:
28.3	1.55	12
25.4	1.45	9

Her er vekt respons, og høyde og alder er uavhengige variable. Betrakt først situasjonen der sammenhengen mellom vekt og høyde beskrives med en enkel lineær regresjonsmodell:

$$Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \dots, 12.$$

- a) Forklar hvilke forutsetninger en slik modell bygger på. Hvilke egenskaper til estimatorene for parametrene  $\beta_0$  og  $\beta_1$  kan utledes av det som i læreboka kalles ”standard statistisk modell”, det vil si uten å anta at feilreddene  $e_1, \dots, e_{12}$  er normalfordelte?
- b) Finn et estimat for  $\beta_1$  og bestem et 95% konfidensintervall for  $\beta_1$ . Her kan du bruke at  $\sum_{i=1}^{12} y_i = 280.869$ ,  $\sum_{i=1}^{12} x_i = 16.078$ ,  $\sum_{i=1}^{12} x_i^2 = 21.873$  og  $\sum_{i=1}^{12} x_i y_i = 381.526$ . Residual kvadratsum er  $RSS_1 = \sum_{i=1}^{12} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 41.645$ , der  $\hat{\beta}_0$  og  $\hat{\beta}_1$  er estimator for parametrene  $\beta_0$  og  $\beta_1$ .

En lineær regresjonsmodell uten konstantledd der høyde og alder er uavhengige variable kan formuleres som

$$Y_i = \alpha_1 x_i + \alpha_2 z_i + e_i, i = 1, \dots, 12.$$

(Fortsettes side 3.)

- c) Forklar hvordan modellen kan uttrykkes på matriseform,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{e},$$

der  $\mathbf{X}$  er en matrise av kjente tall og  $\mathbf{Y}$  og  $\mathbf{e}$  er vektorer av tilfeldige variable.

- d) Finn et estimat for  $\alpha_2$ . Her kan du i tillegg til de størrelsene som er oppgitt i punkt b), bruke at  $\sum_{i=1}^{12} z_i^2 = 976$ ,  $\sum_{i=1}^{12} x_i z_i = 144.247$  og  $\sum_{i=1}^{12} z_i y_i = 2534.908$ . Du kan også trenge

$$\begin{pmatrix} 21.873 & 144.247 \\ 144.247 & 976.0 \end{pmatrix}^{-1} = \begin{pmatrix} 1.804 & -0.267 \\ -0.267 & 0.040 \end{pmatrix}.$$

Residual kvadratsum er  $RSS_2 = \sum_{i=1}^{12} (y_i - \hat{\alpha}_1 x_i - \hat{\alpha}_2 z_i)^2 = 28.3$ , der  $\hat{\alpha}_1$  og  $\hat{\alpha}_2$  er estimatorer for koeffisientene  $\alpha_1$  og  $\alpha_2$ . Hva blir den estimerte standardfeilen til estimatoren for  $\alpha_2$ ?

### Oppgave 3.

Tabellen nedenfor viser antall ulykker per måned på en bestemt veistrekning i løpet av en 20-måneders periode. I 9 av månedene var det altså ingen ulykker, i 7 av månedene var det en ulykke osv.

Table 2: Antall ulykker per måned i en periode på 20 måneder.

Antall ulykker	Antall måneder
0	9
1	7
2	3
3	1

En rimelig modell i denne situasjonen er at antall ulykker antas å være realisasjoner av uavhengige Poisson fordelte variable  $X_1, \dots, X_{20}$  med samme punktsannsynlighet

$$P(X = x) = \frac{\lambda^x}{x!} \exp(-\lambda), \quad x = 0, 1, \dots$$

- a) Finn moment- og sannsynlighetsmaksimeringsestimatoren til  $\lambda$ . Beregn estimatene ved å bruke observasjonene fra tabellen i begynnelsen av oppgaven.

(Fortsettes side 4.)

- b) Utled et tilnærmet 95% konfidensintervall for  $\lambda$ . Beregn intervallet på grunnlag av observasjonene i tabellen.

Se nå på en situasjon der en bare bruker opplysninger om det har funnet sted ulykker i hver av de 20 månedene eller ikke. La  $Y_i = 1$  hvis  $X_i = 0$  og  $Y_i = 0$  ellers. Da er  $Y = \sum_{i=1}^{20} Y_i$  antallet ulykkesfrie måneder.

- c) Begrunn at antallet måneder der det ikke skjer ulykker, det vil si  $Y$ , er binomisk fordelt og angi punktsannsynligheten til  $Y$ . Bestem moment- og sannsynlighetsmaksimeringsestimatoren til  $\lambda$  i dette tilfellet. Hva blir estimatene?
- d) Utled og beregn et tilnærmet 95% konfidensintervall for  $\lambda$  på grunnlag av antall ulykkesfrie måneder.

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamens i: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: Onsdag 5. desember 2007.

Tid for eksamen: 9.00 – 12.00.

Oppgavesettet er på 4 sider.

Vedlegg: Tabell for  $t$ -fordelingen.

Tillatte hjelpeemidler: Godkjent lommeregner og Formelsamling for STK1100 og STK1110.

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

### Oppgave 1.

I USA er Thanksgiving Day, som feires den fjerde torsdagen i november, en nasjonal høytidsdag. På denne dagen spises det tradisjonelt kalkun som festmat. Tabell 1 gir vekten (i pund) for tilfeldig valgte Thanksgiving-kalkuner fra to delstater i USA.

Tabell 1: Slaktevekt for kalkuner fra to delstater i USA.

Delstat	Virginia	Wisconsin
Vekt (i pund)	13.1 12.4 13.2 11.8 13.8	11.5 14.2 15.4 13.1 13.8
Gjennomsnitt	12.625	13.600
Empirisk standardavvik	0.655	1.440

Vi lar  $\mu_1$  være forventet vekt for en tilfeldig valgt kalkun fra Virginia, mens  $\mu_2$  er forventet vekt for en tilfeldig valgt kalkun fra Wisconsin.

(Fortsettes side 2.)

- Bestem et 90% konfidensintervall for  $\mu_2 - \mu_1$  og diskuter hva intervallet sier deg. Beskriv hvilke forutsetninger konfidensintervallet bygger på.
- Angi en test med signifikansnivå 5% for nullhypotesen  $H_0 : \mu_1 = \mu_2$  mot den alternative hypotesen  $H_1 : \mu_1 \neq \mu_2$ . Hva blir konklusjonen av testen?
- Forklar hva vi mener med  $P$ -verdien til en test. Bestem (så godt du kan)  $P$ -verdien for testen i punkt b.

## Oppgave 2.

Thanksgiving-kalkunene i oppgave 1 var ikke like gamle da de ble slaktet. Tabell 2 viser kalkunenes alder (i uker) ved slakting. For oversiktens skyld gir vi også slaktevektene i tabellen.

Tabell 2: Slaktevekt og slaktealder for kalkuner.

Virginia		Wisconsin	
Alder	Vekt	Alder	Vekt
29	13.1	21	11.5
27	12.4	27	14.2
28	13.2	29	15.4
26	11.8	23	13.1
		25	13.8

- Plott vekten av kalkunene mot alderen ved slakting. Utfør plottingen i ett diagram, men benytt forskjellige symboler for de to delstatene. Kommenter plottet.

For å studere hvilken betydning delstat og alder har for slaktevekten til kalkunene, setter vi opp regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i. \quad (1)$$

I modellen (1) er:

- $Y_i$  slaktevekten til kalkun nummer  $i$  (fra begge delstatene sett under ett),
- $x_{i1}$  alderen ved slakting til kalkun nummer  $i$ ,
- $x_{i2} = 1$  hvis kalkun nummer  $i$  er fra Wisconsin,  $x_{i2} = 0$  hvis kalkun nummer  $i$  er fra Virginia.

(Fortsettes side 3.)

- b) Gi en forklaring av parametrene  $\beta_1$  og  $\beta_2$  i regresjonsmodellen (1), og beskriv hvilke forutsetninger regresjonsmodellen bygger på. Syns du modellformuleringen (1) virker rimelig i lys av plottet i punkt a?

Ut fra tallene i tabell 2 finner vi at minste kvadraters estimatorer er  $\hat{\beta}_0 = 0.311$ ,  $\hat{\beta}_1 = 0.448$  og  $\hat{\beta}_2 = 2.094$ , og at estimert standardfeil til  $\hat{\beta}_2$  er  $S_{\hat{\beta}_2} = 0.235$ . (Du skal ikke vise dette.)

- c) Angi en test med signifikansnivå 5% for nullhypotesen  $H_0 : \beta_2 = 0$  mot den alternative hypotesen  $H_1 : \beta_2 \neq 0$ . Hva blir konklusjonen av testen? Kommenter resultatet i lys av testen i punkt b i oppgave 1.

Ut fra tallene i tabell 2 har vi at residualkvadratsummen er

$$RSS = \sum_{i=1}^9 \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \right)^2 = 0.5648.$$

Vi innfører  $9 \times 3$  matrisen

$$\mathbf{X} = \begin{pmatrix} 1 & 29 & 0 \\ 1 & 27 & 0 \\ 1 & 28 & 0 \\ 1 & 26 & 0 \\ 1 & 21 & 1 \\ 1 & 27 & 1 \\ 1 & 29 & 1 \\ 1 & 23 & 1 \\ 1 & 25 & 1 \end{pmatrix}$$

Da er

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} 17.06 & -0.611 & -1.778 \\ -0.611 & 0.0222 & 0.0556 \\ -1.778 & 0.0556 & 0.589 \end{pmatrix}$$

- d) Finn den estimerte standardfeilen  $S_{\hat{\beta}_1}$  til  $\hat{\beta}_1$  og bestem et 95 % konfidensintervall for  $\beta_1$ . Forklar hva intervallet sier deg.

## Oppgave 3.

En bedrift produserer lysstøffrør. Levetiden (i timer) av et tilfeldig valgt lysstøffrør antas å være en stokastisk variabel med sannsynlighetstetthet

$$f(y | \theta) = \begin{cases} \frac{y}{\theta} e^{-y^2/2\theta} & \text{for } y > 0 \\ 0 & \text{ellers} \end{cases} \quad (2)$$

der  $\theta > 0$ .

(Fortsettes side 4.)

- a) La  $F(y | \theta)$  være den kumulative fordelingsfunksjonen som svarer til sannsynlighetstettheten (2). Medianen  $y_{0.50}$  er gitt ved at  $F(y_{0.50} | \theta) = 0.50$ . Vis at  $y_{0.50} = \sqrt{2\theta \log 2}$ . (Her står "log" for den naturlige logaritmen.)

Bedriften ønsker å studere levetiden til lysstoffrørene. For å gjøre det, observerer de levetidene  $Y_1, Y_2, \dots, Y_{20}$  for 20 tilfeldig valgte lysstoffrør. Vi antar at  $Y_1, Y_2, \dots, Y_{20}$  er uavhengige og identisk fordelte med sannsynlighetstetthet gitt ved (2).

- b) Vis at maksimum likelihood estimatoren  $\hat{\theta}$  er gitt ved  $\hat{\theta} = \frac{1}{40} \sum_{i=1}^{20} Y_i^2$ .
- c) Bedriften ønsker å teste nullhypotesen at median levetid er høyest 1000 timer mot alternativet at den er mer enn 1000 timer. Vis at det er det samme som å teste

$$H_0 : \theta \leq \theta_0 \quad \text{mot} \quad H_1 : \theta > \theta_0$$

der  $\theta_0 = 10^6 / (2 \log 2)$ .

- d) En rimelig test forkaster  $H_0$  hvis  $\hat{\theta} > k$ . Bestem  $k$  slik at testen får signifikansnivå 5%. (Vink: En kan vise at  $\frac{1}{\theta} \sum_{i=1}^{20} Y_i^2$  er kji-kvadratfordelt med 40 frihetsgrader. Du kan bruke dette resultatet uten å vise det. Utvalgte fraktiler i kji-kvadratfordelingen med 40 frihetsgrader er gitt i tabell 3 nedenfor.)
- e) Bestem (så godt du kan) styrken for testen i punkt d hvis median levetid er 1250 timer. Hva blir da sannsynligheten for feil av type II?

*Tabell 3: Fraktiler for kji-kvadratfordelingen  
med 40 frihetsgrader.*

$p$	0.025	0.05	0.10	0.20	0.30	0.40	0.50
$\chi_p^2$	24.43	26.51	29.05	32.34	34.87	37.13	39.33
$p$	0.60	0.70	0.80	0.90	0.95	0.975	
$\chi_p^2$	41.62	44.16	47.27	51.81	55.76	59.34	

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdag: Tirsdag 5. desember 2006.

Tid for eksamen: 15.30 – 18.30.

Oppgavesettet er på 3 sider.

Vedlegg: Tabell for normalfordeling, tabell for  $t$ -fordeling.

Tillatte hjelpeemidler: Godkjent lommeregner og Formelsamling for STK1100 og STK1110.

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

### Oppgave 1.

- a) Forklar hva som menes med at en estimator er forventningsrett og at den er konsistent.

I resten av oppgaven er  $X \sim \text{bin}(n, p)$  og  $\hat{p} = \frac{X}{n}$ .

- b) Forklar hvorfor  $\hat{p}$  er forventningsrett og konsistent.

- c) La  $\theta = p(1 - p)$ . En rimelig estimator for  $\theta$  er  $\hat{\theta} = \hat{p}(1 - \hat{p})$ . Er den forventningsrett? Hvis svaret er nei, foreslå en modifikasjon av  $\hat{\theta}$  som er det.

(Fortsettes side 2.)

## Oppgave 2.

Anta at  $X_1, X_2, \dots, X_n$  er uavhengige og identisk fordelte med tetthet

$$f(x|\sigma) = \frac{1}{2\sigma} \exp(-|x|/\sigma), \quad -\infty < x < \infty$$

- a) Bestem sannsynlighetsmaksimeringsestimatoren,  $\hat{\sigma}$ , for  $\sigma$ . Finn  $E(\hat{\sigma})$ .
- b) Vis at sannsynlighetsmaksimeringsestimatoren  $\hat{\sigma}$  er tilnærmet  $N(\sigma, \sigma^2/n)$  når antallet observasjoner,  $n$ , er stort.
- c) Forklar hvorfor både intervaller av formen

$$(\hat{\sigma} - z_{1-\alpha/2} \hat{\sigma}/\sqrt{n}, \hat{\sigma} + z_{1-\alpha/2} \hat{\sigma}/\sqrt{n})$$

og

$$\left( \frac{\hat{\sigma}}{(z_{1-\frac{\alpha}{2}}/\sqrt{n}) + 1}, \frac{\hat{\sigma}}{-(z_{1-\frac{\alpha}{2}}/\sqrt{n}) + 1} \right)$$

er konfidensintervaller med tilnærmet konfidenskoeffisient  $1 - \alpha$  når antallet observasjoner er stort. Her er  $z_{1-\frac{\alpha}{2}}$ ,  $1 - \alpha/2$  kvantilen i standardnormalfordelingen  $N(0, 1)$ .

## Oppgave 3.

Dataene nedenfor er et utsnitt av 17 årige målinger av prosentandel snøinnhold,  $x_i$ , 1. april og gjennomsnittlig vannføring om våren,  $y_i$ , målt i tommer i en elv i USA.

$i$	$x_i$	$y_i$
1	23.1	10.5
2	32.8	16.7
$\vdots$	$\vdots$	$\vdots$
16	21.1	10.5
17	27.6	16.1

Her er  $\sum_{i=1}^{17} x_i = 511.5$ ,  $\sum_{i=1}^{17} y_i = 267.1$ ,  $\sum_{i=1}^{17} x_i^2 = 16628.7$ ,  $\sum_{i=1}^{17} y_i^2 = 4549.43$   
og  $\sum_{i=1}^{17} x_i y_i = 8653.45$ .

(Fortsettes side 3.)

La vannføring være responsvariabel. Vi antar følgende lineære regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, 17$$

der  $\epsilon_1, \dots, \epsilon_{17}$  er uavhengige  $N(0, \sigma^2)$ -fordelte variable.

- a) Beregn minste kvadraters estimatene  $\hat{\beta}_0$  og  $\hat{\beta}_1$  for dataene ovenfor og skisser den tilpassede regresjonslinja.
- b) Du skal også tilpasse en modell uten konstantledd, dvs.

$$Y_i = \gamma x_i + \eta_i$$

der  $\eta_1, \dots, \eta_{17}$  er uavhengige  $N(0, \tau^2)$ -fordelte variable.

Finn minste kvadraters estimatoren  $\hat{\gamma}$ . Beregn estimatet for  $\gamma$  i dataene ovenfor og skisser også denne regresjonslinja i samme diagram som regresjonslinja fra punkt a).

- c) Forklar hvorfor estimatoren  $\hat{\beta}_0$  er normalfordelt. Hvordan kan nullhypotesen

$$H_0 : \beta_0 = 0 \quad \text{mot} \quad H_A : \beta_0 \neq 0$$

testes? Angi en øvre og nedre grense for  $p$ -verdien. Kommenter resultatet av testen i lys av resultatene du fant i punkt a) og b).

Her kan du bruke at kvadratsummen av residualene  $RSS = 45.56$  og uttrykket for  $\text{Var}(\hat{\beta}_0)$  fra formelsamlingen.

## Oppgave 4.

En logistisk fordelt variabel har kumulativ fordelingsfunksjon

$$F(x) = \frac{1}{1 + e^{-x}}, \quad -\infty < x < \infty.$$

Forklar hvordan man beregner et  $Q-Q$  plott for sammenligning av den kumulative fordelingsfunksjonen  $\Phi$  til en  $N(0, 1)$ -fordelt variabel og  $F$  gitt ovenfor. Angi verdiene som svarer til kvartilene og 0.1 og 0.9 kvantilene, dvs. verdiene for  $p = 0.1, 0.25, 0.5, 0.75$  og 0.9. Skisser plottet og kommenter utseendet.

SLUTT

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamensdato: STK1110 — Statistiske metoder  
og dataanalyse 1.

Eksamensdag: Mandag 5. desember 2005.

Tid for eksamen: 09.00 – 12.00.

Oppgavesettet er på 3 sider.

Vedlegg: Tabell for normalfordeling, tabell  
for  $t$ -fordeling.

Tillatte hjelpeemidler: Godkjent lommeregner og Formelsamling for STK1100 og STK1110.

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

### Oppgave 1.

- Forklar hvordan en  $t$ -fordeling er definert i læreboka.
- Anta at  $X_1, \dots, X_n$  er uavhengige og identisk fordelte tilfeldige variable som er  $N(\mu, \sigma^2)$ . Forklar hvorfor variabelen

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

er  $t$ -fordelt med  $n - 1$  frihetsgrader. Her er  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$   
og  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . Du kan her bruke kjente resultater om  
fordelingen til  $\bar{X}$  og  $S^2$ .

(Fortsettes side 2.)

## Oppgave 2.

La  $X_1, \dots, X_n$  være uavhengige og identisk fordelte variable som er Poisson fordelte med forventning  $\lambda$ .

- a) Begrunn at gjennomsnittet,  $\hat{\lambda} = \bar{X}$ , er estimatoren for  $\lambda$  som følger både fra moment- og sannsynlighetsmaksimeringsprinsippet.
- b) Finn  $E(\hat{\lambda})$  og  $\text{Var}(\hat{\lambda})$  og vis at  $\hat{\lambda}$  er konsistent.

På en bestemt veistrekning er de årlige tallene for antall ulykker med personskade i en femårsperiode

1 2 2 2 1 .

Anta i det følgende at dataene kan ses som realisasjon av uavhengige Poisson fordelte variable.

- c) Vi ønsker å teste

$$H_0 : \lambda = 1 \quad \text{mot} \quad H_A : \lambda > 1$$

og bruker en test av formen: Forkast  $H_0$  hvis  $\hat{\lambda} = \frac{1}{5} \sum_{i=1}^5 X_i > k$ .

Hvorfor er det en rimelig test?

Nedenfor finner du en tabell over den kumulative fordelingsfunksjonen til en Poisson fordelt variabel med forventning 5.

$x$	0	1	2	3	4	5	6	7	8
$P(X \leq x)$	0.0067	0.0404	0.1247	0.2650	0.4405	0.6160	0.7622	0.8666	0.9319

	9	10	11	12	13	14	15	16
	0.9682	0.9863	0.9945	0.9980	0.9993	0.9998	0.9999	1.0000

Forklar hvordan man bestemmer  $k$  slik at testen får nivå  $\alpha = 0.05$ .

(Hint: Husk at summen av uavhengige Poisson fordelte variable er Poisson fordelt.)

- d) Gir dataene ovenfor forkastning av  $H_0$ ? Hva er  $p$ -verdien til testen?
- e) Bruk normaltilnærmelsen fra sentralgrenseteoremet til å beregne den tilnærmede styrken for  $\lambda = 3$ .

(Fortsettes side 3.)

## Oppgave 3.

I to vulkanutbrudd tas det henholdsvis 12 og 10 målinger for å bestemme andelen hydrogen i gassen fra utbruddene. Målingene fra de to utbruddene er sammenfattet i følgende tabell.

	antall	gjennomsnitt	empirisk varians
Vulkan A:	$n = 12$	$\bar{x} = 47.5$	$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = (5.2)^2$
Vulkan B:	$m = 10$	$\bar{y} = 46.7$	$s_Y^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 = (6.7)^2$

Vi antar at målingene kan oppfattes som to uavhengige utvalg fra populasjoner med fordeling  $N(\mu_X, \sigma^2)$  og  $N(\mu_Y, \sigma^2)$  henholdsvis. Parametrene  $\mu_X, \mu_Y$  og  $\sigma^2$  er ukjente.

- a) Utled et 95% konfidensintervall for  $\mu_X - \mu_Y$ . Regn ut konfidensintervallet for dataene i begynnelsen av oppgaven. Forklar hvordan resultatet skal tolkes.  
Er det grunnlag for å forkaste hypotesen  $H_0 : \mu_X = \mu_Y$  mot  $H_A : \mu_X \neq \mu_Y$  når nivået er  $\alpha = 0.05$ ?
- b) Hva er den forventede lengden av konfidensintervallet? Hvis det totale antallet målinger,  $N = m + n$ , er gitt, hvilke valg av  $m$  og  $n$  gir det korteste konfidensintervallet?

## Oppgave 4.

La  $X_1, X_2, \dots, X_n$  være uavhengige og identisk fordelte tilfeldige variable med tetthet

$$f_X(x) = \begin{cases} \exp(-(x - \theta)) & x > \theta \\ 0 & \text{ellers} \end{cases}$$

der  $\theta$  er en ukjent parameter.

- a) Finn sannsynlighetsmaksimeringsestimatoren  $\hat{\theta}$ , til  $\theta$ .
- b) Hva er tettheten i fordelingen til estimatoren  $\hat{\theta}$ ?

SLUTT

# UNIVERSITETET I OSLO

## Matematisk Institutt

EKSAMEN I: **STK 1110 – Statistiske metoder og dataanalyse 1**

**Avsluttende eksamen**

TID FOR EKSAMEN: **Mandag 6. desember 2004, kl. 9:00–12:00**

HJELPEMIDLER: **«Formelsamling til STK 1100 og STK 1110», kalkulator**

Dette oppgavesettet utgjør den andre av kursets to eksamener. Det inneholder fire oppgaver og er på tre sider. Kursets første eksamen ble arrangert **14. oktober d.å.**

### Oppgave 1

VI SKAL UT PÅ TUR i en homogen Poisson-skog. Trærne (av den bestemte typen *Tsuga Canadensis*) er fordelt slik at (a) antall trær  $N(A)$  innenfor et område  $A$  med areal  $\text{ar}(A)$  (målt i kvadratmeter) er Poisson-fordelt med parameter  $\lambda \text{ar}(A)$ ; (b) antall trær  $N(A)$  og  $N(B)$  innenfor ikke-overlappende områder  $A$  og  $B$  er stokastisk uavhengige. Her er  $\lambda$  en ukjent parameter som altså svarer til den gjennomsnittlige tre-tettheten pr. kvadratmeter.

- Fra et vilkårlig utgangspunkt i skogen, la  $A$  være en sirkel med radius  $y$ . Hva er sjansen for at det skal være nøyaktig tre trær innenfor  $A$ ? Og hva er sjansen for at det ikke skal være noen trær der?
- La  $Y$  være distansen fra dette utgangspunktet til nærmeste tre. Vis at  $Y$  har sannsynlighetstetthet

$$f(y, \lambda) = e^{-\lambda\pi y^2} 2\lambda\pi y \quad \text{for } y > 0.$$

- I dette punktet kan du benytte deg av formlene

$$\int_0^\infty y^2 e^{-cy^2} dy = \frac{\sqrt{\pi}}{4c\sqrt{c}} \quad \text{og} \quad \int_0^\infty y^3 e^{-cy^2} dy = \frac{1}{2c^2},$$

som holder for  $c$  positiv. Vis at

$$\mathbb{E}Y = \frac{1}{\sqrt{\lambda}},$$

og finn dessuten variansen til  $Y$ .

- (d) I tillegg til  $Y$ , distansen fra utgangspunktet til det nærmeste tre, kan man observere  $Z$ , distansen fra utgangspunktet til det nest nærmeste tre. Finn en formel for sannsynlighetstettheten til  $Z$ .

## Oppgave 2

OG HVOR MANGE TRÆR (av denne bestemte typen) er det i skogen? Man velger via et kart  $n$  forskjellige startsteder i skogen, alle nøyaktig spesifiserte, og fra hver av disse måler man distansen til det nærmeste tre. Dette gir målingene  $Y_1, \dots, Y_n$ .

- (a) I det følgende skal du anta at  $Y_1, \dots, Y_n$  er uavhengige med den samme sannsynlighetstetthet, nemlig den  $f(y, \lambda)$  du fant i Oppgave 1(b). Diskuter kort om disse antagelsene synes rimelige eller ikke.
- (b) Gjennomsnittet  $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$  vil konvergere i sannsynlighet når  $n$  vokser. Mot hvilken grense? Begrunn svaret.
- (c) Finn momentestimatoren for  $\lambda$ , og kall den  $\lambda^*$ . Vis at den er konsistent, altså at  $\lambda^*$  konvergerer i sannsynlighet mot  $\lambda$  når  $n$  vokser.
- (d) Vis så at ML-estimatoren (maximum likelihood-estimatoren, eller rimelighetsfunksjonsmaksimeringsestimatoren) blir

$$\hat{\lambda} = \frac{1}{\pi} \frac{1}{W_n}, \quad \text{der } W_n = \frac{1}{n} \sum_{i=1}^n Y_i^2.$$

- (e) Man foretok  $n = 25$  slike målinger, og fant disse tallene, i meter:

9.55	4.92	7.85	1.27	2.91	3.77	2.94	3.39
4.29	4.14	6.13	0.16	4.94	5.59	2.53	9.16
6.71	4.42	2.64	7.66	0.85	2.21	1.38	0.94
							2.70

For disse målingene finner man (blant annet)

$$\frac{1}{n} \sum_{i=1}^n Y_i = 4.122, \quad \frac{1}{n} \sum_{i=1}^n Y_i^2 = 23.458, \quad \text{empirisk standardavvik } 2.596.$$

Din oppgave nå er å lage et konfidensintervall for  $\lambda$  med konfidensgrad tilnærmet 95%.

## Oppgave 3

EN BIOLOG UNDERSØKER 300 KANINER, og klassifiserer hver av dem i bestemte kategorier 1, 2, 3. Hun har en genetisk-biologisk hypotese som etter noe sannsynlighetsberegning kan formuleres som at sannsynlighetene for at kaninene skal være av type 1, type 2, type 3 kan uttrykkes som

$$p_1 = \theta^2, \quad p_2 = 2\theta(1 - \theta), \quad p_3 = (1 - \theta)^2,$$

for en passende parameter  $\theta$ . Av hennes 300 kaniner viser  $N_1 = 10$  seg å være av type 1,  $N_2 = 80$  av type 2, og  $N_3 = 210$  av type 3. Hun vil gjerne teste sin teori.

- (a) Forklar kort de forutsetninger som skal til for at likelihoodfunksjonen for data, under hennes modell, blir

$$L(\theta) = \frac{300!}{10! 80! 210!} p_1(\theta)^{10} p_2(\theta)^{80} p_3(\theta)^{210}.$$

Anta i det følgende at disse betingelser er oppfylt.

- (b) Finn et estimat for  $\theta$ .  
(c) Bruk data til å teste om modellen hennes holder.

#### Oppgave 4

ER MIN FORVENTNING NULL, eller er den ikke? Vi skal se på den ganske enkle modellen der  $Y_1, \dots, Y_{100}$  er uavhengige observasjoner fra den samme normalfordelingen  $N(\theta, 1)$ , og der man altså vil teste  $H_0: \theta = 0$  mot alternativet at  $\theta \neq 0$ .

- (a) Sett opp likelihoodfunksjonen  $L_{100}(\theta)$  for modellen, og vis at ML-estimatoren under a priori-omstendigheter er  $\hat{\theta} = \bar{Y}_{100}$ , gjennomsnittet av observasjonene.  
(b) Sett opp et generelt uttrykk for  $\text{GLR}_{100}$ , den generaliserte likelihood-ratio-testen for  $H_0$ , og finn en så enkel formel for

$$Z_{100} = -2 \log \text{GLR}_{100},$$

i denne situasjonen, som mulig.

- (c) Hva sier den generelle teorien fra pensum om fordelingen til  $Z_n = -2 \log \text{GLR}_n$ , når  $n$  vokser? Hvordan fungerer dette approksimasjonsresultatet for situasjonen i denne oppgaven?  
(d) Anta denne testen basert på  $Z_{100}$  blir brukt til å teste  $H_0: \theta = 0$ , med forkastningsgrense svarende til at testens nivå blir 0.05. Forklar hvordan man kan beregne sannsynligheten for at testen faktisk forkaster  $H_0$ , når den virkelige  $\theta$  er lik 0.33.