UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen 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 hjelpemidler:	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.

α	0.100	0.050	0.025	0.010	0.001	0.0001	0.00001	0.000001
z_α	1.281	1.645	1.960	2.326	3.090	3.719	4.265	4.753

Oppgave 1

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

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

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

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

(b) Forklar prinsippet bak maksimum likelihood (ML) estimering og vis at ML estimatoren for θ er

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

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

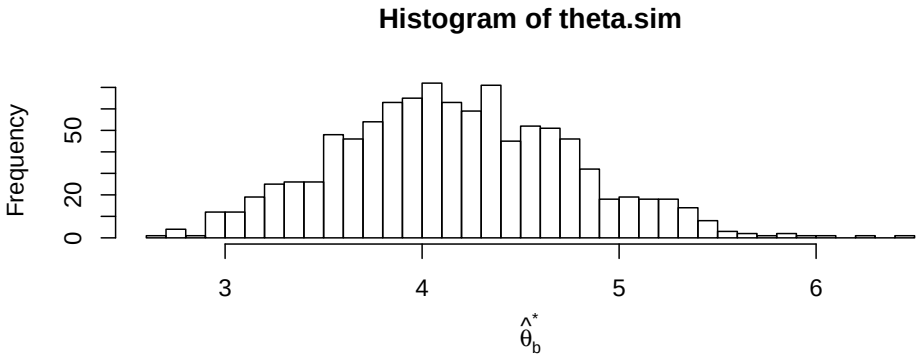
Er $\hat{\theta}$ forventningsrett?

(d) Regn ut $\hat{\theta}$ for de gitte data. Hva blir standardfeilen for $\hat{\theta}$ og estimatet 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 θ. 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:

```
        lpsa     lcavol  lweight age
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

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

der x_{i1} er `lcavol` mens y_i er `lpsa`. Her følger ε_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
lcavol       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 β_0 og β_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

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

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

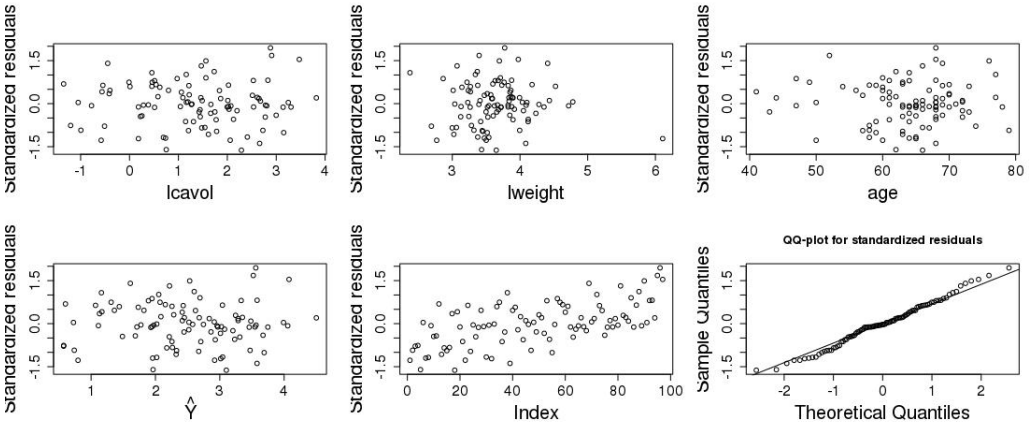
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.146941   0.772372   0.190  0.84953
lcavol       0.687819   0.067418  10.202  < 2e-16
lweight      0.549937   0.163838   3.357  0.00114
age         -0.009486   0.011003  -0.862  0.39081

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 $\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ der $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Vis at $\boldsymbol{E} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$ (der $\boldsymbol{I}$ er identitetsmatrisen mens du selv må finne ut hva $\boldsymbol{H}$ 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 $\boldsymbol{H}$.

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 sykkelulykker (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 sykkelulykker 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

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