UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	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 hjelpemidler:	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:

	Kvantiler				
Fordeling	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

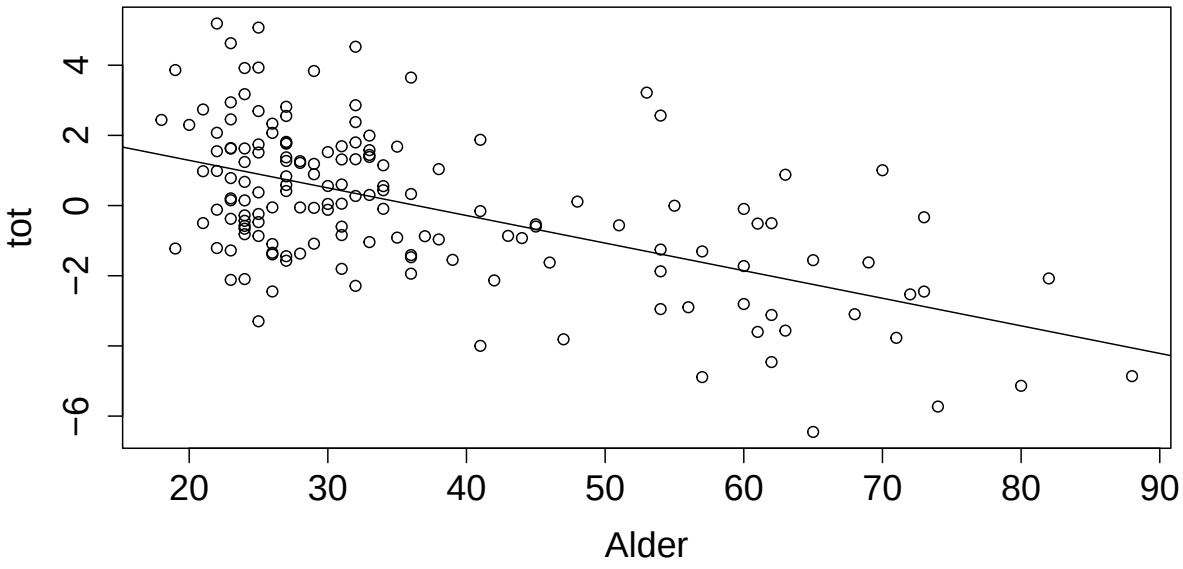
Oppgave 1

Plottet nedenfor viser data fra et studie på funksjonalitet av nyrer på 157 friske frivillige individer. Individenes alder er gitt på x-aksen mens y-aksen gir et sammensatt mål `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:

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

der $\hat{\beta}_0, \hat{\beta}_1$ er beregnet utifra minste kvadraters prinsippet. Her er $x =$ Alder og $Y =$ 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 β_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 tilsynelatende hverken `Alder` eller `Alder`^{2} 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 prediksjonsestimater 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

$X \sim$Gamma(α, β);
$Y \sim$Gamma$(3\alpha, \beta)$.

Anta α er kjent, mens β er ukjent.

(a) Sett opp likelihoodfunksjonen for β basert på observasjonenen X og Y og vis at log-likelihood funksjone (der noen ledd som ikke avhenger av β 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 β, $\hat{\beta}_{ML}$.

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

(Fortsettes på side 4.)

(c) Hvis α 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 skillsmissehyppighet 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?

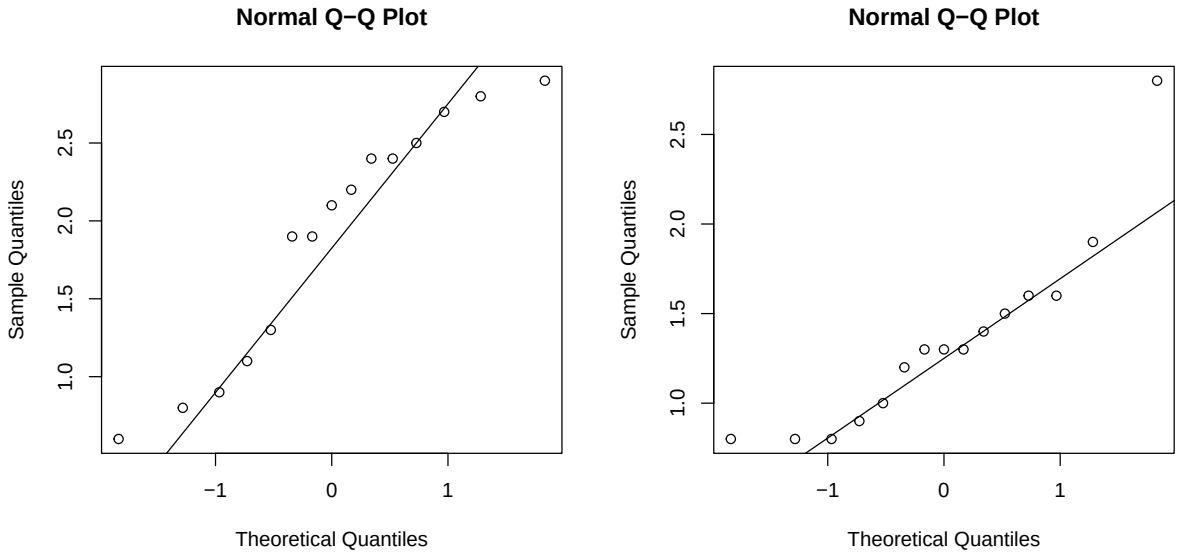
I forhold til konfidensintervallene du fikk i (*b*), virker det rimelig at P-verdien ligger i intervallet $[0.01, 0.05]$?

(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