UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	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 hjelpemidler:	Alle hjelpemidler 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 ν frihetsgrader for noen verdier av α og ν. Du vil få bruk for tabellen i Oppgave 1 og 3.

α :	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 $X_1, \ldots, X_{28} \overset{uif}{\sim} N(\mu, \sigma^2)$.

a

Utled et 99% konfidensintervall for forventet gjennomsnittlig vekt μ 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 størrelsesorden for P-verdien.

Oppgave 2

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

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

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

a

Vis at momentestimatoren for θ 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 $N(\theta, \sigma^2_{\hat{\theta}})$-fordelt og finn et uttrykk for $\sigma^2_{\hat{\theta}}$.

Oppgave 3

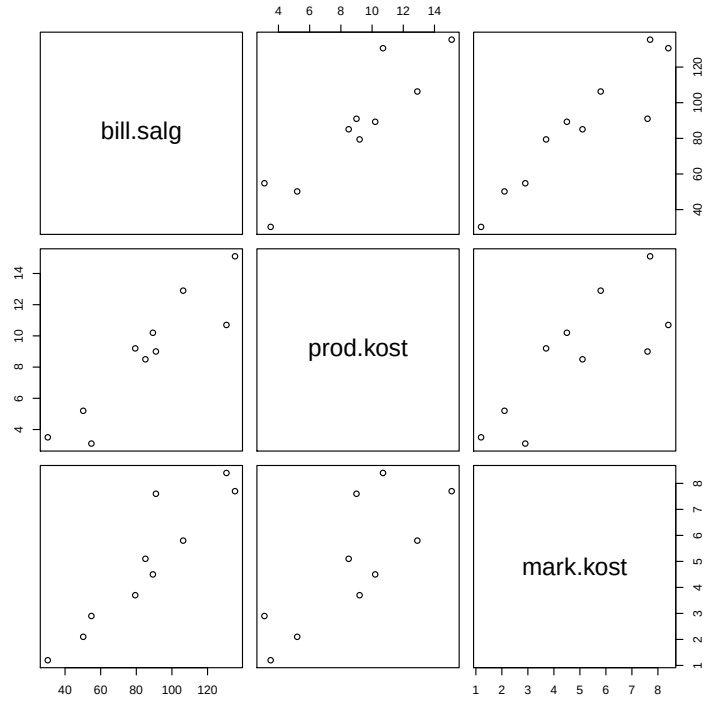
En ønsker å finne ut hvordan billettsalget for en kinofilm avhenger av produksjons- og markedsføringskostnadene. Dataene, som er basert på 10 Hollywood-filmer, består av

- 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:

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

der $\bar{x}_1 = 1/n \sum_{i=1}^{n} x_{i1}$ og vi antar at $\epsilon_1, \ldots, \epsilon_{10} \overset{uif}{\sim} N(0, \sigma^2)$. 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.)

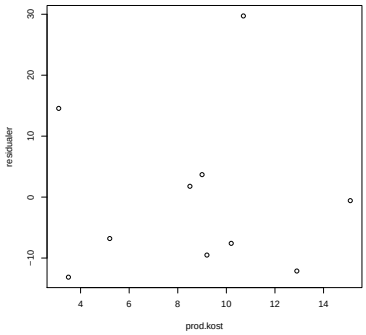
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     85.240      4.509  18.906 6.34e-08 ***
I(prod.kost - mean(prod.kost))   7.978      1.223   6.522 0.000184 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

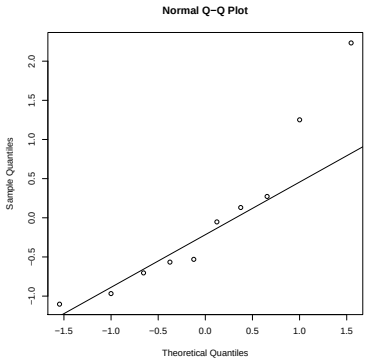
a

Gi en fortolkning av estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$. Lag så et 95% konfidensintervall for β_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?

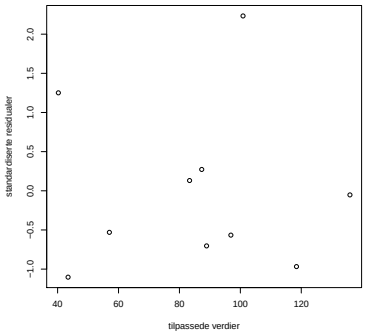
Residualer mot forklaringsvariabel



Normalfordelingsplott av standardiserte residualer



Standardiserte residualer mot tilpassede verdier



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

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

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 $\hat{Y} \sim N(\mu_{Y|x^*}, \sigma^2_{\hat{Y}})$. Vis at $\mu_{Y|x^*} = \mathrm{E}(Y|x^*)$ og $\sigma^2_{\hat{Y}} = \mathrm{Var}(\hat{Y}) = \sigma^2 \left(\frac{1}{10} + \frac{(x_1^* - \bar{x}_1)^2}{S_{xx}} \right)$ (du kan ta for gitt at $\hat{\beta}_0$ og $\hat{\beta}_1$ er forventningsrette for β_0 og β_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 σ^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 stole på resultatene du har fått? Begrunn svaret ditt.

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

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2 x_{i2} + \epsilon_i,\ i = 1, \ldots, 10,\ \epsilon_1, \ldots, \epsilon_{10} \overset{uif}{\sim} N(0, \sigma^2).$$

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 fortolkning 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?
Formulér 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