Løsningsforslag til Eksamen i M4STI1 2015F

Opgave 1

a. Hvis vi antager at stikprøverne er repræsentative for hele produktionen, så er 1.5/25*100% = **6%** af de producerede bremseskiver er defekte.

b. Antal defekte i stikprøverne opfattes som en binomialfordelt stokastisk variabel Y, hvor n = 25 og p = 0.06, så

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

Derfor er $P(Y=0) = \binom{25}{0}(0.06)^0(1-0.06)^{25-0} = 1 \cdot 1 \cdot (0.94)^{25} = \mathbf{0.213}$

Dette kan også beregnes i MatLab: binopdf(0, 25, 0.06) = 0.2129

c. $P(Y \geq 2) = 1 - P(Y < 2) = 1 - P(Y=0) - P(Y=1)$

$P(Y=1) = \binom{25}{1}(0.06)^1(1-0.06)^{25-1} = 25(0.06)(0.94)^{24} = 0.340$

Dermed $P(Y \geq 2) = 1 - 0.213 - 0.340 = \mathbf{0.447}$

d. Middelværdi: $\mu = np = 25 \cdot 0.06 = \mathbf{1.5}$

Varians: $\sigma^2 = np(1-p) = 25 \cdot 0.06 \cdot 0.94 = \mathbf{1.41}$

Spredning: $\sigma = \sqrt{\sigma^2} = \sqrt{1.41} = \mathbf{1.19}$

Opgave 2

a. Trin 1: Vælg hypoteser

H_0: $= \mu_0 = 21.8$

H_a: $\mu \neq \mu_0$

Trin 2: Formuler test statistikken:

Vi kender ikke populationsvariansen, så vi bruger

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

som er t-fordelt med n-1 frihedsgrader

Trin 3: Formuler den kritiske region:

Vi afviser H_0, hvis

$$|t| > t_0 = t_{n-1,\alpha/2}$$

Værdien $t_{n-1,\alpha/2}$ kan findes med MatLab som $-\mathrm{tinv}(\alpha/2\,, n-1) = 2.0639$

Trin 4: Foretag forsøget og beregn t

Datasættet for stikprøven indlæses og beregnes med MatLab. Funktion til indlæsning af kolonne A i Excel regnearket 'M4STI_2015_data.xlsx':

```
xlsread('M4STI_2015_data.xlsx','A:A')
```

Mellemresultater:

Stikprøvemiddelværdi: $\bar{y} = 21.7120$

Stikprøvevarians: $s^2 = 0.0436$

Stikprøvespredning: $s = 0.2088$

Teststatistik: $t = \frac{21.712 - 21.8}{0.2088\sqrt{25}} = -2.1072$

Trin 5: Konklusion

Teststatistikken $|t| = 2.1072$ er større end den kritiske grænse, $t_0 = 2.0639$, så vi kan forkaste nulhypotesen på baggrund af stikprøven. Bremseskivernes tykkelse er altså ikke 21.8 mm som ønsket.

b. 95% konfidensintervallet for middelværdien af bremseskivernes tykkelse er

$$\bar{y} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 21.7120 \pm 2.0639\frac{0.2088}{\sqrt{25}} = 21.7120 \pm 0.0862 = (\mathbf{21.63},\ \mathbf{21.798})$$

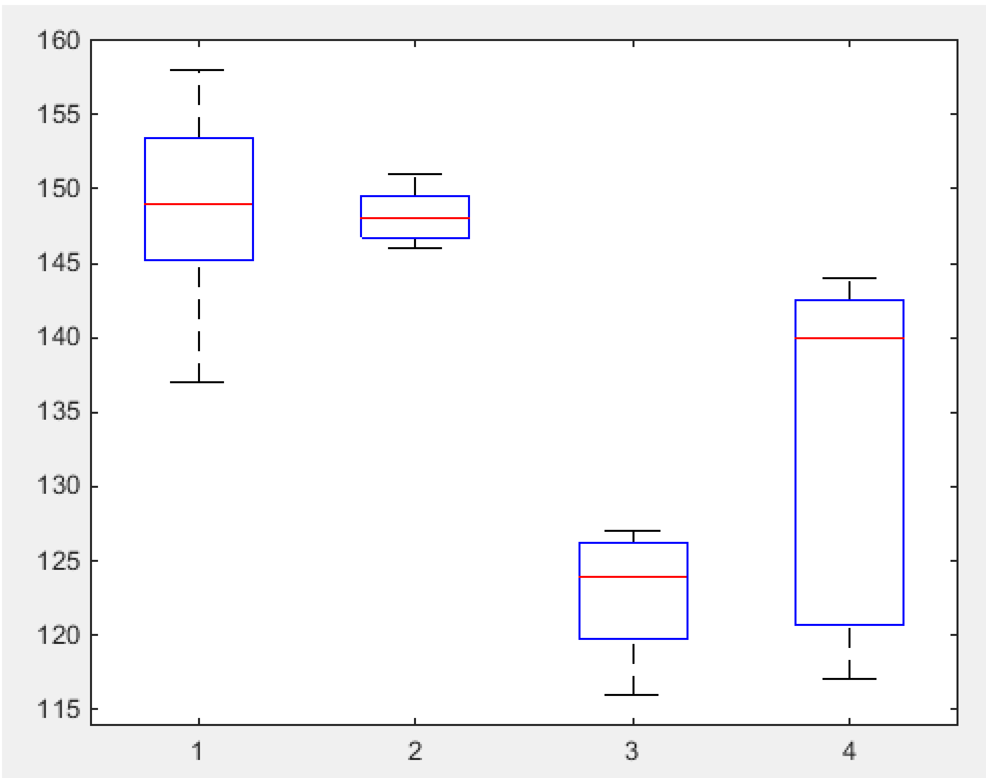
Vi ser at den ønskede middelværdi 21.8 er udenfor konfidensintervallet, men kun lige netop.

c. 95% prediktionsintervallet er

$$\bar{y} \pm t_{n-1,\alpha/2}s\sqrt{1+\frac{1}{n}} = 21.7120 \pm 2.0639 \cdot 0.2088\sqrt{1+\frac{1}{25}} = 21.7120 \pm 0.4395$$
$$= (\mathbf{21.27},\ \mathbf{22.15})$$

Opgave 3

a. Boxplot laves i MatLab med funktionen boxplot(Styrke, Legering), hvor Legering angiver grupperingen. Nedenfor vises boxplots over styrken for de fire legeringer. Vi ser, at legering 1 og 2 har omtrent samme styrke, som er højere end legering 3 og 4. 4 har højere styrke end 3, men også større varians. Legering 1 har desuden større varians end legering 2.



b. ANOVA er beregnet i MatLab med funktionen
`anova1(Styrke, Legering).`
Resultatet ses nedenfor:

Source	SS	df	MS	F	Prob>F
Groups	2382.8	3	794.267	13.29	0.0001
Error	956.4	16	59.775		
Total	3339.2	19			

Der er kraftig evidens for, at legeringerne ikke har samme styrke. P-værdien er 0.0001, langt under de 0.05, som var det valgte signifikansniveau.

c. Den parvise sammenligning vises nedenfor. Den er beregnet med MatLab funktionen
`multcompare(stats,'Alpha',0.05,'CType','lsd')`

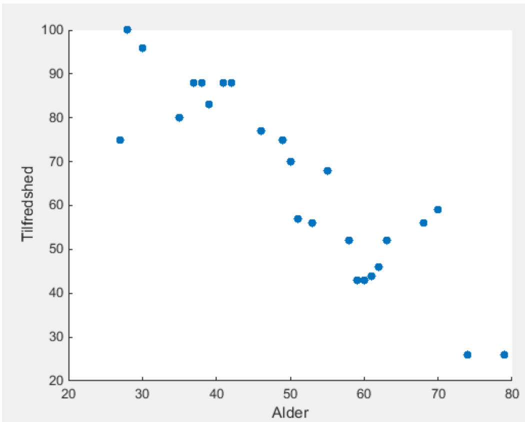
Sammenligningen viser f.eks. (første række i tabellen), at forskellen i gennemsnitlig styrke for observationerne af legering 1 og 2 er 0.6. Konfidensintervallet er (-9.77, 10.97), så det er altså muligt, at der ingen forskel er, da 0 tilhører intervallet. P-værdien på 0.9039 viser da også, at det ikke kan afvises. Til gengæld er legering 1 signifikant forskellig fra 3 og 4, legering 2 er forskellig fra 3 og 4, men legering 3 og 4 er ikke forskellige på signifikantniveau 0.05 (P-værdien er dog kun en anelse over, nemlig 0.0533).

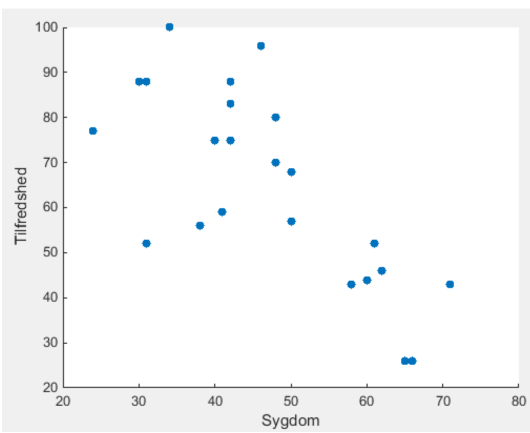
i	j	KI lav	Forskel	KI høj	P-værdi
1	2	-9.77	0.6	10.97	0.9039
1	3	15.63	26	36.37	0.0001
1	4	5.43	15.8	26.17	0.0052
2	3	15.03	25.4	35.77	0.0001
2	4	4.83	15.2	25.57	0.0068
3	4	-20.57	-10.2	0.17	0.0533

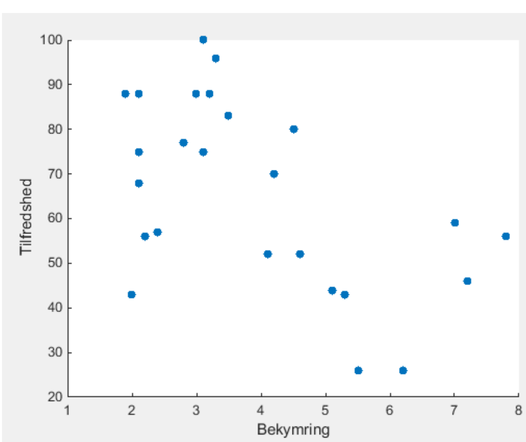
d. Legering 1 og 2 har højeste styrke, på næsten samme niveau; 1 har 148.8 og 2 har 148.2. Som boxplottet i a. viser har observationerne for legering 2 dog væsentligt mindre variation. Derfor vil jeg anbefale legering 2 som en ensartet legering med høj styrke.

Opgave 4

a. Nedenstående scatterplots er lavet i MatLab, f.eks.
`scatter(x(:,1),y,'filled')`







Alle tre variable lader til at have en negativ korrelation med Tilfredshed. Der er dog en del variation på data. Sammenhængen er mest tydelig for Alder og mindst for Bekymring.

b. Funktionen `mdl = fitlm(x, y)` i MatLab giver dette resultat:

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate       SE         tStat        pValue
                   ________    _______      _______     __________

    (Intercept)      144.58     5.9902       24.137     8.4778e-17
    x1              -1.1267    0.13655      -8.2515     4.9959e-08
    x2             -0.58727    0.13246      -4.4335     0.00023049
    x3               1.3448     1.0668       1.2605         0.2213


Number of observations: 25, Error degrees of freedom: 21
Root Mean Squared Error: 7.07
R-squared: 0.901,  Adjusted R-Squared 0.887
F-statistic vs. constant model: 64, p-value = 9.82e-11
```

Det giver regressionsligningen:

Tilfredshed = 144.58 - 1.1267*Alder - 0.58727*Sygdom + 1.3448*Bekymring

c. Tilfredshed = 144.58 - 1.1267*60 - 0.58727*45 + 1.3448*3.0 = 54.5853
Altså et tilfredshedsindeks på **55**

d. Modellen har R^2 = 0.901 og Adjusted R^2 = 0.887. Det vil sige, at modellen beskriver variationen i data godt. F-værdien på 64 og den tilhørende p-værdi på 9.82e-11 siger, at det er ekstremt usandsynligt at alle tre regressorvariable er uden effekt (altså at koefficienterne svarende til x1, x2 og x3 i virkeligheden alle er 0). P-værdierne på estimaterne for koefficienterne er da også alle tæt på 0, bortset fra for x3 (Bekymring). Her er p-værdien 0.2213, så denne koefficient er ikke signifikant forskellig fra 0.

e. Da koefficienten for regressoren Bekymring ikke er signifikant forskellig fra 0 vil jeg teste den reducerede model, hvor kun de signifikante regressorer indgår. En multipel regressionsanalyse med kun Alder og Sygdom giver følgende resultat:

```
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      144.01     6.0522     23.794    3.4318e-17
    x1               -1.038    0.11855    -8.7557     1.278e-08
    x2             -0.55818    0.13217    -4.2231    0.00034986


Number of observations: 25, Error degrees of freedom: 22
Root Mean Squared Error: 7.16
R-squared: 0.894,  Adjusted R-Squared 0.884
F-statistic vs. constant model: 92.7, p-value = 1.91e-11
```

R^2 og Adjusted R^2 er reduceret marginalt til hhv. 0.894 og 0.884, så modellen beskriver variationen i data næsten lige så godt, men til gengæld er modellen simplere med kun to variable, som begge har signifikante koefficienter. Derfor vil jeg foretrække denne model.

f. Estimat og konfidensinterval er beregnet med MatLab funktionen predict:
`[yhat, yci] = predict(mdl,x)`
Funktionen returnerer arrays `yhat` med estimater for Tilfredshed beregnet med regressionsmodellen, og `yci` med 95% konfidenstintervaller.
Resultaterne vises i tabellen nedenfor. De sidste 4 kolonner viser hhv. regressionsligningens estimat for Tilfredshed, residualet, samt den nedre og øvre grænse for konfidensintervallet.

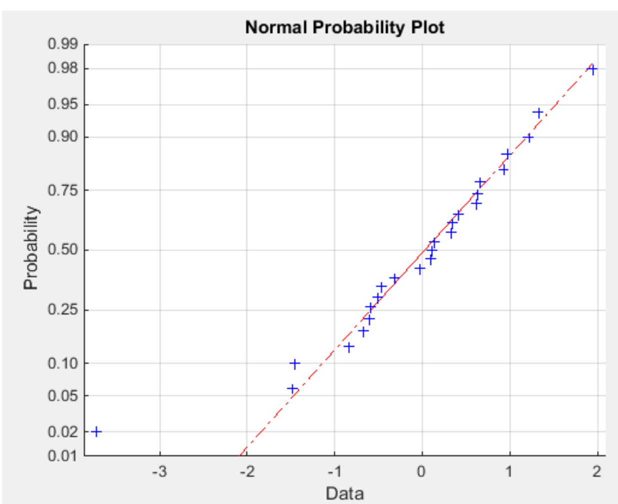
Obs.	Alder	Sygdom	Bekym-ring	Tilfreds-hed	Estimat	Residual	KI lav	KI høj
1	27	42	3.1	75	93.7	-18.7	87.3	100.0
2	51	50	2.4	57	61.0	-4.0	56.2	65.8
3	53	38	2.2	56	65.5	-9.5	60.2	70.9
4	41	30	2.1	88	83.6	4.4	78.5	88.6
5	37	31	1.9	88	87.2	0.8	82.2	92.3
6	28	34	3.1	100	97.2	2.8	91.3	103.2
7	42	30	3	88	83.7	4.3	79.0	88.4
8	50	48	4.2	70	65.7	4.3	62.6	68.8
9	58	61	4.6	52	49.6	2.4	45.0	54.2
10	60	71	5.3	43	42.4	0.6	35.8	49.1
11	62	62	7.2	46	48.0	-2.0	41.2	54.8

12	68	38	7.8	56	56.1	-0.1	47.1	65.2
13	70	41	7	59	51.0	8.0	43.5	58.6
14	79	66	6.2	26	25.1	0.9	18.4	31.9
15	63	31	4.1	52	60.9	-8.9	54.1	67.7
16	39	42	3.5	83	80.7	2.3	76.7	84.7
17	49	40	2.1	75	68.7	6.3	63.9	73.5
18	55	50	2.1	68	56.1	11.9	50.3	61.8
19	46	24	2.8	77	82.4	-5.4	76.2	88.6
20	30	46	3.3	96	88.2	7.8	82.1	94.3
21	35	48	4.5	80	83.0	-3.0	76.8	89.2
22	59	58	2	43	46.7	-3.7	39.5	54.0
23	61	60	5.1	44	47.5	-3.5	43.1	51.9
24	74	65	5.5	26	30.4	-4.4	24.4	36.5
25	38	42	3.2	88	81.4	6.6	77.4	85.5

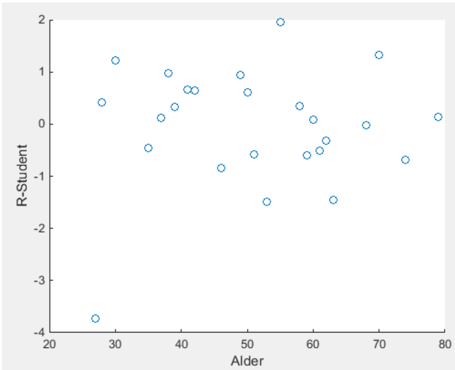
g. Det er bedst at bruge studentiserede residualer (R-Student) til residualplots. Disse er beregnet af MatLab, når man laver regressionsanalysen med fitlm:

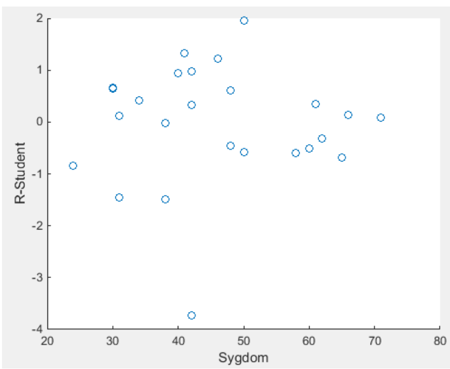
```
rst = mdl.Residuals.Studentized;
```

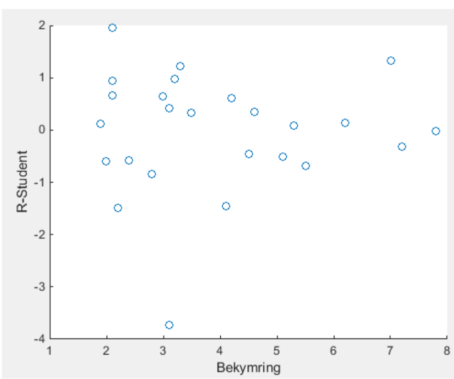
R-Student vist som normalfordelingsplot viser, at residualerne som antaget er fra en fordeling, der ligner normalfordelingen. Dog 'stritter' den første observation ud som outlier. Den ses også i alle residualplots som nederste punkt med R-Student = -3.7.

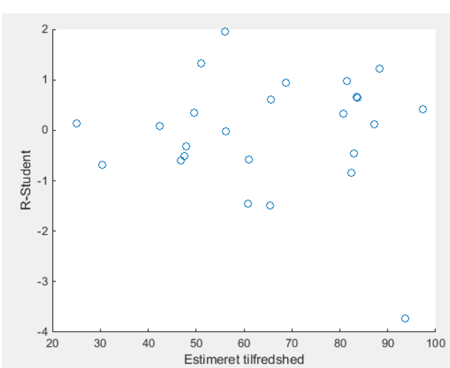


Residualplots giver ikke anledning til bekymring i forhold til modelantagelserne:


R-Student
Alder


R-Student
Sygdom


R-Student
Bekymring


R-Student
Estimeret tilfredshed