Løsningsforslag til Eksamen i M4STI1 2015E

Opgave 1

a. $P(S)$ er sandsynligheden for at en tilfældig stålbjælke er svag, dermed er $P(S) = \mathbf{0.10}$.
$P(S^c) = 1 - P(S) = 1 - 0.10 = \mathbf{0.90}$.
$P(R|S^c)$ er sandsynligheden for at en bjælke udviser det særlige mønster, selv om den er stærk. D.v.s.
$P(R|S^c) = 7\% = \mathbf{0.07}$.
$P(R^c|S^c) = 1 - P(R|S^c) = 1 - 0.07 = \mathbf{0.93}$.

b. Sandsynligheden for at en tilfældig stålbjælke udviser det særlige mønster er $P(R)$. Det følger af loven om total sandsynlighed (lov nr. 6, V&K s. 92):
$P(R) = P(R|S) \cdot P(S) + P(R|S^c) \cdot P(S^c) = 0.87 \cdot 0.10 + 0.07 \cdot 0.90 = \mathbf{0.15}$.

c. Sandsynligheden for at en stålbjælke er svag, givet at den udviser det særlige mønster, er $P(S|R)$. Det kan vi udregne med Bayes' formel (lov nr. 7, V&K s. 93):

$$P(S|R) = \frac{P(R|S) \cdot P(S)}{P(R)} = \frac{0.87 \cdot 0.10}{0.15} = \mathbf{0.58}.$$

d. Her ønskes sandsynligheden for at en bjælke er svag, givet at den ikke udviser det særlige mønster. Udtrykt i symboler er det $P(S|R^c)$. Igen bruges Bayes' formel:

$$P(S|R^c) = \frac{P(R^c|S) \cdot P(S)}{P(R^c)} = \frac{0.13 \cdot 0.10}{1 - 0.15} = \mathbf{0.0153}.$$

MatLab kode:

```
%% M4STI E2015 opgave 1 om stålbjælker
clc; clear all; close all;


%% Oplyst:
p_R_givet_S = 0.87
    % P(R|S) = 0.87, da 87 pct. af de svage (S) viser mønsteret R
p_Rc_givet_S = 1 - p_R_givet_S
    % P(Rc|S) = 0.13. Vi ser kun på de svage (givet S). Af dem viser 87 pct.
    % mønsteret, så resten (13 pct) viser det ikke

%% a
p_S = 0.1
    % P(S) = 0.10, da 10 pct af stålbjælkerne er svage (S)
p_Sc = 1 - p_S
    % P(Sc) = 0.90, da resten (90pct) er stærke
p_R_givet_Sc = 0.07
    % P(R|Sc) = 0.07, da 7 pct. af de stærke (Sc) viser mønsteret R
p_Rc_givet_Sc = 1 - p_R_givet_Sc
    % P(Rc|Sc) = 0.93. Vi ser kun på de stærke (givet Sc). Af dem viser 7 pct
    % mønsteret, så resten (93 pct) viser det ikke.


%% b
% P(R) = P(R|S)*P(S) + P(R|Sc)*P(Sc)
p_R = p_R_givet_S*p_S + p_R_givet_Sc*p_Sc


%% c
% P(S|R) = P(R|S)*P(S)/P(R) (Bayes)
p_S_givet_R = p_R_givet_S * p_S / p_R


%% d
% P(S|Rc) = P(Rc|S)*P(S)/P(Rc) (Bayes)
%         = (1 - P(R|S))*P(S)/(1 - P(R))
p_S_givet_RC = (1 - p_R_givet_S) * p_S / (1 - p_R)
```

Opgave 2

a. Trin 1: Vælg hypoteser.
Da bilproducenten er bekymret for, om middelværdien for bilmodellens NOx udledning er over 0.08 vælger vi den ensidige test:
H_0: $= \mu_0 = 0.08$
H_a: $\mu > \mu_0$

b. Trin 2: Formuler teststatistikken:
Vi kender ikke populationsvariansen, så vi bruger
$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$
som er t-fordelt med $n - 1$ frihedsgrader

c. Trin 3: Formuler den kritiske region:
Vi afviser H_0, hvis
$$|t| > t_0 = t_{n-1,\alpha}$$
Værdien $t_{n-1,\alpha}$ kan findes med MatLab som
$t_0 = \text{tinv}(1 - \alpha, n - 1) = \text{tinv}(1 - 0.05, 10 - 1) = \text{tinv}(0.95, 9) = 1.8331.$

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: $\bar{y} = 0.0841$
Stikprøvevarians: $s^2 = 2.5211\ e - 05$
Stikprøvespredning: $s = 0.0050$

Teststatistik: $t = \frac{0.0841 - 0.08}{0.005/\sqrt{10}} = 2.5822$

Trin 5: Konklusion
Teststatistikken $t = 2.5822$ er større end den kritiske grænse, $t_0 = 1.8331$, så vi kan forkaste nulhypotesen på baggrund af stikprøven. Bilerne af denne model udleder altså for meget NOx.

d. 95% konfidensintervallet for middelværdien af bilmodellens NOx udledning er
$$\bar{y} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 0.0841 \pm 2.2622\frac{0.0050}{\sqrt{10}} = 0.0841 \pm 0.0036 = (\mathbf{0.0805},\ \mathbf{0.0877})$$
Vi ser at den ønskede middelværdi på 0.08 er udenfor konfidensintervallet, men kun lige netop.

e. 95% prediktionsintervallet er

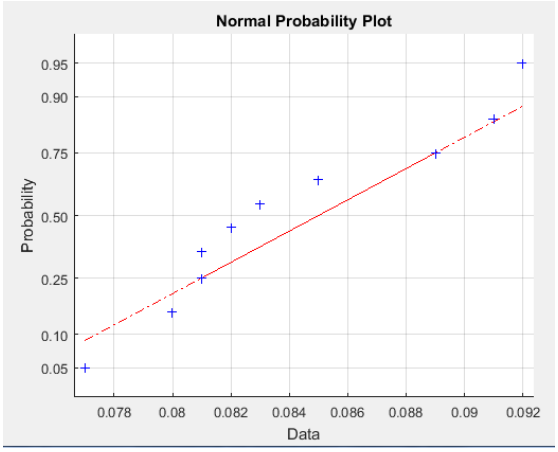
$$\bar{y} \pm t_{n-1,\alpha/2} \cdot s \cdot \sqrt{1+\frac{1}{n}} = 0.0841 \pm 2.2622 \cdot 0.0050 \cdot \sqrt{1+\frac{1}{10}} = 0.0841 \pm 0.0119 = (\mathbf{0.0722},\ \mathbf{0.0960})$$

f. Vores konklusioner i hypotesetesten bygger på antagelsen om den centrale grænseværdisætning, som siger, at teststørrelsen *t* følger t-fordelingen hvis n er 'tilstrækkelig stor'. Hvor stor n behøver at være afhænger af, hvor pæn fordelingen, som stikprøven kommer fra, er. Vi kan se hvordan stikprøvefordelingen ser ud med et Stem-and-leaf plot med MatLab funktionen stemleafplot(y,-3), hvor y er et array med NOx målingerne. Parameteren -3 vælges, fordi vi gerne vil have sidste ciffer i observationerne som blade i plottet, og dette ciffer er det tredje efter decimaltegnet.

```
>> stemleafplot(y,-3)
  7 | 7
  8 | 0 1 1 2 3 5 9
  9 | 1 2
key: 36|5 = 0.365
stem unit: 0.010
leaf unit: 0.001
```

Stem-and-leaf plottet viser at data kommer fra en pæn fordeling med et enkelt toppunkt, nogenlunde symmetrisk og med hurtigt uddøende haler.

Vi kan desuden teste antagelsen med et normalfordelingsplot, som vi får med MatLab funktionen normplot(y).



Normalfordelingsplottet viser ikke en særligt lineær sammenhæng, så vi er ikke overbevist om, at stikprøvens fordeling ligner normalfordelingen. Det er lidt bekymrende, at der kun er 10 observationer i stikprøven. Med over 30 observationer ville vi være næsten sikre på, at den centrale grænseværdisætning holder. Det kan derfor anbefales, at bilproducenten foretager NOx målingen på flere biler for at være mere sikker på, at modelantagelserne holder.

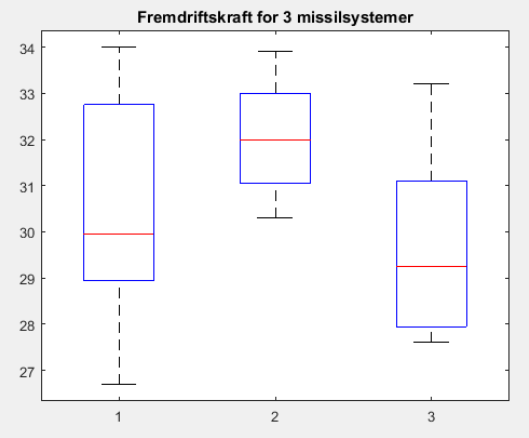
MatLab kode

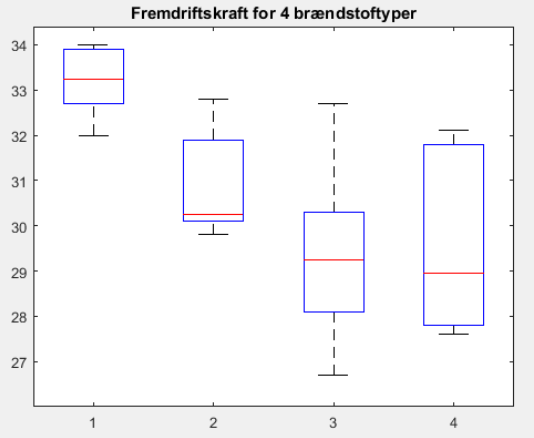
```
%% M4STI E2015 opgave 2 om bilers NOx udledning
clc; clear all; close all;


%% a Hypotesetest
%  skridt 1
mu0 = 0.08
%  H0:  my = mu0
%  Ha:  my > mu0
% Bilproducenten er bekymret for om middelværdien er over den tilladte på
% 0.08, så vi vælger ensidig test med Ha:  my > mu0

%% b
% Skridt 2
% Vi kender ikke populations-spredningen, så vi estimerer den som s ud fra
% stikprøven. Derfor er teststørrelsen t t-fordelt med n-1 frihedsgrader:
% t = (y_streg - mu0)/(s/sqrt(n))

% skridt 3
alpha = 0.05
% Vi ved at der testes på n=10 biler
n = 10
df = n-1

%% c
% Kritisk grænse t0 for ensidig hypotesetest til højre
t0 = tinv(1-alpha, n-1)

% skridt 4
% Nu kan vi indlæse data og beregne teststørrelsen
y = xlsread('M4STI1_2015E_data.xlsx', 'A:A')
n_test = size(y,1) % tester at der er indlæst 10 observationer

y_streg = mean(y)
sumy2 = sum(y .* y)
s2 = (n*sumy2 - (sum(y))^2)/(n*(n-1))
s = sqrt(s2)
% Teststørrelsen
t = (y_streg - mu0)/(s/sqrt(n))

% Skridt 5
% teststørrelsen t = 2.5822 er større end den øvre kritiske grænse,
% t0 = 1.8331, så vi forkaster H0 på baggrund af stikprøven. Bilerne
% udleder for meget NOx.

% p-værdi:
pValue = 1 - tcdf(t, n-1)

%% d
% 95 pct. konfidensinterval
t_alphahalf = -tinv(alpha/2, n-1)
CI_width = t_alphahalf*s/sqrt(n)
```

```
CI_low = y_streg - CI_width
CI_high = y_streg + CI_width


%% e
% 95 pct. prædiktionsinterval
PI_width = t_alphahalf*s*sqrt(1 + 1/n)

PI_low = y_streg - PI_width
PI_high = y_streg + PI_width


%% f
% Antagelser
% Vi har antaget den centrale grænseværdisætning, som siger, at
% teststørrelsen følger t-fordelingen hvis n er tilstrækkelig stor. Hvor
% stor n behøver at være afhænger af, hvor pæn fordelingen, som stikprøven
% kommer fra, er.

stemleafplot(y,-3)
% Stem-and-leaf plottet viser at data kommer fra en pæn fordeling med et
% enkelt toppunkt, nogenlunde symmetrisk og med hurtigt uddøende haler

normplot(y)
% Normalfordelingsplottet viser ikke en særligt lineær sammenhæng, så vi
% er ikke overbevist om, at stikprøvens fordeling ligner normalfordelingen.
% Det er lidt bekymrende, at der kun er 10 observationer i stikprøven.
% Med over 30 observationer ville vi være næsten sikre på, at den centrale
% grænseværdisætning holder. Det kan derfor anbefales, at bilproducenten
% foretager NOx målingen på flere biler for at være mere sikker på, at
% modelantagelserne holder.
```

Opgave 3

a. Boxplot laves i MatLab med funktionen boxplot(X, G), hvor X indeholder data og G grupperingen. Nedenfor vises boxplots over fremdriftskraften for de tre missilsystemer (til venstre) og for de fire brændstoftyper (til højre):





Vi ser at missilsystem 2 har mindre variabilitet og højere median end de andre to systemer. Brændstoftyperne 1 og 2 har mindre variabilitet end 3 og 4, men 1 har højere median end 2. Brændstoftype 3 og 4 har median på samme niveau.

b. Variansanalysen med MatLab funktionen anovan giver følgende resultat:

Analysis of Variance

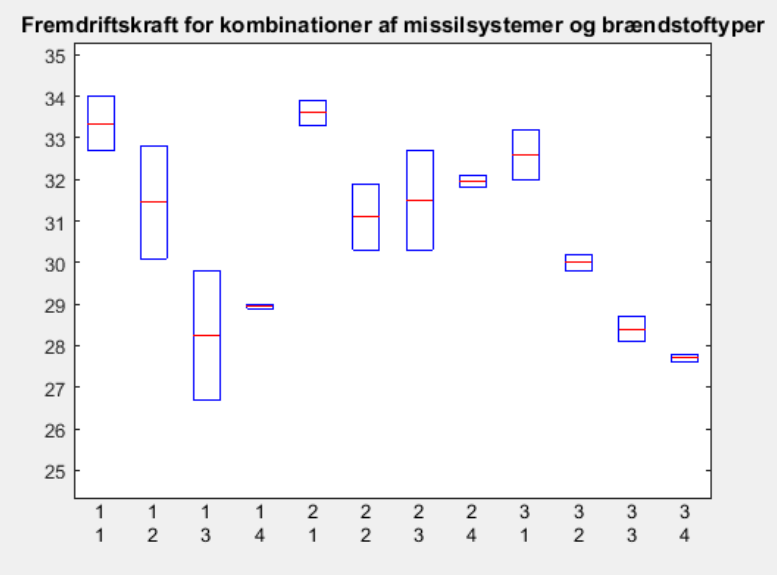
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
MissilSystem	23.002	2	11.5012	9.4	0.0035
Braendstof	55.671	3	18.5571	15.16	0.0002
MissilSystem*Braendstof	12.918	6	2.1529	1.76	0.1906
Error	14.685	12	1.2237		
Total	106.276	23			

Analysen viser, at missilsystemet har en signifikant effekt på fremdriftskraften (p-værdi 0.0035). Ligeledes er der signifikant effekt af brændstof (p-værdi 0.0002). Selv om der lader til at være interaktion imellem missilsystem og brændstof, så viser analysen at interaktionen ikke er signifikant (p-værdi 0.1906).

c. Missilsystemerne og brændstoftyperne undersøges parvist med MatLab funktionen multcompare. Den viser, at missilsystem 2 er forskellig fra både 1 og 3, men missilsystem 1 og 3 er ikke signifikant forskellige. For de fire brændstoftyper gælder der:
Brændstoftype 1 er signifikant forskellig fra 2, 3 og 4.
Brændstoftype 2 er signifikant forskellig fra 1 og 3, men ikke 4.

Brændstoftype 3 er signifikant forskellig fra 1 og 2, men ikke 4.
Brændstoftype 4 er signifikant forskellig fra 1, men ikke 2 og 3.

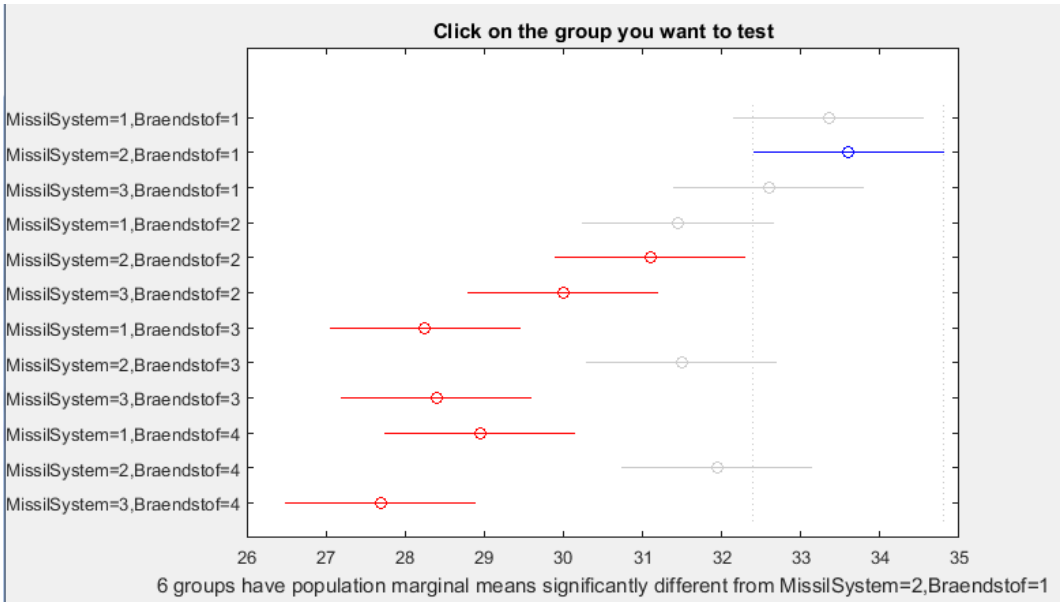
d. Nedenfor vises et boxplot med alle kombinationer af missilsystem og brændstoftype. Det ses, at den højeste median fås med missilsystem 2 og brændstoftype 1 (femte kasse fra venstre). Missilsystem 1 og brændstoftype 1 giver også en høj median, men med lidt større variabilitet.



Alle kombinationerne kan også undersøges parvist med kommandoen:

```
multcompare(stats,'Alpha',0.05, 'CType','lsd', 'Dimension', [1,2])
```

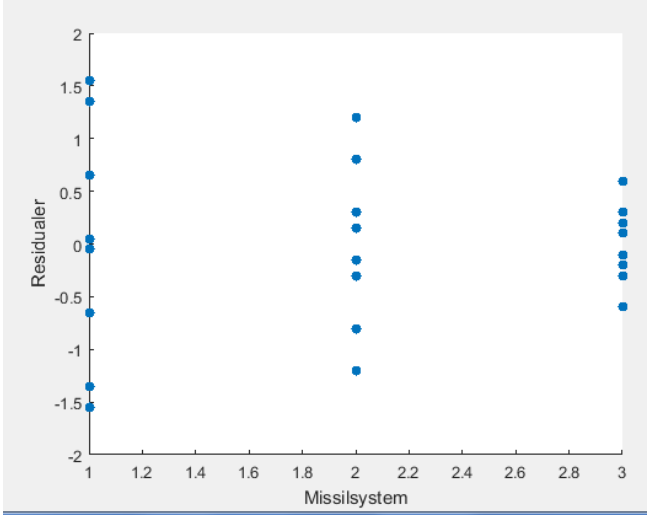
Det giver følgende resultat:

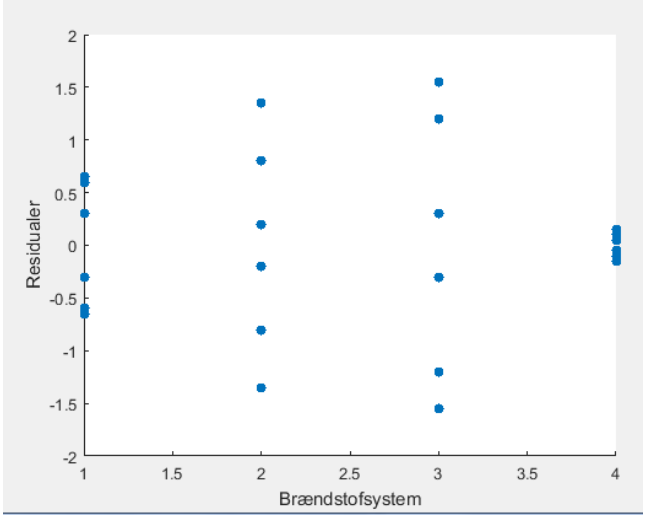


Vi ser også her, at kombinationen missilsystem 2 og brændstoftype 1 (markeret blå i figuren) har den højeste forventede fremdriftskraft, men der er 5 kombinationer (markeret grå), der ikke er signifikant forskellige.

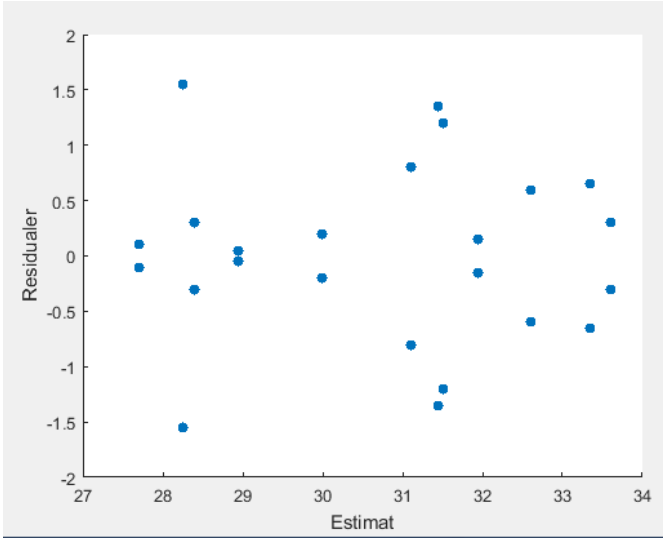
Jeg vil vælge missilsystem 2 og brændstoftype 1, fordi det både har den højeste forventede fremdriftskraft og desuden mindre variabilitet end de andre kombinationer med høj fremdriftskraft, altså højere pålidelighed.

e. Vi har antaget, at fremdriftskraften kan beskrives som en overordnet middelværdi med en effekt af missilsystem, en effekt af brændstofsystem og en effekt af interaktion mellem de to faktorer. Residualet har vi antaget er normalfordelt med middelværdi 0 og samme varians for alle faktorer. Residualerne skal derfor gerne være tilfældige og uafhængige af faktorniveau i residual plots. Det lader ikke til at være tilfældet i nedenstående residualplots for missilsystem og brændstofsystem. Residualerne for missilsystem 3 er mindre end for de andre, og residualerne for brændstoftype 1 og især 4 er mindre end for 2 og 3.

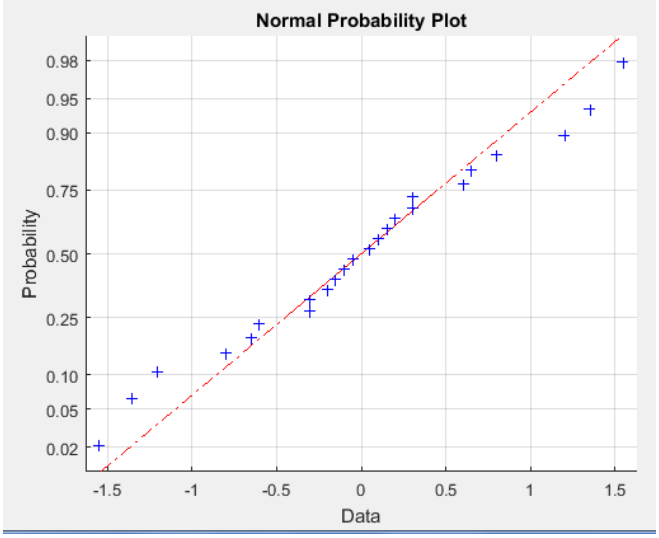




Residualerne skal også være uafhængig af estimatets størrelse (y hat). Som det ses af næste residualplot lader det til at være opfyldt, da der ikke lader til at være systematik i residualernes størrelse:



Vi kan teste om residualerne er normalfordelte med et normalfordelingsplot. Residualerne fordeler sig nogenlunde lineært i plottet, men der er problemer med yderområderne, dvs. for residualer med numerisk stor værdi.



Residualanalysen er ikke helt overbevisende. Jeg er ikke helt tryg ved, om vores antagelser ikke holder.

MatLab kode

```
%% M4STI E2015 opgave 3 om missilsystemer
clc; clear all; close all;

M = xlsread('M4STI1_2015E_data.xlsx', 'C:E')

MissilSystem = M(:,1)    % Faktor 1
Braendstof = M(:,2)      % Faktor 2
Fremdriftskraft = M(:,3)    % Respons

%% a
figure(1)
boxplot(Fremdriftskraft, MissilSystem)
title('Fremdriftskraft for 3 missilsystemer')

figure(2)
boxplot(Fremdriftskraft, Braendstof)
title('Fremdriftskraft for 4 brændstoftyper')

%% b
[p,table,stats] = anovan(Fremdriftskraft, [MissilSystem,Braendstof], 'model',
'interaction', 'varnames',{'MissilSystem','Braendstof'})

%% c
%[c,m,h,gnames] = multcompare(stats,'Alpha',0.05, 'CType','lsd')
figure(4)
[c,m] = multcompare(stats,'Alpha',0.05, 'CType','lsd', 'Dimension', [1])

figure(5)
[c,m] = multcompare(stats,'Alpha',0.05, 'CType','lsd', 'Dimension', [2])

%% d
figure(6)
boxplot(Fremdriftskraft, [MissilSystem,Braendstof])
title('Fremdriftskraft for kombinationer af missilsystemer og brændstoftyper')

figure(7)
[c,m] = multcompare(stats,'Alpha',0.05, 'CType','lsd', 'Dimension', [1,2])

% Svar: kombinationen af missilsystem 2 og brændstoftype 1 er bedst, både
% med højeste median og med fhv. lille varians.
% Men der er 5 andre kombinationer, som ikke er signifikant forskellige

%% e
% resudualer findes i stats objektet, lavet af anovan
resid = stats.resid

figure(8)
normplot(resid)

figure(9)
scatter(MissilSystem, resid, 'filled')
xlabel('Missilsystem')
```

```
ylabel('Residualer')

figure(10)
scatter(Braendstof, resid, 'filled')
xlabel('Brændstofsystem')
ylabel('Residualer')

y_hat = [Fremdriftskraft - resid]
% For hver observation beregnes estimatet for fremdriftskraften

figure(11)
scatter(y_hat, resid, 'filled')
xlabel('Estimat')
ylabel('Residualer')

% Antagelserne er ikke særligt godt underbygget; Normplottet er ikke
% lineært (fig. 8), residualerne er ikke ens for de enkelte faktorer (fig.
% 9 og 10, men residualerne lader dog ikke til at afhænge af y_hat
% Det er tvivlsomt om eksperimentet kan bruge til at konkludere,
% hvilken kombination af missilsystem og brændstoftype, der bør vælges.
```

Opgave 4

a. Vi laver en simpel lineær regression med MatLab funktionen fitlm, der giver dette resultat:

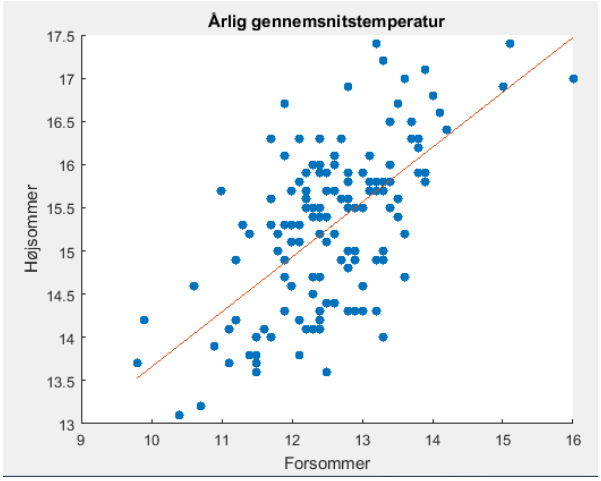
```
Estimated Coefficients:
                   Estimate        SE       tStat       pValue
                   ________    ________    ______    _________

    (Intercept)      7.2963      0.8232    8.8633    3.3397e-15
    x1               0.63605    0.065483    9.7133    2.4276e-17


Number of observations: 141, Error degrees of freedom: 139
Root Mean Squared Error: 0.737
R-squared: 0.404,  Adjusted R-Squared 0.4
F-statistic vs. constant model: 94.3, p-value = 2.43e-17
```

Det giver følgende regressionsligning, hvor x er temperaturen i forsommeren og y er den estimerede temperatur for højsommeren:
$y = 7.2963 + 0.63605x$

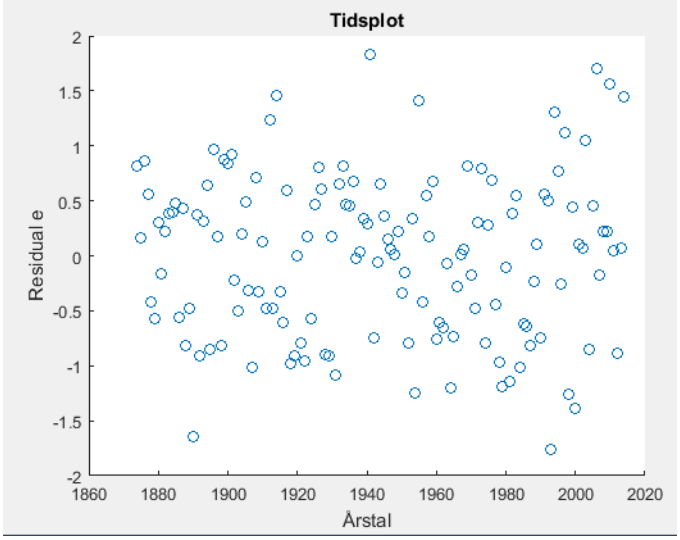
b. Vi ser, at både koefficienten for skæring og for hældningskoefficient er signifikant forskellige fra 0 (begge p-værdier er næsten 0). Der er altså en klar sammenhæng mellem temperaturen i forsommeren og i højsommeren. Der er dog meget usikkerhed i forudsigelserne, hvilket kommer til udtryk i, at R-squared og Adjusted R-Squared er temmelig lave, kun hhv. 0.404 og 0.4. Det vil sige, at modellen kun kan forudsige omkring 40 pct. af variabiliteten i data. Det er et udtryk for, at det er svært at lave vejrudsigter i Danmark, at vejret er meget uforudsigeligt.

c. $y = 7.2963 + 0.63605 \cdot 11.25 = \mathbf{14.45}$

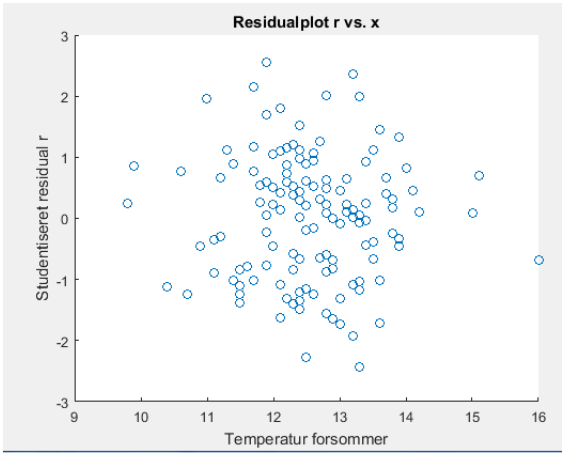
d. Denne figur viser et scatter plot over de målte gennemsnitstemperaturer i forsommeren og højsommeren for hvert år. Desuden er regressionsligningen tegnet ind. Figuren viser, at det ser fornuftigt ud med en lineær model, men også at der er stor variabilitet i data.

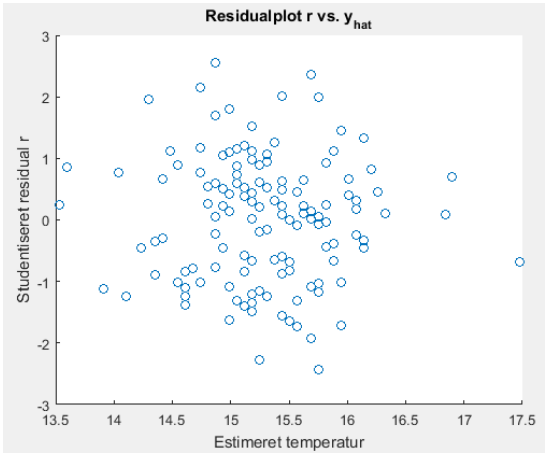


e. Figuren nedenfor viser de rå residualer plottet mod årstal. Residualerne lader til at fordele sig tilfældigt over tid, men måske er der en tendens til, at der er flere residualer med høj, positiv værdi i de seneste år. Det ville muligvis kunne forklares med klimaeffekten - klimaet bliver varmere med årene, og modellen er parameteriseret med data helt tilbage fra 1874. Derfor kan man frygte, at modellen estimerer for lavt for fremtidige år.

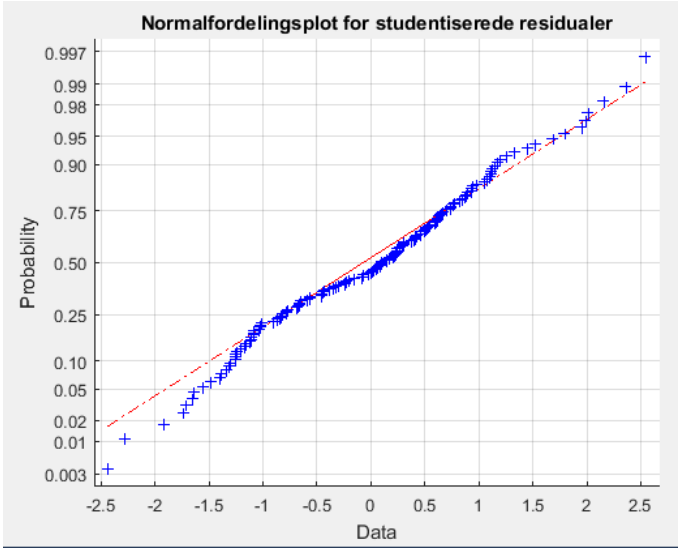


De næste to figurer viser studentiserede residualer, hhv. som funktion af temperaturen i forsommeren og den estimerede temperatur i højsommeren. Residualerne lader til at fordele sig tilfældigt. Der er måske en tendens til større spredning i residualerne i midterfeltet, men det er også her, de typiske data ligger.





Residualerne skal gerne være normalfordelte. Det tester vi med et normalfordelingsplot:



Residualplottet viser, at residualerne kommer fra en pæn fordeling, der ligner normalfordelingen. Vores antagelser er tilfredsstillende overholdt.

MatLab kode

```
%% M4STI E2015 opgave 4 om prognose af sommervejr
clc; clear all; close all;


%% a
M = xlsread('M4STI1_2015E_data.xlsx', 'G:I')

aar = M(:,1);
x = M(:,2);
y = M(:,3);

mdl = fitlm(x,y)
anova_sti(mdl)
% y = 7.2963 + 0.63605*x


%% b
% Begge koefficienter har en meget lille p-værdi tæt på 0. Der er
% utvivlsomt en sammenhæng mellem middeltemperaturen i forsommeren og
% højsommeren. Anova testens F-værdi og den tilhørende p-værdi siger det
% samme. Dog er både R2 og Adjusted R2 temmelig lave, kun omkring 0.4.
% Modellen forudsiger kun 40% af variationen i data. Det fortæller os,
% at der er meget usikkerhed i modellen. Trods en signifikant
% sammenhæng mellem regressor og respons kan modellen ikke forudsige det
% kommende vejr særligt præcist


%% c
y_2015 = 7.2963 + 0.63605*11.25
% y_2015 = 14.4519


%% d
y_hat = 7.2963 + 0.63605*x

figure(1);
hold on;
scatter(x,y,'filled');
plot(x, y_hat)
title('Årlig gennemsnitstemperatur');
xlabel('Forsommer');
ylabel('Højsommer');
hold off;

% Alternativ figur med den indbyggede matlabfunktion plot:
figure(2);
plot(mdl);
title('Årlig gennemsnitstemperatur');
xlabel('Forsommer');
ylabel('Højsommer');
```

```
%% e
% Brug enten almindelige residualer e eller studentiserede r (bedst)
e = y - y_hat
r = mdl.Residuals.Studentized

figure(3);
scatter(aar, e);
title('Tidsplot');
xlabel('Årstal');
ylabel('Residual e');
% Residualerne fordeler sig tilfældigt over tid

figure(4);
scatter(x, r);
title('Residualplot r vs. x');
xlabel('Temperatur forsommer');
ylabel('Studentiseret residual r');
% Residualerne er uafhængige af x, dog er der måske en tendens til større
% residualer i midterområdet.

figure(5);
scatter(y_hat, r);
title('Residualplot r vs. y_{hat}');
xlabel('Estimeret temperatur');
ylabel('Studentiseret residual r');
% Residualerne er uafhængige af y_hat, dog er der måske en tendens til større
% residualer i midterområdet.

figure(6);
normplot(r);
title('Normalfordelingsplot for studentiserede residualer');
% Residualerne kommer fra en 'pæn' fordeling, da de ligger nogenlunde på en
% ret linje i normalfordelingsplottet.

stemleafplot(r, -1)

figure(7);
histogram(r,7)
% Både stem-and-leaf plot og histogram viser at residualerne kommer fra en
% 'pæn' fordeling med et toppunkt og hurtigt uddøende haler. Nogenlunde
% symmmetrisk.
```