

OPGAVE 1

Følgende betegnelser bruges for hændelsen om en tilfældigt fremstillet ventil:

- A: Ventilen er fremstillet på maskine A
 B: Ventilen er fremstillet på maskine B
 In: Ventilen er intakt, dvs. den kan anvendes i den videre produktion
 D: Ventilen er defekt, dvs. den kasseres

- a. Udfyld antal i nedenstående tabel, idet nødvendige mellemregninger medtages.

	Maskine A	Maskine B	I alt
Intakt ventil, In	$576 - 226 = 350$	226	$600 - 24 = 576$
Defekt ventil, D	$360 - 350 = 10$	$24 - 10 = 14$	$0.04 \cdot 600 = 24$
I alt	$0.60 \cdot 600 = 360$	$600 - 360 = 240$	600

En tilfældig blandt de fremstillede ventiler udtages.

- b. Beregning af sandsynlighederne for:

Ventilen er fremstillet på maskine B, $P(B)$:

$$P(B) = \frac{240}{600} = \mathbf{0.4000}$$

Ventilen er intakt, $P(\text{In})$:

$$P(\text{In}) = \frac{576}{600} = \mathbf{0.9600}$$

Ventilen er defekt, $P(D)$:

$$P(D) = \frac{24}{600} = \mathbf{0.04}$$

eller

$$P(D) = 1 - P(\text{In}) = 1 - 0.9600 = \mathbf{0.04}$$

Ventilen er fremstillet på maskine B og er defekt, $P(B \cap D)$:

$$P(B \cap D) = \frac{14}{600} = \mathbf{0.0233}$$

Ventilen er intakt, når den er fremstillet på maskine B, $P(\text{In}|B)$:

$$P(\text{In}|B) = \frac{226}{240} = \mathbf{0.9417}$$

Ventilen er defekt, når den er fremstillet på maskine B, $P(D|B)$:

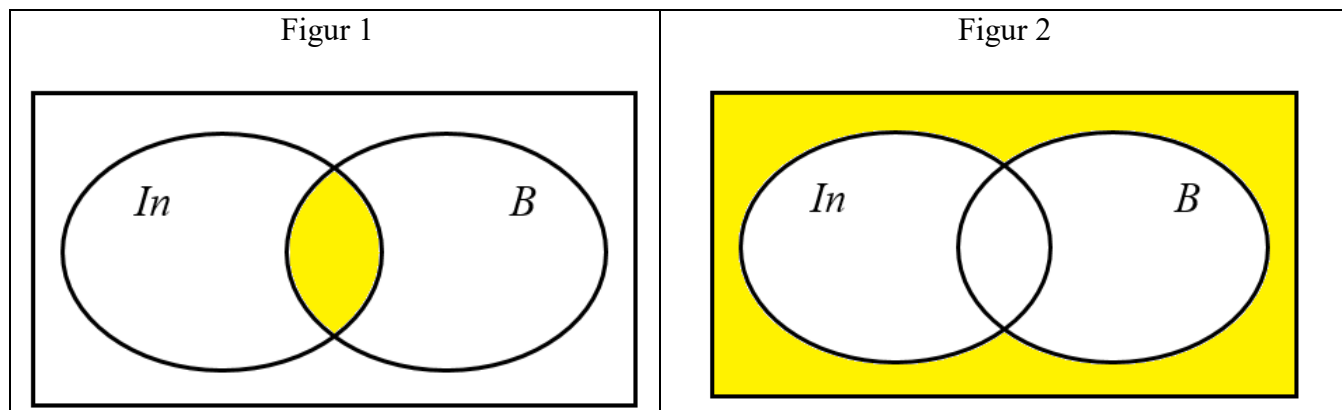
$$P(D|B) = \frac{14}{240} = \mathbf{0.0583}$$

eller

$$P(D|B) = 1 - P(\text{In}|B) = 1 - 0.9417 = \mathbf{0.0583}$$

OPGAVE 1 fortsat

- c. I figur 1 og i figur 2 ses et Venn-diagram med hvide og gule områder.



- Beskrivelse af hvilke ventiler det gule område svarer til i figur 1.
Det gult markerede område er fælleshændelsen mellem de intakte ventiler, der er fremstillet og de ventiler, der er fremstillet på maskine B: $In \cap B$. Dvs. intakte ventiler fremstillet på maskine B.
- De gult markerede hændelser svarer det gule område til i figur 2:
 - $(In \cap B)^c$
 - $(In \cup B)^c$
 - $In^c \cap B$
 - $In^c \cap B^c$
 - $In^c \cup B^c \cup (In \cap B)$

Opgave 2

I en produktion fyldes der solcreme på små tuber. Erfaringsmæssigt er massen af en tube solcreme normalfordelt med en standardafvigelse på 2.8 g.

Som led i en kvalitetskontrol udtages der en tilfældig stikprøve på 9 tuber solcreme, som efterfølgende afvejes. De afvejede masser, i enheden g, er indført i nedenstående tabel:

68.6	69.0	66.0	70.5	70.6	67.2	65.1	69.6	68.1
------	------	------	------	------	------	------	------	------

- a. Beregning af stikprøvens middelværdi: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{9} \sum_{i=1}^9 y_i = \frac{1}{9} \cdot 614.7 = \mathbf{68.3}$

I MATLAB: $\text{sum}(y) = 614.7$ $\bar{y} = \text{mean}(y) = 68.3$

- b. Bestemmelse et 95%-konfidensinterval for populations- middelværdien, μ :

Konfidensintervallet bestemmes vha. formlen: $\bar{y} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

hvor værdien af: $\alpha = 1 - 0.95 = 0.05$, $\sigma = 2.8$ $n = 9$

Vha. MATLAB findes: $z_{\frac{\alpha}{2}} = -\text{norminv}(\frac{\alpha}{2}) = -\text{norminv}(0.025) = 1.9600$

Dvs. 95%-konfidensintervallet for middelværdien er:

$$\bar{y} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 68.3 \pm 1.9600 \cdot \frac{2.8}{\sqrt{9}} = 68.3 \pm 1.8293, \text{ dvs. } [66.4707; 70.1293] \approx [\mathbf{66.5; 70.1}]$$

- c. Hvad beskriver et 95%-konfidensinterval for middelværdien? Forklar kort med ord.

Et 95%-konfidensinterval for middelværdien angiver det interval, hvor vi med 95% sikkerhed finder populationens middelværdi.

- d. Bestemmelse af mindste stikprøvestørrelse så, at 95%-konfidensintervallet har en intervalbredde på højst 3.0 g, dvs. at 95%-konfidensintervallet er: $\bar{y} \pm 1.5$ g.

Vi har: $\bar{y} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ og $\bar{y} \pm 1.5$

Dvs.

$$z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 1.5 \Rightarrow n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{1.5} \right)^2 = \left(\frac{1.9600 \cdot 2.8}{1.5} \right)^2 = 13.3853 \text{ Der rundes op til nærmeste heltal, dvs. } \mathbf{n = 14}$$

En stikprøvestørrelse på 14 vil give et 95%-konfidensinterval med den ønskede bredde.

Opgave 3

En virksomhed producerer solceller. Varedeklarationen for virksomhedens solceller angiver, at en solcelle har en effekt på mindst 100W (watt) i klart solskinsvejr.

Virksomheden ønsker, at undersøge om middeleffekten af de fremstillede solceller mindst er 100W.

Der udtages derfor en tilfældig stikprøve på 32 solceller, som testes i en teststander, der simulerer klart solskinsvejr. De målte effekter i W er indført i nedenstående tabel:

106	92	101	105	106	102	94	105	96	103	101
105	98	100	98	102	108	102	97	107	101	92
103	101	108	99	98	99	102	105	100	110	

Opgaven drejer sig om at undersøge ved hjælp af en hypotesetest, om middelværdien af effekten for de fremstillede solceller opnår den ønskede værdi på mindst 100 W, når der vælges et signifikansniveau på 5%.

- a. Nulhypotese og alternativ hypotese for hypotesetesten.

$$H_0: \mu = 100$$

$$H_a: \mu < 100$$

Der vælges en en-sidet test, da effekten for de fremstillede solceller ikke må være under de ønskede 100 W.

- b. Formel for teststørrelsen (teststatistikken), og angivelse af hvilken fordeling den følger.

Da populationsvariansen ikke er kendt, så bestemmes standardafvigelsen, s , ud fra stikprøven.

Teststørrelsen, $t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$ som er t -fordelt med $n-1$ frihedsgrader

- c. Bestemmelse af den kritiske værdi og angivelse af det kritiske område for testen.

Signifikansniveau er $\alpha = 0.05$

Der testes på $n = 32$ solceller og antal frihedsgrader er: $df = n - 1 = 32 - 1 = 31$

Testen er en-sidet, så H_0 afvises, hvis $t < t_{n-1, \alpha}$

Værdien af den kritiske værdi $t_{n-1, \alpha} = t_{31, 0.05}$ findes vha. MATLAB:

$$t_{31, 0.05} = \text{tinv}(\alpha, n-1) = \text{tinv}(0.05, 32 - 1) = \text{tinv}(0.05, 31) = -1.6955$$

Dvs. det kritiske område: H_0 afvises, hvis $t < t_{31, 0.05} = -1.6955$

Opgave 3 fortsat

- d. Beregning teststørrelsens (teststatistikken) værdi, idet nødvendige mellemregninger medtages.

Dette gøres vha MATLAB:

Først beregnes:

$$\text{Stikprøvens middelværdi: } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{32} \sum_{i=1}^{32} y_i = 101.4375$$

$$\text{Stikprøvens varians: } s^2 = \frac{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}{n(n-1)} = \frac{32 \cdot 329890 - 3246^2}{32 \cdot (32-1)} = 20.125$$

$$\text{Stikprøvens standardafvigelse: } s = \sqrt{20.125} = 4.4861$$

I MATLAB:

$$\bar{y} = \text{mean}(y) = 101.4375$$

$$s^2 = \text{var}(y) = 20.1250$$

$$s = \text{std}(y) = 4.4861$$

$$\text{Teststørrelsen (teststatistikken): } t = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{101.4375 - 100}{4.4861/\sqrt{32}} = 1.8127 \approx 1.81$$

- e. Konklusion på hypotesetesten.

Teststørrelsen (teststatistikken) $t = 1,81$, dvs $t = 1.81 > t_{31,0.05} = -1.6955$ og dermed ligger t ikke i det kritiske område, men i H_0 -accept-området

Dermed kan **H_0 -hypotesen ikke forkastes** på baggrund af stikprøven.

Dvs. det kan ikke afvises, at middeleffekten for de fremstillede solceller har en ønsket værdi på mindst 100 W.

Dette spørges der ikke om:

p -værdien kan findes vha. MATLAB:

$$p - \text{værdi} = \text{tcdf}(t, n - 1) = \text{tcdf}(1.8127, 31) = 0.9602$$

- f. Hvilken antagelse er der foretaget for at udføre hypotesetesten?
Er antagelsen rimelig?

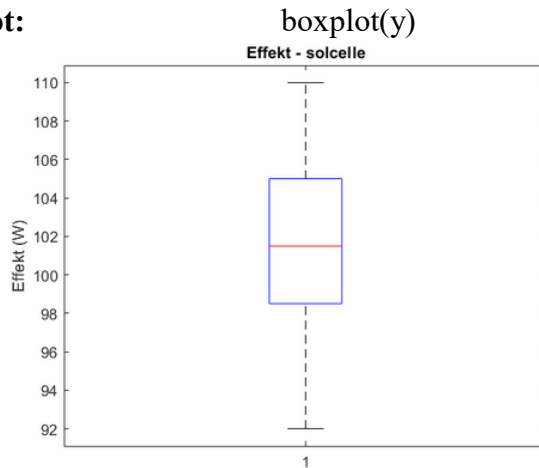
Det er antaget, at den centrale grænseværdisætning (CLT) gælder. Det betyder at teststørrelsen (teststatistikken) følger en t -fordeling, hvis n er tilstrækkelig stor.

Med en stikprøvestørrelse $n = 32 \geq 30$, så er n tilstrækkelig stor, og den centrale grænseværdisætning gælder.

Opgave 3 fortsat

g. Der udføres følgende plot i MATLAB:

Boxplot:



Boxplottet har en median på 101.5, som stort set er sammenfaldende med middelværdien på 101.4.

Det interkvartile range ligger pænt symmetrisk omkring medianen. Med nedre kvartil på 98.5 og øvre kvartil på 105.0. Den øvre del er lidt bredere end den nedre del.

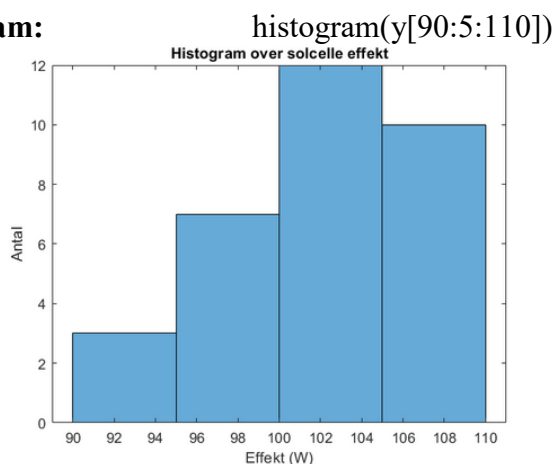
Den nedre kost er lidt længere end den øvre. Dvs. fordelingen af data er lidt venstre skæv.

Der er ingen outliers.

Dvs. data kommer fra en stort set pæn fordeling.

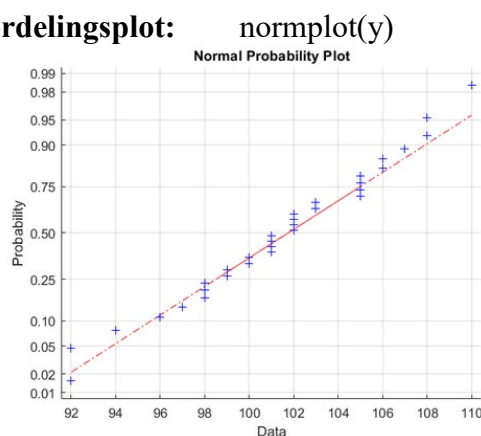
$$Q = \text{prctile}(y,[25,50,75]) = [98.5,101.5,105.0]$$

Histogram:



Histogrammet viser, at data kommer fra en pæn fordeling med en enkelt top, lidt venstre skæv og med hurtigt uddøende haler.

Normalfordelingsplot:



Normalfordelingsplottet viser en pæn lineær sammenhæng. Dette støtter antagelsen om en pæn fordeling, dvs. at stikprøvens fordeling ligner en normalfordeling.

Boxplot, histogram og normalfordelingsplot viser alle, at stikprøven kommer fra en pæn fordeling, der godt kunne ligne en normalfordeling. Så også her er kravet for at den centrale græseværdi sætning gælder, opfyldt.

OPGAVE 4

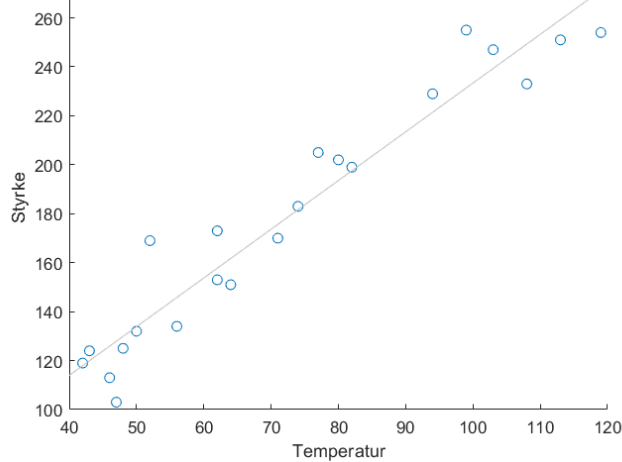
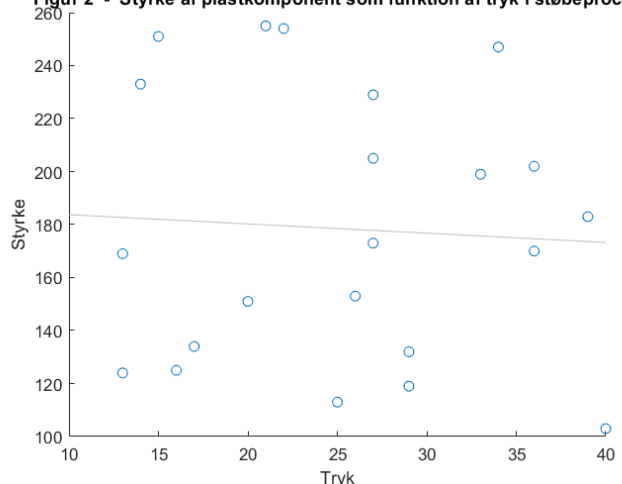
En plastkomponent til pumper fremstilles ved hjælp af sprøjtstøbning på en maskine. Maskinen giver mulighed for at kontrollere temperaturen af støbeform og trykket i støbeprocessen.

For at undersøge om styrken af plastkomponenten afhænger af temperaturen af støbeform og trykket i støbeprocessen udføres en styrketest i et standardudstyr.

Der foretages sammenhørende målinger af temperatur af støbeform og trykket i støbeprocessen for 22 plastkomponenter, og efterfølgende testes styrken af hver plastkomponent.

De målte data er indført i tabel, hvor temperaturen af støbeform betegnes T , trykket i støbeprocessen betegnes p , og styrken af plastkomponent betegnes S .

- a. I MATLAB laves der laves et scatterplot af data:
 med temperaturen af støbeform, T , som regressorvariabel og styrken af plastkomponent, S , som responsvariabel, figur 1
 med trykket i støbeprocessen, p , som regressorvariabel og styrken af plastkomponent, S , som responsvariabel, figur 2

Scatterplot:	
<p>Figur 1 - Styrke af plastkomponent som funktion af temperatur på støbeform</p> 	<p>Scatterplot viser, at styrken af plastkomponent, S, øges, når temperaturen af støbeform, T, øges. Hermed er sammenhængen mellem temperatur, T, og styrke, S, positivt korreleret. Sammenhængen synes lineær.</p>
<p>Figur 2 - Styrke af plastkomponent som funktion af tryk i støbeproces</p> 	<p>Scatterplot viser tilfældig variation af data. Der synes ingen sammenhæng mellem styrke af plastkomponent, S, og tryk i støbeproces, p. Dermed er tryk, p, og styrke, S, ukorrelerede.</p>

OPGAVE 4 fortsat

- b. Multipel lineær regressionsanalyse, der beskriver styrke af plastkomponent (S) som funktion af temperatur af støbeform (T) og af trykket i støbeprocessen (p),
og opskrivning af regressionsligningen: $S = b_0 + b_1T + b_2p$

Der udføres en multipel lineær regression i MATLAB vha: `mdl = fitlm([T,p], S)`

Og der fås følgende output:

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

Estimated Coefficients:

```

	Estimate	SE	tStat	pValue
(Intercept)	37.464	13.869	2.7012	0.014153
x1	1.9884	0.12955	15.349	3.6695e-12
x2	-0.1177	0.36857	-0.31934	0.75295

```

Number of observations: 22, Error degrees of freedom: 19
Root Mean Squared Error: 14.5
R-squared: 0.926, Adjusted R-Squared: 0.918
F-statistic vs. constant model: 118, p-value = 1.9e-11

```

Dermed fås regressionsligningen:

$$S = 37.464 + 1.9884 \cdot T - 0.1177 \cdot p$$

- c. Vurdering om modellen beskriver observationerne godt ud fra regressionsanalysens statistikker (f.eks. R^2 , F og p -værdi). Der anvendes et signifikansniveau, $\alpha = 0.05$ ved vurderingen.

$R^2 = 0.926$, dvs. 92.6% af variationen forklares af modellen. R^2 større end 90% er sædvanligvis tilfredsstillende.

Adjusted $R^2 = 0.918$. Adjusted R^2 justerer for antal parametre i modellen, her 2: T og p .

Adjusted R^2 er kun en anelse lavere end R^2 , dvs. vi straffes kun en anelse ved at bruge både T og p i modellen.

Altså modellen fitter godt til målingerne.

$F = 118$ har en p -værdi på $1.9 \cdot 10^{-11}$. Med en nulhypotese, H_0 , der siger, at data er ukorrelerede, en alternativ hypotese, H_a , der siger, at data er korrelerede, og med et signifikansniveau, α , på 5%, så kan H_0 forkastes. Dvs. at mindst en af regressorvariablene har effekt på styrken.

Koefficienterne b_1 og b_2 har en p -værdi på henholdsvis $3.6695 \cdot 10^{-12}$ og 0.75295, som henholdsvis er mindre og større end 0.05. Nulhypotese, H_0 , siger, at koefficienten er 0, den alternative hypotese, H_a , siger, at koefficienten er forskellig fra 0, og med et signifikansniveau, α , på 5%, fås:

For b_1 er p -værdien på $3.6695 \cdot 10^{-12}$ dvs. mindre end 0.05, så H_0 forkastes. b_1 er signifikant forskellig fra 0.

OPGAVE 4 c fortsat

For b_2 er p -værdien på 0.75295 dvs. større end 0.05, så H_0 kan ikke forkastes, og b_2 er ikke signifikant forskellig fra 0. Det kan ikke afvises, at b_2 er 0.

Dette er i god overensstemmelse med de to scatterplot.

(Skæring med y -aksen $b_0 = 37.464$ har en p -værdi på $0.014153 < 0.05$ er derfor signifikant forskellig fra 0.)

- d. Udvidelse af modellen ved inddragelse af kvadratled og interaktionsled:

$$S = b_0 + b_1T + b_2p + b_{11}T^2 + b_{22}p^2 + b_{12}Tp$$

hvor $b_0, b_1, b_2, b_{11}, b_{22}$ og b_{12} er konstanter.

Der udføres en multipel lineær regression i MATLAB, hvor der anvendes Wilkinson notation:

```
mdl1 = fitlm([T, p], S, 'y ~ x1 + x2 + x1^2 + x2^2 + x1:x2 ')
```

Og der fås følgende output:

```
mdl1 =
```

```
Linear regression model:
```

```
y ~ 1 + x1*x2 + x1^2 + x2^2
```

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	23.082	53.183	0.43401	0.67007
x1	2.913	1.2198	2.3882	0.029607
x2	-1.4558	2.455	-0.59299	0.56148
x1:x2	0.028193	0.015829	1.7811	0.093885
x1^2	-0.0099556	0.006671	-1.4924	0.15506
x2^2	-0.017189	0.041697	-0.41224	0.68563

```
Number of observations: 22, Error degrees of freedom: 16
```

```
Root Mean Squared Error: 12.6
```

```
R-squared: 0.953, Adjusted R-Squared: 0.938
```

```
F-statistic vs. constant model: 64.3, p-value = 5.04e-10
```

$R^2 = 0.953$ og er en anelse øget fra 0.926

Adjusted $R^2 = 0.938$ er ligeledes øget en anelse fra 0.918.

Dvs. variationen forklares en anelse bedre.

p -værdier for koefficienterne er alle med undtagelse af b_1 for x_1 større end 0.05. Det mindst signifikante led fjernes. Det er x_2^2 , som har en p -værdi på 0.68563.

Der udføres en ny multipel lineær regression i MATLAB:

OPGAVE 4 d fortsat

```
mdl2 = fitlm([T, p], S, 'y ~ x1 + x2 + x1^2 + x1:x2 ')
```

Og der fås følgende output:

```
mdl2 =
```

```
Linear regression model:
```

```
y ~ 1 + x1*x2 + x1^2
```

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	32.061	47.319	0.67754	0.50718
x1	2.9509	1.1862	2.4876	0.023536
x2	-2.3605	1.073	-2.2	0.041926
x1:x2	0.028396	0.01543	1.8403	0.083248
x1^2	-0.010199	0.0064805	-1.5738	0.13396

```
Number of observations: 22, Error degrees of freedom: 17
```

```
Root Mean Squared Error: 12.3
```

```
R-squared: 0.952, Adjusted R-Squared: 0.941
```

```
F-statistic vs. constant model: 84.5, p-value = 5.51e-11
```

$R^2 = 0.952$ er stort set uændret.

Adjusted $R^2 = 0.941$ er øget en anelse mere.

Koefficienterne b_1 for x_1 og b_2 for x_2 er signifikant forskellige fra 0 med p -værdier mindre 0.05.

p -værdier for de øvrige koefficienter er større end 0.05.

Det mindst signifikante led fjernes. Det er x_1^2 , som har en p -værdi på 0.13396.

Der udføres en ny multipel lineær regression i MATLAB:

```
mdl3 = fitlm([T, p], S, 'y ~ x1 + x2 + x1:x2 ')
```

```
mdl3 =
```

```
Linear regression model:
```

```
y ~ 1 + x1*x2
```

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	95.338	25.955	3.6732	0.0017394
x1	1.1594	0.34721	3.3393	0.0036505
x2	-2.7413	1.0874	-2.5211	0.02135
x1:x2	0.037575	0.01486	2.5286	0.021017

```
Number of observations: 22, Error degrees of freedom: 18
```

```
Root Mean Squared Error: 12.8
```

```
R-squared: 0.945, Adjusted R-Squared: 0.936
```

```
F-statistic vs. constant model: 103, p-value = 1.55e-11
```

OPGAVE 4 d fortsat

$R^2 = 0.945$ er en anelse mindsket fra forrige model på 0.952, og er en anelse øget fra 0.926 på den oprindelige model i (4b)

Adjusted $R^2 = 0.936$ er en anelse mindsket fra forrige model på 0.941, og er en anelse øget fra 0.918 på den oprindelige model i (4b)

Dvs. variationen kan forklares en anelse bedre fra den oprindelige model i (4b).

Endvidere er alle led signifikant forskellige fra 0, da p -værdierne alle er mindre end 0.05.

Jeg vælger denne model.

- e. Opskrivning af ligningen for den foretrukne model: $S = b_0 + b_1T + b_2p + b_{12}Tp$ er:

$$S = 95.338 + 1.1594 \cdot T - 2.7413 \cdot p + 0.037575 \cdot T \cdot p$$

- f. Undersøgelse for "unormale" datapunkter, dvs. løftestangspunkter, outliers og indflydelsespunkter. Svaret begrundes.

Jeg laver analysen med den bedste model, mdl3.

For at undersøge for "unormale" punkter beregnes hat-diagonaler og studentiserede residualer, (rst), i MATLAB:

```
lev = mdl3.Diagnostics.Leverage;  
rst = mdl3.Residuals.Studentized;
```

Resultatet ses i tabellen næste side.

Løftestangspunkter (leverage) er "unormale" værdier i x -retningen, og det måles med hat-diagonalen. Grænsen beregnes vha. formlen:

$\text{lev}_{\text{limit}} = \frac{2(c+1)}{n}$, hvor c er antal regressorvariable, og n er antal observationer.

Der er 3 regressorvariable i den valgte model: Der er T , p og Tp

Dvs.

$$\text{lev}_{\text{limit}} = \frac{2(3+1)}{22} = 0.3636$$

I tabellen ses, at **punkterne nr. 2, nr. 12 og nr. 22 har hat-diagonal på henholdsvis: 0.3835, 0.4530 og 0.3644, alle større end 0.3636, så de er løftestangspunkter.**

Outliers er "unormale" værdier i y -retningen, og det måles på den numeriske værdi af det studentiserede residual, rst . Grænsen er: $|rst| > 3$.

I tabellen ses ingen punkter med $|rst| > 3$. **Dvs. der er ingen outliers.**

Et punkt skal være både et løftestangspunkt og outlier for at være et indflydelsespunkt. Der er derfor **ingen indflydelsespunkter.**

OPGAVE 4 f fortsat

Nr	Temperatur af støbeform T	Tryk i støbeproces p	Styrke af plast-komponent S	Hat diagonal Lev	Studentiseret residual rst
1	82	33	199	0.1152	-0.2121
2	103	34	247	0.3835	-0.5998
3	71	36	170	0.1180	-0.4069
4	43	13	124	0.2996	-0.6010
5	56	17	134	0.1215	-1.3110
6	62	27	173	0.0574	1.3934
7	62	26	153	0.0554	-0.2753
8	119	22	254	0.2233	-1.6049
9	64	20	151	0.0709	-0.9543
10	52	13	169	0.2175	2.3244
11	94	27	229	0.1041	0.2671
12	47	40	103	0.4530	-0.8178
13	50	29	132	0.1177	0.3000
14	80	36	202	0.1509	0.3619
15	74	39	183	0.1719	0.0278
16	48	16	125	0.1776	-0.9429
17	77	27	205	0.0506	1.3324
18	42	29	119	0.1755	0.7383
19	46	25	113	0.1078	-0.8488
20	99	21	255	0.1136	2.2285
21	108	14	233	0.3507	-0.5695
22	113	15	251	0.3644	0.1978