02402 Projekt 1 Sønderborg

Daniel Brasholt s214675

Oktober 2022

Kapitel 1: Beskrivelse

a)

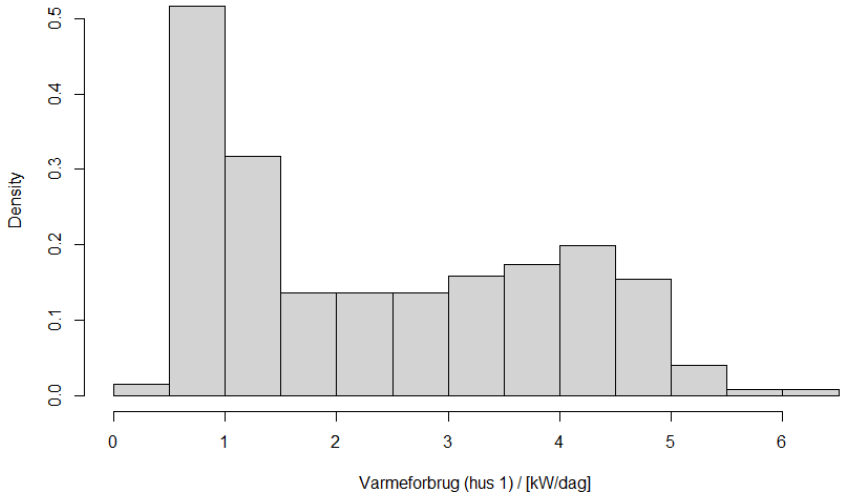
De variable, som indgår i datamaterialet, er t, som er tiden givet ved en dato af formatet $YYYY$-mm-dd; Ta, som er temperaturen udenfor til tiden givet i $°C$; G, som er den globale indstråling til målingstidspunktet givet i W/m^2; Ws, som er vindhastigheden givet i m/s; og til sidst $Q1$-$Q4$, som er varmeforbruget i de 4 huse givet i kW/dag. Alle disse variable er kvantitative pånær t, som er en dato-variabel. Der er således ingen kategoriske variable i datamaterialet.

Der er 973 observationer i alt. Med `head`- og `tail`-kommandoerne kan man se, at observationerne strækker sig fra d. 2. oktober 2008 til d. 1. juni 2011.

Ud af de 973 observationer er der dog mange manglende. Ta, G og Ws indeholder hver 61 manglende observationer. $Q1$ mangler 431, $Q2$ mangler 362, $Q3$ 511 og $Q4$ mangler 384.

b)

På nedenstående figur 1 kan man se fordelingen af varmeforbruget i hus 1 over tidsperioden, hvilken der er foretaget målinger. Figuren viser en assymetrisk, højreskæv fordeling. Alle observationer ligger på 0 eller derover. Der kan siges at være stor spredning i data, da sandsynlighedsdensiteten er næsten ens for observationer fra $1.5kW/dag$ til $5kW/dag$. Dog er der klart flest observationer fra $0.5kW/dag$ til $1.5kW/dag$ og ret få $< 0.5kW/dag$ og $> 5kW/dag$.



Figur 1: *Density histogram for varmeforbruget i hus 1.*

c)

På nedenstående figur 2 kan et tidsseriedram over den indsamlede data ses. De 4 huse er vist på samme diagram, hvilket afslører nogle ligheder i de tendenser, husenes varmeforbrug har.

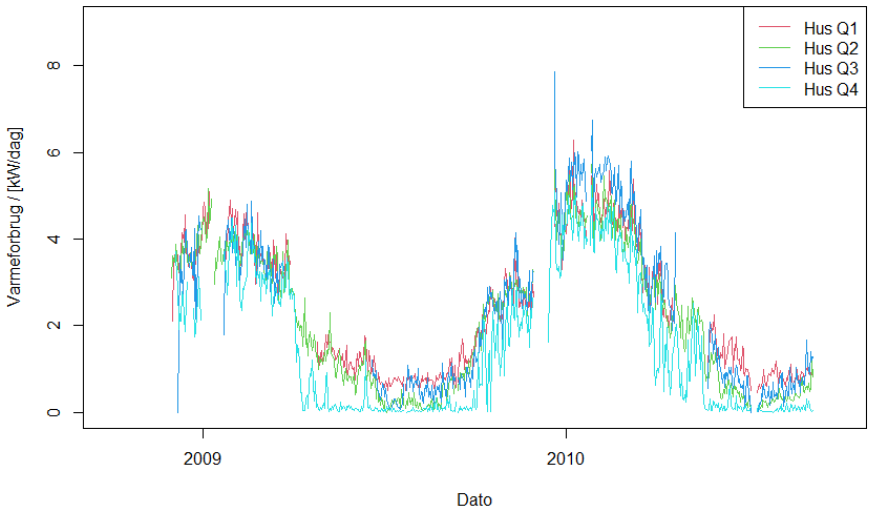
De 4 huse følger alle det, at de i slutningen af 2008 starter med et varmeforbrug mellem ca. 2 og 5. Dette falder dog markant for alle husene efter omkring første kvartal af 2009. Hus 4 falder mere end de andre huse, men de skærer alle markant ned.

Det begynder at stige igen okmring slutningen af år 2010, hvor alle 4 huses varmeforbrug vokser til mere, end det var i starten af observationerne. Dette varer igen nogle måneder, hvorefter varmeforbruget falder ligesom det gjorde efter et stykke tid i år 2009.

Ud fra dette kan man se, at fyringssæsonen ligger mellem slutningen af året og starten af andet kvartal, hvilket stemmer overens med virkeligheden.

På figur 2 ser det ud til, at varmeforbruget for hus 4 svinger mere fra dag til dag end for de 3 andre huse. Hus 3 har dog de mest ekstreme målinger på henholdsvis 0 midt i den periode, hvor der ellers var generel stigning i varmeforbruget (slutningen af 2008); og en måling omkring 8 i slutningen af 2009, hvor de andre huses varmeforbrug sjældent oversteg 6.

Figur 2 viser også 2 perioder, hvori der ikke blev udført observationer for nogen af de 4 huse; Omkring starten af 2009 og slutningen af 2009. Derudover er der, som det blev nævnt i a), manglende observationer indimellem for de 4 huse. For eksempel er der ingen observationer for hus 3 i den tid, der omtrent svarer til 2. kvartal af 2009.



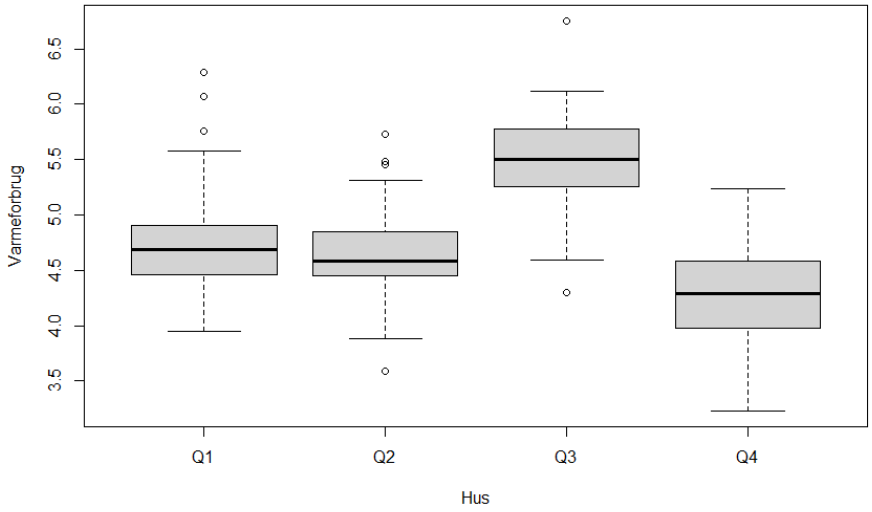
Figur 2: *Varmeforbruget i de 4 huse i perioden 2. oktober* 2008 *til 1. oktober* 2010.

d)

Nedenstående figur 3 viser 4 boxplot; hvert af disse viser varmeforbruget i ét hus i perioden januar til februar 2010.

Det kan ses på figuren, at husene 1-3 alle har outliers i deres varmeforbrug; dog er det kun 2-3 ud af de ca. 55 observationer (se nedenstående tabel 1, som knytter sig til sektion e)).

Det ses også, at varmeforbruget i hus 4 i perioden er tilnærmelsesvist symmetrisk. For hus 1 ser det ud til, at fordelingen er højreskæv. Hus 2 ser også ud til at være højreskæv, da de de fleste af observationerne ligger mellem 4 og 4.5. Hus 3 ser ud til at være tættere på at være normalfordelt end hus 1 og 2, men dog stadig lidt venstreskæv. Der er da forskel på fordelingen af de fire huses varmeforbrug i denne periode, hvilket udforskes yderligere i næste kapitel.



Figur 3: *Fordeling af varmeforbrug i de 4 huse i perioden januar-februar* 2010.

e)

Nedenstående tabel 1 viser, at antallet af observationer for de 4 huse er tæt på at være ens. Tabellen gør det også muligt at se på varians og spredning for de 4 huses varmeforbrug. Boxplottet på figur 3 tyder på, at hus 4 har den højestre spredning, da kvartilsættet ser ud til at ligge mere spredt; men ser man på tabel 1, kan man se, at varmeforbruget i dette hus faktisk har den laveste standardafvigelse. Hus 1 har den højeste standardafvigelse, hus 3 den næsthøjeste, hus 2 dernæst og hus 4 har den mindste.

Derudover kan man bruge tallene for stikprøvegennemsnittet og medianen for de 4 huse til at se, at forbruget i hus 4 er mest symmetrisk, hvilket boxplottet også viste. Gennemsnittet i hus 1 ligger til højre for medianen, hvorfor fordelingen nok er højreskæv. Ligeledes kan tallene for hus 2 og 3 anvendes til at se, at forbruget i hus 2 muligvis følger en højreskæv fordeling og hus 3 en venstreskæv fordeling. Stikprøvegennemsnittet i disse to huse er dog tættere på medianen end tallene for hus 1, hvorfor det ikke nødvendigvist siger lige så meget.

Hus	Antal obs.	Stikprøve-gennemsnit	Stikprøve-varians	Stikprøve-standard-afvigelse	Nedre kvartil	Median	Øvre kvartil
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
1	55	4.757108	0.2105385	0.4588448	4.456250	4.686806	4.916667
2	56	4.609263	0.1850795	0.4302087	4.452083	4.580903	4.852778
3	55	5.471087	0.1919077	0.4380728	5.241667	5.498611	5.777778
4	57	4.281017	0.174925	0.4182404	3.975000	4.289583	4.289583

Tabel 1: *Tabel med statistiske nøgletal for hver af de 4 huses varmeforbrug i perioden januar-februar* 2010.

Kapitel 2: Statistisk analyse

f)

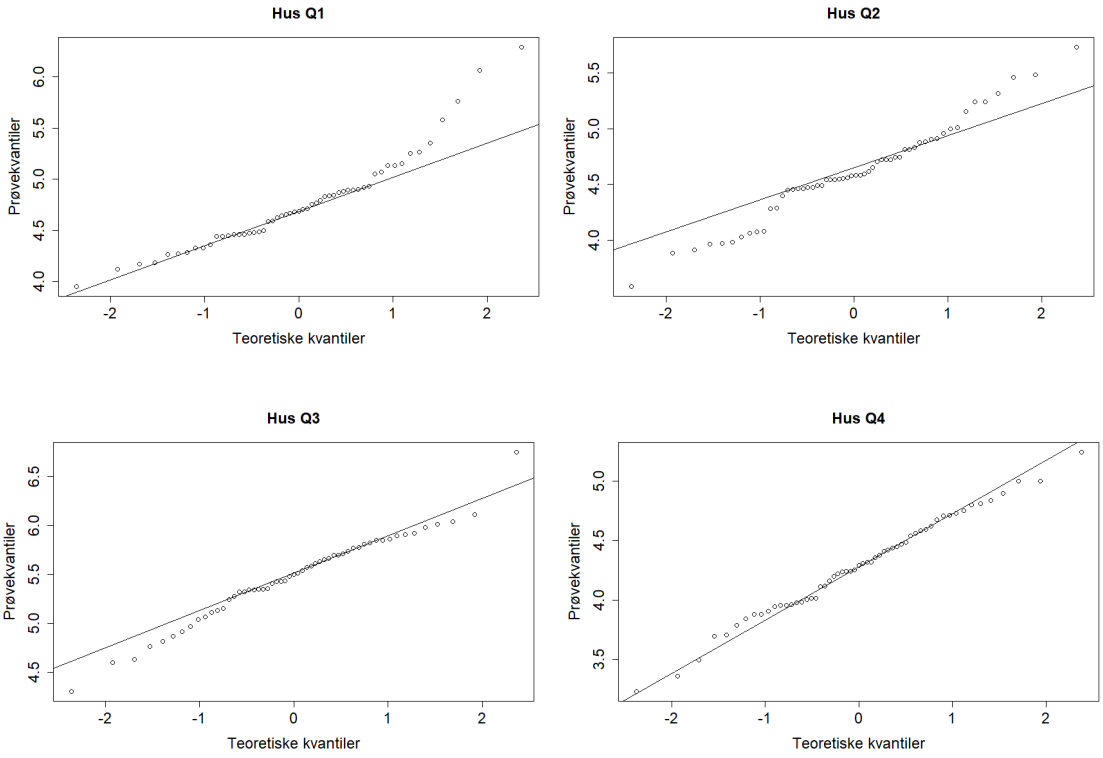
Figur 4 viser QQ-plot for observationerne af varmeforbruget i hver af de fire huse i den givne periode. Ligesom boxplottene på figur 3 og tabel 1 ses det, at forbruget i hus 4 bedst følger en normalfordeling. Hus 1 har nogle observationer, der pænt følger den rette linje, men afviger kraftigt i de øvre kvantiler. Hus 2 afviger ligeså i både de øvre og nedre kvantiler.

Observationerne på qq-plottet for hus 3 afviger også fra den forventede rette linje, og det ser ud til at observationerne afviger systematisk. Som nævnt tilnærmer observationerne for varmeforbruget i hus 4 bedst normalfordelingen. Dog kan det siges for observationerne for alle fire

huse, at observationerne ligger tilstrækkeligt tæt på en ret linje til, at tallene kan siges at være normalfordelte.

For at opstille modeller for de fire fordelinger, antages det, at observationerne er normalfordelte. Derudover antages det, at alle observationerne er uafhængige. Uden at kende måden, hvorpå forsøget blev udført, kan dette være svært at eftervise[1]. Ifølge den Centrale Grænseværdisætning (Theorem 3.14) vil middelværdien af stikprøven følge en normalfordeling, når n er tilstrækkeligt stor. Da n i alle tilfælde her er > 50, kan dette siges at være opfyldt[2]. Modeller for de fire huse under antagelse af normalfordeling kan da opstilles ved brug af de allerede fundne oplysninger i tabel 1:

$$\begin{aligned}
\text{Hus } 1 &: X_1 \sim N(4.76, 0.211) \\
\text{Hus } 2 &: X_2 \sim N(4.61, 0.185) \\
\text{Hus } 3 &: X_3 \sim N(5.47, 0.192) \\
\text{Hus } 4 &: X_4 \sim N(4.28, 0.175)
\end{aligned}$$



Figur 4: *QQ-plot for hvert af de 4 huse i perioden januar-februar* 2010.

g)

Ifølge bogens metode 3.9, kan 95%-konfidensintervallet for middelværdien beregnes således:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}} \tag{1}$$

$\bar{x}$, s og n er vist i tabel 1 for de fire huses varmeforbrug. Dette giver følgende konfidensinterval for det første hus:

[1] Det kunne for eksempel være, at observationer er afhængige ved, at et højt varmeforbrug én dag vil give et lavere varmeforbrug den næste, da huset allerede er varmt.

[2] I forelæsning til uge 4 blev det sagt, at n som tommelfingerregel blot skulle være over 30.

$$\begin{aligned} &4.7571088 \pm t_{0.975} \cdot \frac{0.4588448}{\sqrt{55}} \\ =&4.7571088 \pm 2.004879 \cdot \frac{0.4588448}{\sqrt{55}} \\ =&4.7571088 \pm 0.1240431 \\ =&\left[4.633065; 4.881151\right] \end{aligned}$$

Udføres samme beregning for de resterende huse fås følgende tabel over 95%-konfidensinterval for middelværdierne:

	Nedre grænse af KI	**Øvre grænse af KI**
Hus 1	4.633	4.881
Hus 2	4.494	4.724
Hus 3	5.353	5.590
Hus 4	4.170	4.392

Tabel 2: *Konfidensintervaller for middelværdier af varmeforbruget i hver af de 4 huse i perioden januar-februar* 2010.

Dette er netop det samme som fås med den i opgaven givne `R`-kommando (se bilag).

h)

For at tjekke, om middelværdien af det daglige varmeforbrug for hus 1 i januar-februar 2010 afviger signifikant fra 2.38, udføres en test ud fra følgende hypotese:

$$\begin{aligned} H_0 &: \ \mu_{Hus1} = 2.38 \\ H_1 &: \ \mu_{Hus1} \neq 2.38 \end{aligned}$$

Hypotesetesten udføres med et signifikansniveau $\alpha = 5\%$. p-værdien skal da gerne være under 0.05, hvis nulhypotesen skal forkastes. Formlen for udregning af teststørrelsen af således:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{2}$$

p-værdien kan da udregnes med formlen:

$$p = 2 \cdot P(T > |t_{obs}|) \tag{3}$$

I denne følger T en t-fordeling med $(n-1) = 54$ frihedsgrader (se tabel 1). Indsættes tal i (2) og (3) fås:

$$\begin{aligned} t_{obs} &= 38.421 \\ p &= 0 \end{aligned}$$

Udføres testen med det indbyggede værktøj i `R`, fås $p < 2.2 \cdot 10^{-16}$.

Det er da meget usandsynligt, at H_0 er sand, hvorfor denne kan forkastes. Middelværdien af det daglige varmeforbrug i januar-februar er større end 2.38, hvilket kan ses i tabel 1. Alt dette kunne man dog have fundet ud af uden at udføre hypotesetesten, da 2.38 ikke ligger indenfor konfidensintervallet for middelværdien for hus 1, hvilket kan ses i tabel 2.

i)

For at undersøge, om der kan påvises en forskel mellem middelværdien af varmeforbruget for hus 1 og hus 2 i perioden januar-februar 2010, laves en two-sample t-test. Her anvendes igen signifikansniveauet $\alpha = 0.05$. Her betragter vi forskellen i middelværdi, hvilket giver anledning til nulhypotesen:

$$\delta = \mu_2 - \mu_1$$
$$H_0 : \ \delta = \delta_0 = 0$$

Vi antager da, at der ikke er forskel mellem middelværdierne. Formlen for teststørrelsen ser således ud:

$$t_{obs} = \frac{(\bar{x_1} - \bar{x_2}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{4}$$

Denne følger approksimativt en t-fordeling med v frihedsgrader, hvor

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1} +} \tag{5}$$

Udtrykket for p-værdien er det samme som i (3). Følgende værdier kan da fås (se bilag):

$$t_{obs} = 1.751$$
$$v = 108.3$$
$$p = 0.0829$$

Da $p > 0.05$ kan nulhypotesen ikke forkastes, hvorfor der ikke kan siges at være forskel i middelværdierne for varmeforbruget i hus 1 og hus 2. Vi kan da heller ikke sige, om forbruget er højere i det ene hus end i det andet.

j)

Det ses i tabel 2, at konfidensintervallerne for middelværdierne for hus 1 og hus 2 overlapper. Fra bemærkning `3.39` ses det dog, at dette ikke i sig selv er tilstrækkeligt til at sige, at varmeforbruget er ens; kun hvis konfidensintervallerne *ikke* overlapper, kan konklusionen om signifikant forskel drages. Når intervallerne overlapper som her, er man nødt til at foretage en hypotestest.

k)

Formlen for korrelation er givet således (`1.19`):

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y} \tag{6}$$

I (6) er s_x standardafvigelsen for x, s_y standardafvigelsen for y og s_{xy} kovariansen mellem x og y. Disse tre værdier er fundet til:

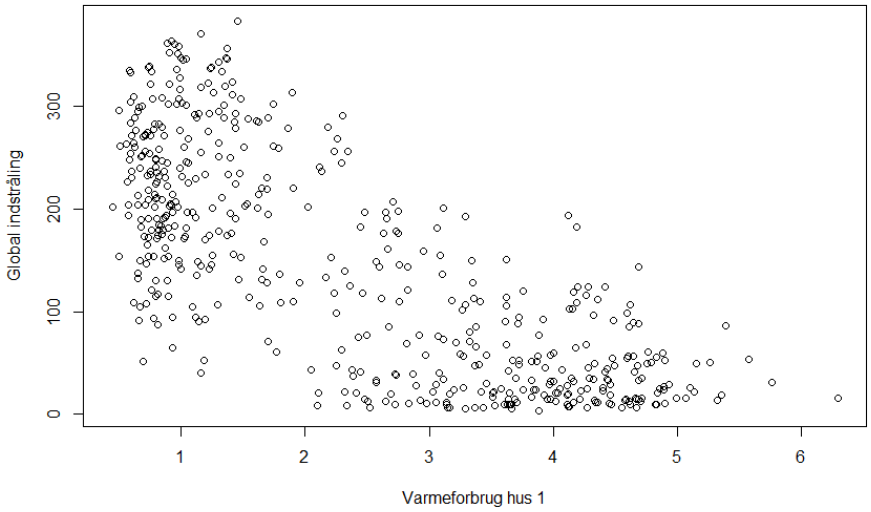
$$s_{xy} = -118.395$$
$$s_x = 1.490$$
$$s_y = 103.365$$

Indsættes disse værdier i (6) fås:

$$r = \frac{-118.395}{1.490 \cdot 103.365} = -0.769$$

Anvendes den i projektbeskrivelsen givne kommando, fås i stedet -0.754. Der er da forskel i måden, hvorpå korrelationen bliver udregnet med den indbyggede kommando og ved at bruge `cov` og `sd` (se bilag). Forskellen er dog ikke så stor, at den betyder noget i denne sammenhæng.

Da kovariansen er negativ, må der være en invers tendens mellem de to observationssæt. En høj værdi af den ene vil da give en lav værdi af den anden. Dette er også forventet, da et hus nok vil bruge mere energi på at varme, når solen ikke varmer huset. Denne tendens og sammenhæng kan også ses på nedenstående figur 5 - jo højere varmeforbruget er, des lavere er den globale indstråling.



Figur 5: *Scatterplot, der viser sammenhængen mellem varmeforbruget i hus 1 og den globale indstråling i hele perioden, hvori målinger er taget.*