Projekt 2: Varmeforbrug i Sønderborg II

02402 Introduktion til Statistik

Daniel Brasholt s214675

8. november 2022

Indhold

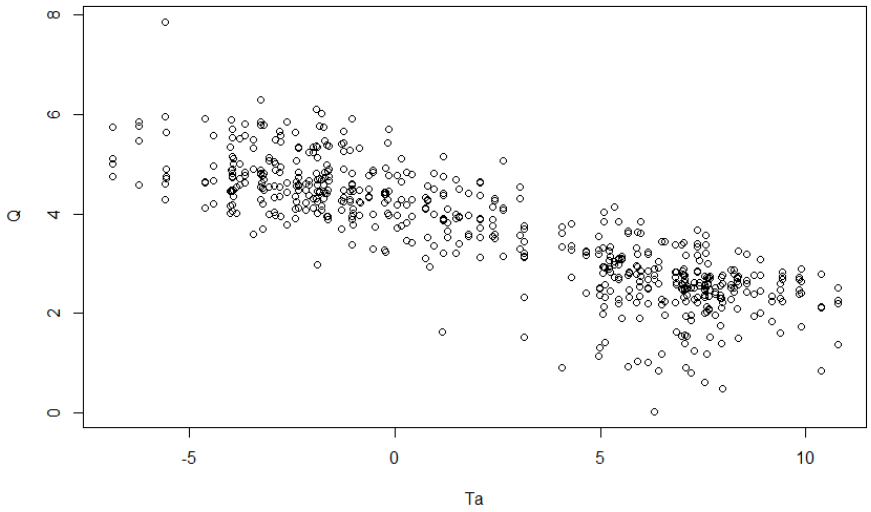
A)

	Antal obs.	Gennemsnit	Std. afvigelse	25%-fraktil	Median	75%-fraktil
Q	576	3.61	1.20	2.64	3.69	4.55
Ta	576	2.39	4.65	−1.85	2.08	7.00
G	576	68.00	61.77	18.59	48.23	109.05

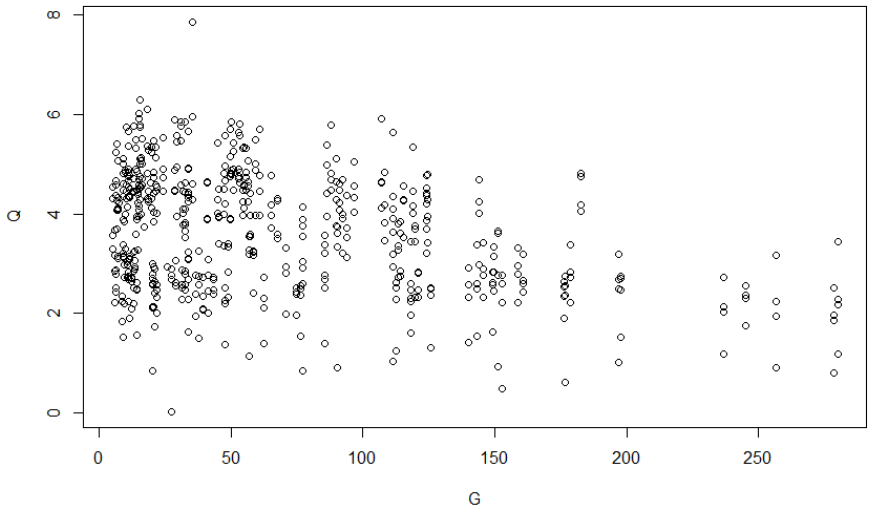
Tabel 1: *Tabel over opsummering af parametre for de 3 variable*

Ovenstående tabel 1 viser, at variablene `Ta` og `G` har en relativt høj standardafvigelse sammenlignet med `Q`. Nedenstående figurer 2 viser, at Q er tættest på at være normalfordelt. Ta ser ud til at være fordelt med to kurver; en lige under 0° og en omkring 6°. G er en højreskæv fordeling. Derudover ser der ud til at være stærkest sammenhæng mellem Q og Ta (i forhold til Q og G), hvilket fremgår af 1. Der er rigeligt antal observationer til at lave en analyse med 576 observationssæt. Til sidst kan det ses på boxplottene, at Q og G har ekstreme målinger - begge udelukkende høje, ekstreme målinger. Ta har ingen ekstreme målinger.

God introduktion og forsmag på hvilke sammenhænge vi vil prøve at afspejle i vores model senere.

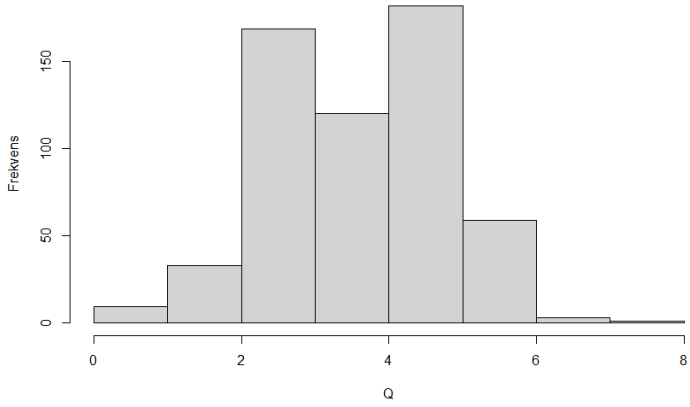


(a) *Ta*

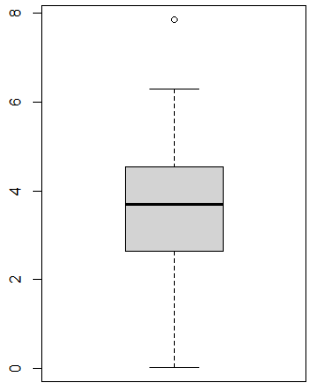


(b) *G*

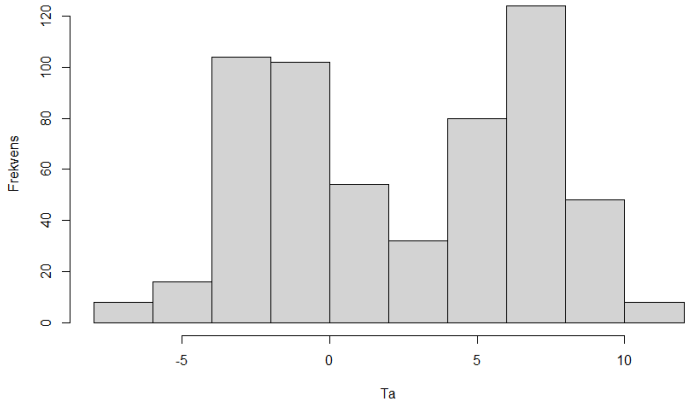
Figur 1: *Scatterplot af varmeforbrug mod 1a udendørs temperatur og 1b solstråling*



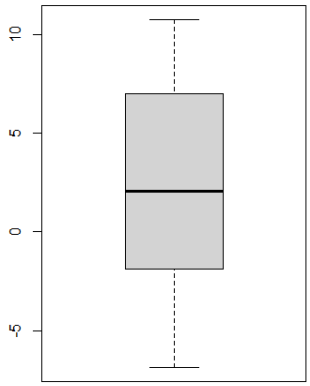
(a) Histogram over `Q`



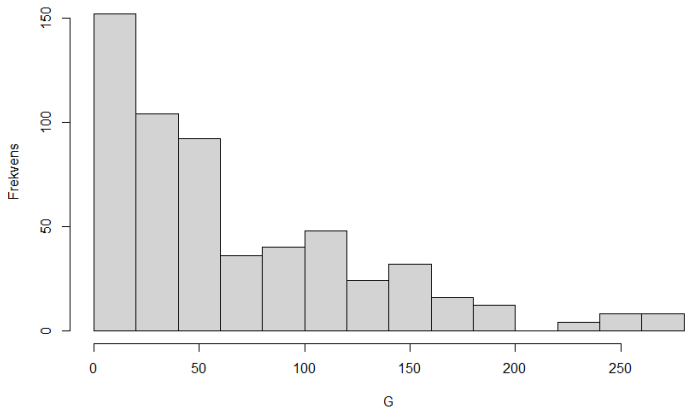
(b) Boxplot over `Q`



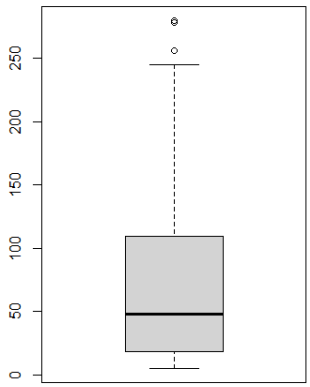
(c) Histogram over `Ta`



(d) Boxplot over `Ta`



(e) Histogram over `G`



(f) Boxplot over `G`

Figur 2: *Histogram over `Q`, `Ta` og `G`*

B)

En model med varmeforbruget som responsvariabel (Y_i) og med udendørstemperatur og solens indstråling som forklarende variable (henholdsvis $x_{1,i}$ og $x_{2,i}$) vil se således ud:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.} \tag{1}$$

Det vigtige er variansen er den samme for alle residualerne.

Når denne laves, antages der dog, at residualerne er uafhængige og identisk distribuerede normale variable med en middelværdi på 0 og en ukendt varians σ^2. Det antages altså, at udendørstemperaturen og solens indstråling er uafhængige af hinanden, hvilket ikke nødvendigvist er sandt. Derudover antages det, at usikkerheden er normalfordelt, hvilket heller ikke nødvendigvist gør sig gældende.

Ikke helt. Der antages ikke at de to forklarende variable er uafhængige. De må gerne korrelere til en vis grad. Hvis der er for meget kolinearitet, så er parameteren mere eller mindre den samme, og så er det overflødigt at have to, men det er det eneste.

C)

Med `R` fås følgende værdier som estimater for parametrene givet i modellen vist i (1):

$$\begin{aligned} \hat{\beta}_0 &= 4.2788 \, , \quad \sigma_{\beta_0} = 0.039 \\ \hat{\beta}_1 &= -0.2090 \, , \quad \sigma_{\beta_1} = 0.0058 \\ \hat{\beta}_2 &= -0.0025 \, , \quad \sigma_{\beta_2} = 0.0004 \\ \hat{\sigma}^2 &= 0.621 \end{aligned} \tag{2}$$

Det er standardafvigelsen!

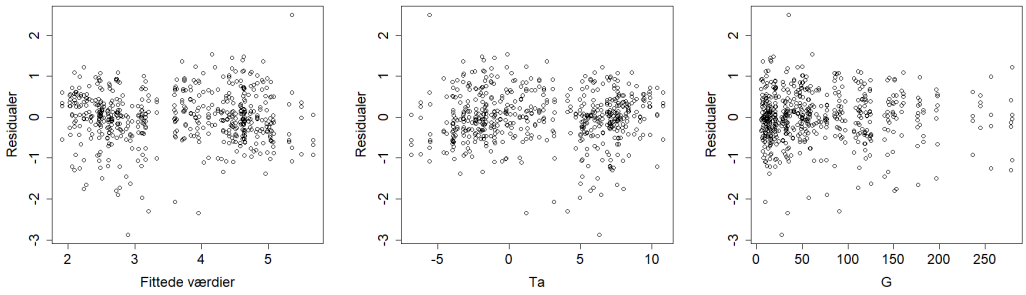
Antallet af frihedsgrader brugt til estimatet af residualernes varians er 573.

Modellens forklarede varians er fundet til $R^2 = 0.7325$.

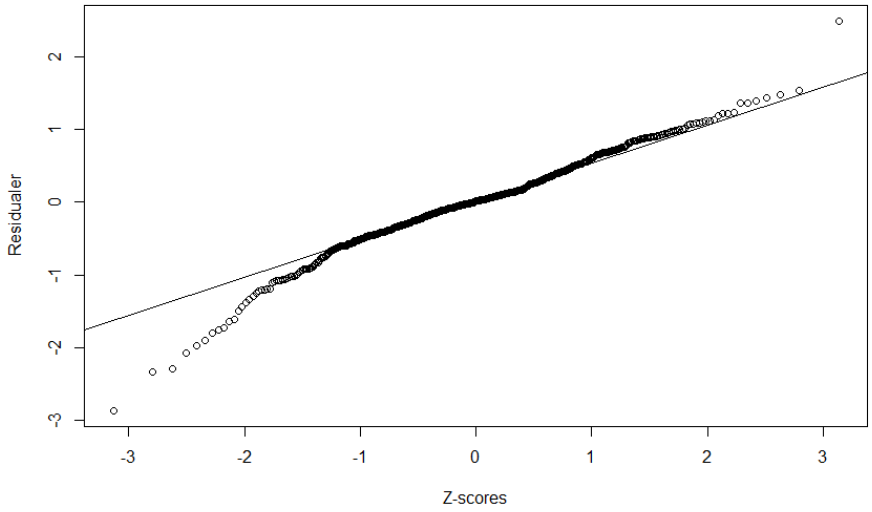
Da β_0 er positiv og β_1 samt β_2 negative, kan det siges, at Ta og G påvirker Q negativt; dette giver også mening, da man vil forvente et lavere varmeforbrug de dage, hvor der er varmere udenfor eller solen bidrager mere til varmen.

Du må gerne fortolke lidt mere konkret på parameterestimaterne. Hvor meget stiger varmeforbruget pr. grad udetemperaturen stiger, og hvor stor en forøgelse i solindstråling svarer til en 1 grads forøgelse i udetemperatur? Virker det rimeligt? Også gerne denne slags ting, da tallene svarer til fysiske størrelse og det er sådan den endelige model vil komme til at fungere.

D)



Figur 3: *Residualer mod hver af de forklarende variable samt mod fittede værdier*



Figur 4: *QQ-plot af residualerne*

Figur 3 viser, at residualerne for de to indgående variable samt for de fittede værdier ikke ser ud til at have en sammenhæng og da er tilfældige. Dette forstærker gyldigheden af antagelserne.

Figur 4 viser, at residualerne ikke helt ser ud til at ligge på en ret linje som man ville forvente, hvis residualerne var normalfordelte. Residualerne ser ud til at afvige fra linjen systematisk, hvilket muligvis betyder, at modellen er opstillet forkert. Dog er det relativt få observationer (sammenlignet med det antal observationer, der ligger på linjen), der afviger fra den forventede rette linje, hvorfor afvigelsen ikke nødvendigvist er systematisk. Derfor vurderes forudsætningerne for modellen opfyldte.

Jeg savner et plot af de varmeforbruget mod de fittede værdier, og så savner jeg også at du tage stilling til residualantagelserne individuelt. Du undersøger kun om de er uafhængige af hinanden, men ikke om de har middelværdi 0 og konstant varians.

E)

Formlen for at bestemme $(1-\alpha/2)$konfidensintervallet for en parameter (Method 6.5) er givet ved:

$$\hat{\beta}_i \pm t_{1-\alpha/2}\hat{\sigma}_{\beta_i} \tag{3}$$

I dette tilfælde bestemmer vi den for koefficienten for udendørstemperaturen, β_1. Med `R` fås værdierne:

$$\begin{aligned}\hat{\beta}_i &= -0.2089526\\ \hat{\sigma}_{\beta_1} &= 0.0058260\end{aligned}$$

t-fordelingen anvendes med 573 frihedsgrader som givet i del c). Dette giver:

$$\begin{aligned}&-0.2089526 \pm 1.964113 \cdot 0.0058260\\ &= [-0.2203955\ ;\ -0.1975097]\\ &\approx [-0.220\ ;\ -0.200]\end{aligned}$$

Med `R` (se bilag), fås samme værdier med tilstrækkelig præcision. Med `R` er de resterende parametrene bestemt til:

	2.5%	97.5%
β_0	4.202	4.355
β_1	−0.220	−0.200
β_2	−0.0033	−0.0016

Tabel 2: *Tabel over konfidensintervaller for de estimerede parametre*

F)

Vi er interesserede i, om β_1 kunne have værdien -0.25. Nulhypotesen og den alternative hypotese ser da således ud:

$$H_{0,1} : \beta_1 = -0.25$$
$$H_{1,i} : \beta_1 \neq -0.25$$

Teststørrelsen er givet ved formlen:

$$t_{obs,\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$$

Her er $\beta_{0,1} = -0.25$, $\sigma_{\beta_1} = 0.0058260$ og $\beta_1 = -0.2090$. Dette indsættes i formlen:

$$t_{obs,\beta_1} = \frac{-0.2090 - (-0.25)}{0.0058260} = 7.037418$$

Dernæst skal p-værdien findes ved formlen:

$$p - \text{value} = 2 \cdot P(T > |t_{obs,\beta_i}|)$$

I dette tilfælde er der igen 573 frihedsgrader. Da fås ovenstående med `R` til:

$$p \approx 5.61 \cdot 10^{-12} < \alpha = 0.05$$

Da p-værdien er tilstrækkeligt lav, kan vi forkaste nulhypotesen. Da kunne β_1 ikke have haft værdien -0.25. Så skulle vi have været meget uheldige.

G)

Det er ikke valgt at reducere modellen, da begge parametre er statistik signifikante. Derudover viste et forsøg på at fjerne G fra modellen (den med højest p-værdi), at R^2-værdien kun blev mindre af at fjerne en parameter. Derudover blev fejlen på residualerne også større, hvilket fremgår af `R`-koden. Derfor er den endelige model sat til at være:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.} \tag{4}$$

ligesom i (1). Estimaterne for modellens parametre er da igen som i (2).

H)

Med slutmodellen vist i (1) og (2) fås følgende prædikterede værdier (udregnet i `R`, se bilag):

	id	Q	fit	lwr	upr
66	3	3.53	3.41	2.19	4.63
1117	5	4.09	3.85	2.63	5.07
5027	10	3.13	3.07	1.84	4.29
11858	17	2.60	2.76	1.54	3.98

Rigtig god rapport! Jeg savnede kun at du tog stilling til alle antagelser om residualerne i modelkontrollen. Ellers rigtig godt.

Tabel 3: *Tabel over prædikterede- og egentlige værdier samt prædiktionsintervaller*

Det kan ses fra tabel 3, at de egentlige værdier ligger tæt på de med modellen prædikterede værdier. Værdierne ligger alle indenfor 95%-prædiktionsintervallet - også pænt indenfor intervallerne og ikke i nærheden af grænserne. Det ser da ud til, at modellen godt kan prædiktere varmeforbruget i et hus ud fra den på det pågældende tidspunkt givne udendørstemperatur og solindstråling.