Projekt 1: Varmeforbrug i Sønderborg

Formaliteter, opgavestruktur og forventninger til 1. obligatoriske opgave

Opgaven består af to dele. I første del skal der laves en deskriptiv analyse af data. Anden del handler primært om konfidensintervaller og hypotesetests.

Der er lagt op til, at man skal arbejde med opgaven i små "lette" trin. Opgaven skal i praksis løses ved hjælp af programmet R. Der er udarbejdet R-kode, som gør det nemt at komme i gang med projektet. Koden er dog ikke fuldstændig, og I opfordres til at udforske R samtidig med at I laver projektet. F.eks. kan I arbejde med at lave "pæne" titler til graferne eller benytte R's indbyggede funktioner til beregning af konfidensintervaller og test af hypoteser.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, matematisk notation, samt tekst der beskriver analysens resultater. Relevante grafer og tabeller skal indgå i sammenhæng med teksten - ikke som bilag. Præsenter resultaterne fra jeres analyser på samme måde, som I ville videreformidle dem til andre fagfæller.

Inddel besvarelsen i et underafsnit for hver af de stillede spørgsmål.

Besvarelsen skal afleveres som pdf-fil. R-kode bør ikke indgå i besvarelsen, men vedlægges som bilag (i form af en .R-fil). Besvarelsen samt bilag afleveres under Opgaver på Learn ved: `Projekt 1:Varmeforbrug i Sønderborg`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En normal side udgør 2400 anslag.

Grafer og tabeller kan IKKE stå alene - det er altså vigtigt, at I beskriver og fortolker

outputtet fra R med ord.

Grafer og tabeller indgår ikke i opgørelsen af besvarelsens længde. Det er dog IKKE i sig selv en fordel at medtage mange plots, hvis de ikke er relevante!

I må gerne arbejde sammen i grupper, men besvarelsen af opgaven skal skrives individuelt. Spørgsmål omkring projektet kan rettes til hjælpelæren, se retningslinjerne på siden `Projects` på kursets hjemmeside.

Problemstilling

I dette projekt kigges der på varmeforbruget for fire huse i Sønderborg i perioden oktober 2008 til juni 2011. Ved at sammenholde de klimatiske forhold med varmeforbruget er det blandt andet muligt at undersøge husenes faktiske isoleringsgrad. Ved at sammenholde med vindhastigheder kan man se, om nogle huse er mere utætte/påvirkede end andre.

Ideelt kan man, baseret på denne type data, lave en empirisk energimærkning af huse. Så længe beboerne ikke ændrer deres adfærd for meget i løbet af perioden vil analysen være uafhængig af indendørstemperaturen. Husene er fritstående enfamilieshuse.

Indlæsning af data

Lav en mappe til projektet på din computer. Download materialet til projektet fra Learn og udpak ("unzip") det til den mappe, du lige har lavet.

Åbn derefter data-filen `soenderborg1_data.csv` (f.eks. i RStudio, File → Open File) for at se filens indhold. Bemærk at første linje indeholder variabelnavne (kaldes en *header*), og at de efterfølgende linjer indeholder de egentlige observationer. Observationerne af de enkelte variable er adskilt af et ';' (deraf `.csv`: "comma separated values", her dog semikolon).

Observationerne i datasættet består af daglige værdier af varmeforbruget i fire huse samt daglige gennemsnit af centralt observerede klimatiske variable. I data-filen er der følgende søjler/variable:

Variabel	Forklaring
t	Dato
Ta	Udendørstemperatur (°C)
G	Global indstråling (W/m^2)
Ws	Vindhastighed (m/s)
Q1	Varmeforbrug i hus 1 (kW/dag)
Q2	Varmeforbrug i hus 2 (kW/dag)
Q3	Varmeforbrug i hus 3 (kW/dag)
Q4	Varmeforbrug i hus 4 (kW/dag)

Åbn filen `soenderborg1_dansk.R`, som indeholder R-kode der kan bruges til analysen. Først skal "working directory" sættes til den mappe på computeren, hvor filerne til projektet er gemt:

```
## I RStudio kan man nemt sætte working directory med menuen
## "Session -> Set Working Directory -> To Source File Location"
## Bemærk: i R bruges kun "/" til separering i stier
## (altså ingen backslash).
setwd("Erstat her med stien til den mappe, hvor projektfilerne er gemt.")
```

Nu kan datasættet indlæses i R med følgende kode:

```
## Indlæs data fra soenderborg1_data.csv
D <- read.table("soenderborg1_data.csv", header=TRUE, sep=";",
                as.is=TRUE)
```

Bemærk, at der i R-koden bruges engelsk. Det er generelt ikke en god ide at bruge æøå osv. ved programmering. `D` bliver en "data.frame" (en slags tabel), som indeholder den indlæste data (se R-introen i kapitel 1.5 i bogen).

Beskrivende analyse (descriptive analysis)

Første del af projektet går ud på at lave en beskrivende analyse af data. I en rapport er det vigtigt at præsentere og beskrive data for læseren. Dette kan f.eks. gøres ved hjælp af opsummerende størrelser/nøgletal ("summary statistics") og passende figurer.

En simpel opsummering af det indlæste datasæt fås ved at køre følgende kode:

```
## Dimensionen af D (antallet af rækker og søjler)
dim(D)
## Søjle-/variabelnavne
names(D)
## De første rækker/observationer
head(D)
## De sidste rækker/observationer
tail(D)
## Udvalgte opsummerende størrelser
summary(D)
## En anden type opsummering af datasættet
str(D)
```

a) Lav en kort beskrivelse af datamaterialet: Hvilke variable indgår i datasættet? Er der tale om *kvantitative* og/eller *kategoriserede* variable (eller dato-variable)? (Kategoriserede variable dukker først op i kapitel 8, men det er bare variable, som inddeler observationerne i kategorier - f.eks. tre kategorier: lav, mellem og høj). Hvor mange observationer er der? Hvilken periode dækker observationerne over (hvornår er første hhv. sidste observation foretaget)? Er der manglende værdier for nogen af variablene?

Et "density histogram" der beskriver den empiriske tæthed af observationerne af det daglige varmeforbrug i hus 1 (se kapitel 1.6.1) kan laves ved hjælp af følgende kode:

```
## Histogram der beskriver den empiriske tæthed for obs.
## af varmeforbruget for hus 1 (histogram for de daglige
## målinger normaliseret så arealet er lig 1)
hist(D$Q1, xlab="Varmeforbrug (hus 1)", prob=TRUE)
```

b) Lav et density histogram for observationerne af varmeforbruget for hus 1. Beskriv fordelingen af observationerne ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan varmeforbruget være negativ? Er der stor spredning i observationerne?

Bemærk: I en *skæv* fordeling er sandsynlighedsmassen ikke symmetrisk fordelt omkring medianen. For en venstreskæv fordeling gælder der, at den længste hale ligger til venstre for midten (almindeligvis vil gennemsnittet også ligge til venstre for medianen). Tilsvarende gælder der, at for en højreskæv fordeling ligger den længste hale til

højre for midten (almindeligvis med gennemsnit til højre for medianen).

Ved observationer foretaget regelmæssigt over tid omtales data ofte som en *tidsrække*. Data for varmeforbruget i hvert hus udgør således en tidsrække. For tidsrækker er det ofte relevant at lave grafer, der viser udviklingen over tid. Her er det først nødvendigt at fortælle R, at variablen `t` skal opfattes som en dato-variabel. Dette gøres med følgende kode:

```
## Konverterer variablen 't' til en dato-variabel i R
D$t <- as.Date(x=D$t, format="%Y-%m-%d")
## Tjekker resultatet
summary(D$t)
```

Et plot der viser varmeforbruget over tid for perioden 2. oktober 2008 til 1. oktober 2010 for hvert hus (og farvet efter hus) kan nu laves med følgende R-kode:

```
## Plot varmeforbruget over tid
plot(D$t, D$Q1, type="l", xlim=as.Date(c("2008-10-02","2010-10-01")),
     ylim=c(0,9), xlab="Dato", ylab="Varmeforbrug", col=2)
lines(D$t, D$Q2, col=3)
lines(D$t, D$Q3, col=4)
lines(D$t, D$Q4, col=5)
## Tilføj legend
legend("topright", legend=paste0("Q",c(1,2,3,4)), lty=1, col=2:5)
```

Det bemærkes, at der er manglende observationer - vist som huller i tidsrækken.

c) Lav et plot der illustrerer varmeforbruget over tid for perioden 2. oktober 2008 til 1. oktober 2010 for hvert hus (og farvet efter hus). Beskriv udviklingen i varmeforbruget over tid med ord. Udvikler varmeforbruget sig på samme måde for de fire huse? Er varmeforbruget stabilt? Kan man se hvornår der er fyringssæson? Er der nogen steder, hvor varmeforbruget opfører sig uventet?

En meget anvendelig operation i dataanalyse er at opdele data i delmængder. Funktionen `subset` kan bruges til at udtage en delmængde, se også bemærkningen på side 10.

I den videre analyse er det kun data fra januar-februar 2010, der skal benyttes. Brug den følgende R-kode til at lave et nyt deldatasæt, der kun indeholder observationerne fra disse måneder.

```
## Udvælg data fra januar og februar 2010
Dsel <- subset(D, "2010-01-01" <= t & t < "2010-3-01")
```

Følgende R-kode laver et boxplot for varmeforbruget i januar-februar 2010 opdelt efter hus:

```
## Boxplot for varmeforbrug opdelt efter hus
boxplot(Dsel[ ,c("Q1","Q2","Q3","Q4")],
        xlab="Hus", ylab="Varmeforbrug")
```

d) Lav et boxplot for varmeforbruget i januar-februar 2010 opdelt efter hus. Benyt derefter plottet til at beskrive den observerede fordeling af varmeforbruget i de fire huse. Er fordelingerne symmetriske eller skæve? Ser det umiddelbart ud til, at der er forskelle mellem fordelingerne (hvis ja, hvilke)? Er der ekstreme observationer/outliers?

Man kan også beskrive den empiriske fordeling af hvert af husenes varmeforbrug i januar-februar 2010 ved hjælp af opsummerende størrelser/nøgletal som i følgende tabel:

Hus	**Antal obs.**	**Stikprøve-gennem-snit**	**Stikprøve-varians**	**Stikprøve-standard-afvigelse**	**Nedre kvartil**	**Median**	**Øvre kvartil**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
Hus 1							
Hus 2							
Hus 3							
Hus 4							

For at udfylde de tomme celler i tabellen kan man f.eks. benytte R-kode som følgende (se også bemærkning på side 11 for tricks til udregningerne):

```
## Antal observationer af dagligt varmeforbrug for hus 1 i jan-feb 2010
## (medregner ej eventuelle manglende værdier)
sum(!is.na(Dsel$Q1))
## Stikprøvegennemsnit af dagligt varmeforbrug for hus 1 i jan-feb 2010
mean(Dsel$Q1, na.rm=TRUE)
## Stikprøvevarians af dagligt varmeforbrug for hus 1 i jan-feb 2010
var(Dsel$Q1, na.rm=TRUE)
## osv.
## Argumentet 'na.rm=TRUE' sørger for at størrelsen
## udregnes selvom der eventuelt er manglende værdier
```

e) Udfyld tabellen ovenfor med de opsummerende størrelser for varmeforbruget i hvert af de fire huse i januar-februar 2010. Beskriv hvilken ekstra information kan udledes fra tabellen sammenlignet med boxplottet?

Statistisk analyse

Andel del af projektet går ud på at lave en simpel statistisk analyse vedrørende husenes daglige varmeforbrug. Der skal opstilles statistiske modeller for varmeforbruget. Modellernes parametre skal estimeres, og der skal udføres hypotesetests og beregnes konfidensintervaller.

Konfidensintervaller og hypotesetests

Følgende R-kode kan benyttes til at lave et qq-plot med henblik på at vurdere, om det daglige varmeforbrug i hus 1 (i den udvalgte periode) kan antages at være normalfordelt:

```
## qq-plot for varmeforbruget i hus 1
qqnorm(Dsel$Q1)
qqline(Dsel$Q1)
```

f) Opskriv separate statistiske modeller for varmeforbruget i hvert af de fire huse (se bemærkning 3.2). Estimer parametrene i de fire modeller (middelværdi og standardafvigelse). Foretag modelkontrol af de antagede forudsætninger (se kapitel 3 samt afsnit 3.1.8 i bogen). Idet konfidensintervaller og hypotesetests her involverer fordelingen af gennemsnit kan det være nyttigt også at inddrage den centrale grænseværdisætning (sætning 3.14) i argumentationen.

I praksis vil der opstå situationer, hvor man på baggrund af f.eks. modelkontrollen *ikke* kan tillade sig at antage, at en models forudsætninger er opfyldte. Da vil man ofte overveje, om det kunne hjælpe at foretage en transformation af data (se kapitel 3.1.9 i bogen). Bemærk at efter en transformation ændres fortolkningen af resultaterne på den oprindelige skala. Det er *ikke* meningen, at I skal lave en transformation af data i dette projekt.

g) Angiv formlen for et 95% konfidensinterval (KI) for middelværdien af det daglige varmeforbrug i januar-februar 2010 for hus 1 (se sektion 3.1.2 i bogen). Indsæt tal og beregn intervallet. Beregn tilsvarende konfidensintervaller for de tre andre huse og udfyld tabellen nedenfor.

	Nedre grænse af KI	**Øvre grænse af KI**
Hus 1		
Hus 2		
Hus 3		
Hus 4		

Sammenlign det beregnede konfidensinterval for hus 1 med resultatet af følgende R-kode:

```
## Konfidensinterval for middelværdi af dagligt varmeforbrug i hus 1
t.test(Dsel$Q1, conf.level=0.95)$conf.int
```

Man har vurderet, at det gennemsnitlige daglige varmeforbrug for hus 1 over et helt år er omkring 2.38 kW/dag.

h) Udfør et hypotesetest med henblik på at undersøge, om middelværdien af det daglige varmeforbrug for hus 1 i januar-februar 2010 afviger signifikant fra 2.38 kW/dag. Dette kan gøres ved at teste følgende hypotese:

$$H_0 : \mu_{\text{Hus1}} = 2.38,$$
$$H_1 : \mu_{\text{Hus1}} \neq 2.38.$$

Angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Hvis der er en signifikant afvigelse, er middelværdien af det daglige varmeforbrug i januar-februar så større eller mindre end 2.38 kW/dag? Kommenter på om det var nødvendigt at udføre det statistiske test, eller om samme konklusion kunne opnås ved konfidensintervallet alene.

Sammenlign resultaterne for test af hypotesen med resultaterne af følgende R-kode:

```
## Test af hypotesen mu=2.38 for dagligt varmeforbrug i hus 1
t.test(Dsel$Q1, mu=2.38)
```

Vi ønsker nu også at undersøge, om der er forskel på middelværdien af det daglige varmeforbrug for hus 1 og hus 2 i januar-februar 2010.

i) Undersøg ved et hypotesetest, om der kan påvises en forskel mellem middelværdien af varmeforbruget for hus 1 og hus 2 i de første to måneder af 2010. Opskriv hypotesen og angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Er varmeforbruget i hus 1 signifikant forskelligt fra hus 2? Hvis ja, er det højere eller lavere?

Sammenlign resultaterne for test af hypotesen med resultaterne af følgende R-kode:

```
## Sammenligning af varmeforbrug i hus 1 og 2
t.test(Dsel$Q1, Dsel$Q2)
```

j) Kommenter om det var nødvendigt at udføre hypotesetestet i forrige spørgsmål, eller om samme konklusion kunne opnås ud fra konfidensintervallerne alene? (Se Bemærkning 3.59 i bogen).

Korrelation

I forbindelse med efterfølgende opbygning af modeller til beskrivelse af varmeforbruget, f.eks. i forbindelse med projekt 2, vil vi sætte fokus på yderligere sammenhænge mellem udvalgte variable.

k) *I dette spørgsmål skal data fra hele tidsperioden oktober 2008 til juni 2011 benyttes (ikke kun data fra januar-februar 2010).* Angiv formlen for beregning af korrelationen mellem varmeforbruget for hus 1 ($Q1$) og global indstråling (G). Indsæt tal og beregn korrelationen G) (indsæt kun i korrelationsformlen, dvs. sæt kun tre tal ind!). Lav desuden et scatterplot der illustrerer sammenhængen mellem de to variable. Vurder om sammenhængen mellem plottet og korrelationen er som forventet.

Sammenlign den beregnede korrelation med resultatet fra følgende R-kode:

```
## Beregning af korrelation mellem Q1 og G
cor(D[, c("Q1","G")], use="pairwise.complete.obs")
```

Remark 2.1 Ekstra tips til R

Dette er en valgfri ekstra bemærkning om R-kodning (ikke nødvendig for at løse opgaven). Der er mange måder hvorpå man kan udtage en delmængde i R.

```
## Ekstra bemærkning om måder at udtage delmænger i R
##
## En logisk (logical) vektor med sandt (TRUE) eller falsk (FALSE) for
## hver række i D, f.eks. de dage hvor der har været frost
D$Ta < 0
## Den kan bruges, når man vil udvælge f.eks. varmeforbrug for hus 1
## på frostdage
D$Q1[D$Ta < 0]
## Alternativt kan man bruge funktionen 'subset'
subset(D, Ta < 0)
## Mere komplekse logiske udtryk kan laves, f.eks. udvælg kun
## observationer fra før 1. januar 2010 hvor der er frostvejr
subset(D, t < "2010-01-01" & Ta < 0)
```

Remark 2.2 Ekstra tips til R

Endnu en bemærkning med ekstra R-tips for de interesserede. Man kan f.eks. lave tabellen mere effektivt med en for-løkke.

```
## Lav en for-løkke med beregning af et par opsummerende størrelser
## og gem resultatet i en ny data.frame
Tbl <- data.frame()
## Find søjlerne med varmeforbrug
udvalgte <- c("Q1","Q2","Q3","Q4")
## Beregn nøgletallene for hvert hus
for(i in udvalgte){
  ## Udtag den relevante søjle
  x <- Dsel[, i]
  ## Tag stikprøvegennemsnittet
  Tbl[i, "mean"] <- mean(x, na.rm=TRUE)
  ## Tag stikprøvevariansen
  Tbl[i, "var"] <- var(x, na.rm=TRUE) }
## Se hvad der er i Tbl
Tbl

## I R er der endnu mere kortfattede måder hvorpår sådanne udregninger
## kan udføres. For eksempel
apply(Dsel[, udvalgte], 2, mean, na.rm=TRUE)
## eller flere ad gangen i et kald
apply(Dsel[, udvalgte], 2, function(x){
  c(mean=mean(x, na.rm=TRUE),
    var=var(x, na.rm=TRUE)) })
## Se flere smarte funktioner med: ?apply, ?aggregate og ?lapply
## og for ekstremt effektiv databehandling se f.eks. pakkerne: dplyr,
## tidyr, reshape2 og ggplot2.

## LaTeX tips:
## R-pakken "xtable" kan generere LaTeX tabeller og skrive dem direkte
## ind i en fil, som derefter kan inkluderes i et .tex dokument.
## R-pakken "knitr" kan anvendes meget elegant til at lave et .tex
## dokument der inkluderer R koden direkte i dokumentet. Dette
## dokument og bogen er lavet med knitr.
```