Projekt 2: Varmeforbrug i Sønderborg II

Formaliteter, struktur og forventninger – 2. obligatoriske opgave

I dette projekt er målet at opstille en passende multipel lineær regressionsmodel, der beskriver varmeforbruget om vinteren i fire huse i Sønderborg. Opgaven skal i praksis løses ved hjælp af den statistiske software R. Rundt omkring i opgaven er der givet forslag til R-koden, men udover det er det en god idé at se på R-koden fra projekt 1 samt f.eks. kapitel 5 og 6 i bogen.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, passende matematisk notation og 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 hvert 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 2:Varmeforbrug i Sønderborg II`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En 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.

Når I bliver bedt om at angive en formel, indsætte tal og derefter foretage en beregning er det vigtigt, at I viser I har gjort dette ved at inkludere nogle mellemregninger. (Disse steder er det ikke nok at anføre resultater aflæst i R). Husk også at et hypotesetest

består af følgende elementer: Angivelse af hypotese og signifikansniveau (α), teststørrelse inkl. dennes fordeling og p-værdi, samt en konklusion 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.

Data

Indlæs datasættet `soenderborg2_data.csv`. Følgende R-kode kan benyttes:

```
# Indlæs 'soenderborg2_data.csv' filen med data
D <- read.table("soenderborg2_data.csv", sep = ";", header = TRUE)
```

I dette projekt vil vi analysere data fra Sønderborgdatabasen, som består af målinger af varmeforbruget i bygninger samt klimatiske målinger som udendørstemperatur, solindstråling og vindhastighed. Bygningerne er typiske parcelhuse bygget i mursten i halvtredserne og tresserne. Husene er placeret i området omkring Borgmester Andersens vej i Sønderborg og er med varierende bebygget areal. De klimatiske målinger er registreret af en vejrstation placeret på det lokale fjernvarmeværk.

Datasættet til dette projekt omfatter observationer af fem variable:

- `t`: Observationsdatoen
- `houseId`: Hus id-nr.
- `Q`: Varmeforbrug (kW)
- `Ta`: Udendørs temperatur (°C)
- `G`: Solstråling på vandret overflade (W/m^2)

I dette projekt vil vi kun analysere data fra vinteren 2009/2010, her defineret som perioden 15. oktober 2009 til 15. april 2010. Desuden vil vi kun se på de fire huse med id hhv. 3, 5, 10 og 17. Vi skal derfor udvælge den relevante delmængde af det indlæste datasæt, til brug i den videre analyse. Dette kan f.eks. gøres ved hjælp af følgende R-kode:

```r
# Lav 't' om til en datovariabel i R
D$t <- as.Date(D$t, format = "%d/%m/%Y")

# Udvælg data for perioden 15. okt 2009 til 15. apr 2010 for de fire huse
D_model <- subset(D, ("2009-10-15" <= t & t < "2010-04-16") &
                  (houseId %in% c(3, 5, 10, 17)))
```

Datasættet har mange observationer, hvor en eller flere variable har manglende værdier (NA'er). Disse observationer skal fjernes fra datasættet inden den statistiske analyse. Dette kan f.eks. gøres med følgende R-kode:

```r
# Fjern observationer med manglende værdier
D_model <- na.omit(D_model)
```

Statistisk analyse

I det følgende skal der benyttes det datasæt, som indeholder observationer for hus 3, 5, 10, 17 i perioden 15. oktober 2009 til 15. april 2010, og som ikke har manglende værdier for nogen af variablene.

a) Lav en kort deskriptiv analyse og opsummering af data for variablene `Q`, `Ta` og `G`. Inkluder scatterplots af varmeforbruget mod de to andre variable, samt histogrammer og boxplots af alle tre variable. Der skal også være en tabel med opsummerende størrelser, som for hver variabel inkluderer antal observationer, gennemsnit, standardafvigelse, median samt 25%- og 75%-fraktiler.

b) Opstil en multipel lineær regressionsmodel med varmeforbruget som responsvariabel (Y_i), og med udendørstemperaturen og solens indstråling som forklarende variable (hhv. $x_{1,i}$ og $x_{2,i}$). Husk at angive forudsætningerne/de statistiske antagelser for modellen. (Se bemærkning 5.6, ligning (6-1) og eksempel 6.1).

c) Estimer modellens parametre, som består af regressionskoefficienterne, her kaldet β_0, β_1, β_2, og residualernes varians, σ^2. Brug evt. følgende R-kode:

```r
# Estimer multipel lineær regressionsmodel
fit <- lm(Q ~ Ta + G, data = D_model)

# Vis estimerede parametre mm.
summary(fit)
```

Giv en fortolkning af estimaterne $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, hvor du forklarer, hvad de siger om relationen mellem varmeforbruget og de to forklarende variable i modellen. (Se bemærkning 6.14). Angiv også de estimerede standardafvigelser for $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, frihedsgraderne anvendt til estimatet af residualernes varians $\hat{\sigma}^2$, samt modellens forklarede varians, R^2.

d) Foretag modelkontrol for at undersøge, om forudsætningerne for modellen (modellens antagelser) er opfyldte. Benyt de plots, der kan laves ved hjælp af R-koden nedenfor, som udgangspunkt for din vurdering. (Se afsnit 6.4 om residualanalyse).

```
# Plots til modelkontrol

# Observationer mod fittede værdier
plot(fit$fitted.values, D_model$Q, xlab = "Fittede værdier",
       ylab = "Varmeforbrug")

# Residualer mod hver af de forklarende variable
plot(D_model$FORKLARENDE_VARIABEL, fit$residuals,
        xlab = "INDSÆT TEKST", ylab = "Residualer")

# Residualer mod fittede værdier
plot(fit$fitted.values, fit$residuals, xlab = "Fittede værdier",
     ylab = "Residualer")

# Normal QQ-plot af residualerne
qqnorm(fit$residuals, ylab = "Residualer", xlab = "Z-scores",
       main = "")
qqline(fit$residuals)
```

e) Angiv formlen for et 95% konfidensinterval for koefficienten for udendørstemperaturen, her kaldet β_1. (Se metode 6.5). Indsæt tal i formlen og beregn konfidensintervallet. Benyt derefter nedenstående R-kode til at kontrollere resultatet og til at bestemme konfidensintervaller for de to andre koefficienter i modellen.

```r
# Konfidensintervaller for modellens koefficienter
confint(fit, level = 0.95)
```

f) Man er interesseret i, om β_1 kunne have værdien -0.25. Opstil den tilsvarende hypotese. Anvend signifikansniveauet $\alpha = 0.05$. Angiv formlen for den relevante teststørrelse (se metode 6.4), indsæt tal og beregn teststørrelsen. Angiv fordelingen af teststørrelsen (inkl. frihedsgrader), beregn p-værdien og konkluder.

g) Undersøg ved *backward selection* om modellen kan reduceres. (Se eksempel 6.13). Husk at reestimere modellen undervejs, hvis der kan foretages reduktion af modellen. Angiv slutmodellen og estimater for modellens parametre.

Benyt nu nedenstående R-kode til at lave et deldatasæt med én observation for hver af de fire huse fra vinteren 2008/2009. Bemærk at dette deldatasæt ikke indeholder nogen af de observationer, der blev benyttet i den statistiske analyse ovenfor. Dette datasæt skal benyttes til at vurdere slutmodellens prædiktionsevner.

```r
# Udvælg valideringsdatasæt
D_test <- subset(D, (t == "2008-12-06" & houseId == 3)|
                    (t == "2009-02-22" & houseId == 5)|
                    (t == "2009-03-12" & houseId == 10)|
                    (t == "2009-04-01" & houseId == 17))
```

h) Tag udgangspunkt i din slutmodel fra forrige spørgsmål. Bestem prædikterede værdier og 95% prædiktionsintervaller for varmeforbruget for hver af de fire observationer i valideringsdatasættet (`D_test`). Se eksempel 6.8, metode 6.9 og R-koden nedenfor. Sammenlign prædiktionerne med de observerede varmeforbrug for disse fire observationer og lav en vurdering af modellens evne til at prædiktere.

```r
# Prædiktioner og 95% prædiktionsintervaller
pred <- predict(SLUTMODEL, newdata = D_test, interval = "prediction",
                level = 0.95)

# Observerede værdier sammen med prædiktioner
cbind(id = D_test$houseId, Q = D_test$Q, pred)
```

Dvs. skriv ikke formlerne ind i rapporten, men istedet, at I har brugt R funktionen `predict` til beregningerne. Formlerne kræver en matrix formulering, som rækker ud over pensum (for at udlede formlerne kan ligningerne (6-48) og (6-49) bruges sammen med udledningerne der fører til ligningerne (5-57) og (5-58)).