Kursus 02402/02323 Introducerende Statistik

Forelæsning 9: Multipel lineær regression

Klaus K. Andersen og Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: klaus@cancer.dk

Oversigt

1. Warm up med lidt simpel lineær reg.
2. Multipel lineær regression
3. Model udvælgelse
4. Residual analyse (model kontrol)
5. Kurvelinearitet
6. Konfidens- og prædiktionsintervaller
7. Kollinearitet

Eksempel: Ozon koncentration

Vi har givet et sæt af sammenhængende målinger af: logaritmen af ozon koncentration ((log(ppb))), temperatur, solindstråling og vindhastighed:

	ozone	temperature	radiation	wind
1	3.45	67	190	7.4
2	3.30	72	118	8.0
3	2.29	74	149	12.6
⋮	⋮	⋮	⋮	⋮
110	2.62	76	131	8.0
111	2.71	68	223	11.5

Eksempel: Ozonkoncentration

- Lad os se på sammenhængen mellem log ozon koncentrationen og temperaturen
- Brug en *simpel lineær regressionsmodel*

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

hvor

- Y_i er log ozonkoncentrationen for måling i
- x_i er temperaturen ved måling i

Fit modellen i R

```r
## Read the data
Air <- read.table(file="air.txt", sep=",", header=TRUE)
## What is in Air?
str(Air)
Air
head(Air)

## See the relation between ozone and temperature
plot(Air$temperature, Air$ozone, xlab="Temperature", ylab="Ozon")

## Correlation
cor(Air$ozone, Air$temperature)

## Fit a simple linear regression model
summary(lm(ozone ~ temperature, data=Air))

## Add a vector with random values, is there a significant linear relation?
## JUST shown for illustration!! (not something you do in real cases)
Air$noise <- rnorm(nrow(Air))
plot(Air$ozone, Air$noise, xlab="Noise", ylab="Ozon")
cor(Air$ozone, Air$noise)
summary(lm(ozone ~ noise, data=Air))
```

Simpel lineær regressionsmodel til de to andre

Vi kan også lave en simpel lineær regressionsmodel med de to andre

```
## Simpel lineær regressionsmodel med vindhastigheden
plot(Air$ozone, Air$wind, xlab="Ozone", ylab="Wind speed")
cor(Air$ozone, Air$wind)
summary(lm(ozone ~ wind, data=Air))

## Simple linear regression model with the radiation
plot(Air$ozone, Air$radiation, xlab="Ozone", ylab="Radiation")
cor(Air$ozone, Air$radiation)
summary(lm(ozone ~ radiation, data=Air))
```

Multipel lineær regression

- Y er den *afhængige variabel* (dependent variable)
- Vi er interesseret i at modellere Y's afhængighed af de *forklarende eller uafhængige variabler* (explanatory eller independent variables) $x_1, x_2, ..., x_p$
- Vi undersøger en *lineær sammenhæng* mellem Y og $x_1, x_2, ..., x_p$, ved en regressionsmodel på formen

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

- Y_i og ε_i er stokastiske variabler og $x_{j,i}$ er variabler

Mindste kvadraters metode (least squares)

- Residualerne findes ved at prædiktionen

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$$

indsættes

$$y_i = \hat{y}_i + e_i$$
"observation = prædiktion + residual"

og trækkes fra

$$e_i = y_i - \hat{y}_i$$
"residual = observation − prædiktion"

Mindste kvadraters metode (least squares)

- Ved det bedste estimat for $\beta_0, \beta_1, ..., \beta_p$ forstås de værdier $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ der minimerer residual sum of squares (RSS)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- og estimatet for afvigelsernes (ε_i) standard afvigelse er

$$\widehat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} e_i^2$$

- Find og læs sektion med Theorem 6.2

Mindste kvadraters metode

$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ findes ved at løse de såkaldte normalligninger, der for $p = 2$ er givet ved

$$\begin{aligned}
\sum_{i=1}^{n} y_i =& n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2} \\
\sum_{i=1}^{n} x_{i,1} y_i =& \hat{\beta}_0 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,1} x_{i,2} \\
\sum_{i=1}^{n} x_{i,2} y_i =& \hat{\beta}_0 \sum_{i=1}^{n} x_{i,2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} x_{i,2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2}^2
\end{aligned}$$

Man skal gange nogle matricer sammen.

Udvid modellen (forward selection)

- *Ikke beskrevet i eNoten*
- Start med *mindste model* med den mest signifikante (mest forklarende) variabel
- *Udvid modellen* med de andre forklarende variabler (inputs) en ad gangen
- *Stop* når der ikke er flere signifikante udvidelser

```
## Forward selection:
## Tilføj vind til modellen
summary(lm(ozone ~ temperature + wind, data=Air))
## Tilføj indstraaling til modellen
summary(lm(ozone ~ temperature + wind + radiation, data=Air))
```

Formindsk modellen (model reduction eller backward selection)

- *Beskrevet i eNoten, sektion 6.5*
- Start med den fulde model
- Fjern den mest insignifikante forklarende variabler
- Stop hvis alle prm. estimater er signifikante

```
## Fit den fulde model
summary(lm(ozone ~ temperature + wind + radiation + noise,
           data=Air))
## Fjern det mest ikke-signifikante input, er alle nu sigifikante?
summary(lm(ozone ~ temperature + wind + radiation, data=Air))
```

Model udvælgelse

- Der er ikke noget sikker metode til at finde den bedste model!
- Det vil kræve subjektive beslutninger at udvælge en model
- Forskellige procedurer, enten forward eller backward, afhænger af forholdene
- Statistiske tests mål til at sammenligne modeller
- Her i kurset kun backward procedure beskrevet

Residual analyse (model kontrol)

- Model kontrol: Analyser residualerne for at checke at forudsætningerne er opfyldt
- $e_i \sim N(0, \sigma^2)$ og er independent and identically distributed (i.i.d.)
- Samme som for simpel lineær model

Antagelse om normalfordelte residualer

- Lav et qq-plot (normal score plot) for at se om de ikke afviger fra at være normalfordelt

```
## Gem det udvalgte fit
fitSel <- lm(ozone ~ temperature + wind + radiation, data=Air)

## qq-normalplot
qqnorm(fitSel$residuals)
qqline(fitSel$residuals)
```

Antagelse om identisk distribution

- Plot residualerne (e_i) mod de prædikterede (fittede) værdier ($\hat{y}_i$)

```
plot(fitSel$fitted.values, fitSel$residuals, xlab="Prædikteret værdi", ylab="Residualer")
```

Det ser ud som om modellen godt kan forbedres...

- Plot residualer mod de forklarende variabler

```
pairs(cbind(fitSel$residuals, Air[,c("temperature","wind",
                "radiation")]), panel = panel.smooth)
```

Kurvelineær (Curvilinear)

Hvis vi ønsker at estimere en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

kan vi benytte multipel lineær regression i modellen

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

hvor

- $x_{i,1} = x_i$
- $x_{i,2} = x_i^2$

og benytte samme metoder som ved multipel lineær regression.

Udvid ozon modellen med passende kurvelineær regression

```r
## Lav den kvadrerede vind
Air$windSq <- Air$wind^2
## Tilføj den til modellen
fitWindSq <- lm(ozone ~ temperature + wind + windSq + radiation, data=Air)
summary(fitWindSq)

## Gør tilsvarende for temperatur
Air$temperatureSq <- Air$temperature^2
## Tilføj
fitTemperatureSq <- lm(ozone ~ temperature + temperatureSq + wind + radiation, data=Air)
summary(fitTemperatureSq)

## Gør tilsvarende for indstråling
Air$radiationSq <- Air$radiation^2
## Tilføj
fitRadiationSq <- lm(ozone ~ temperature + wind + radiation + radiationSq, data=Air)
summary(fitRadiationSq)

## Hvilken en var bedst!?
summary(fitWindSq)
summary(fitTemperatureSq)
```

Udvid ozon modellen med passende kurvelineær regression (fortsat)

```
summary(fitWindSqTemperatureSq)

## Model kontrol
qqnorm(fitWindSq$residuals)
qqline(fitWindSq$residuals)
plot(fitWindSq$fitted.values, fitWindSq$residuals, pch=19)


################################
## Plot residualerne vs. de forklarende variabler
pairs(cbind(fitWindSq$residuals, Air[,c("temperature","wind","radiation")]), panel=panel.smooth)
```

Konfidens- og prædiktionsintervaller

```
## Generer et nyt data.frame med konstant temperatur og instråling, men varierende vindhastighed
wind<-seq(1,20.3,by=0.1)
setTemperature <- 78
setRadiation <- 186
AirForPred <- data.frame(temperature=setTemperature, wind=wind, windSq=wind^2, radiation=setRadiation)

## Udregn konfidens- og prædiktionsintervaller (-bånd)
## Læg mærke til at der tilbage transformeres
CI <- predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95)
PI <- predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95)

## Plot them
plot(Air$wind, Air$ozone, ylim=range(CI,PI,Air$ozone), xlab="", ylab="")
title(xlab="vindhastighed (MpH)", ylab="ozon (ppb)",
main=paste("Ved temperatur =",setTemperature, "F og indstraaling = ",setRadiation,"Langleys"))
lines(wind, CI[,"fit"])
lines(wind, CI[,"lwr"], lty=2, col=2)
lines(wind, CI[,"upr"], lty=2, col=2)
lines(wind, PI[,"lwr"], lty=2, col=3)
lines(wind, PI[,"upr"], lty=2, col=3)
## legend
legend("topright", c("Prædiktion","95% konfidensbånd","95% prædiktionsbånd"), lty=c(1,2,2), col=1:3)
```

Kollinearitet (Colinearity)

Der er opstår problemer hvis de forklarende variabler er stærkt korrelerede

```
## Generer nogle værdier til brug for MLR
n <- 100

## Første forklarende variabel en sinus
x1 <- sin(0:(n-1)/(n-1)*2*2*pi) + rnorm(n, 0, 0.1)
plot(x1, type="b")

## Den anden forklarende variabel er x1 med lidt støj
x2 <- x1 + rnorm(n, 0, 0.1)

## x1 og x2 er altså meget korrelerede
plot(x1,x2)
cor(x1,x2)

## Simuler en MLR
beta0=20; beta1=1; beta2=1; sigma=1
y <- beta0 + beta1 * x1 + beta2 * x2 + rnorm(n,0,sigma)

## Se scatter plots for y mod x1, og y mod x2
par(mfrow=c(1,2))
plot(x1,y)
plot(x2,y)

## Fit en MLR
summary(lm(y ~ x1 + x2))
```

Kollinearitet (Colinearity) (fortsat)

```
## Hvis det var et eksperiment og man havde adskilt påvirkningerne i designet
x1[1:(n/2)] <- 0
x2[(n/2):n] <- 0

## Plot dem
plot(x1, type="b")
lines(x2, type="b", col="red")

## Nu meget lav korrelation
cor(x1,x2)

## Simuler MLR igen
y <- beta0 + beta1 * x1 + beta2 * x2 + rnorm(n,0,sigma)

## og fit MLR
summary(lm(y ~ x1 + x2))
```

Det er vigtigt hvordan man designer sit eksperiment!!

Oversigt

1. Warm up med lidt simpel lineær reg.
2. Multipel lineær regression
3. Model udvælgelse
4. Residual analyse (model kontrol)
5. Kurvelinearitet
6. Konfidens- og prædiktionsintervaller
7. Kollinearitet