Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 9: Multipel lineær regression

Jan Kloppenborg Møller
DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.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
8. Skive fjord

Oversigt

Eksempel: Ozon koncentration

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

ozone	radiation	wind	temperature	month	day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
⋮	⋮	⋮	⋮	⋮	
18	131	8.0	76	9	29
20	223	11.5	68	9	30

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

Oversigt

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 variable og $x_{j,i}$ er variable

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) varians 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.

Eller Matrix- formulering

$$\mathbf{0} = \frac{\partial RSS}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1}$$

$$= 2(\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}). \tag{2}$$

eller (normalligningerne)

$$\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \tag{3}$$

Med løsningen

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{4}$$

Matrix formulation

The estimators of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{5}$$

and the covariance matrix of the estimates is

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{6}$$

and central estimate for the residual variance is

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)} \tag{7}$$

Hypotese test (partial t-test)

The estimate of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{8}$$

and the covariance matrix of the estimates is

$$\hat{\boldsymbol{\Sigma}}_\beta = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{9}$$

The obeserved t-statistic for the hypothesis: $H_0 : \beta_i = \beta_{i,0}$ is

$$t_{obs,i} = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{(\hat{\boldsymbol{\Sigma}}_\beta)_{ii}}}. \tag{10}$$

Should be compared with a t-distribution with $n - (p + 1)$ degrees of freedom.

Konfidens og prædiktions interval

$\boldsymbol{x}_{new} = [1, x_{1,new}, \ldots, x_{p,new}]$:
Konfidensinterval for middelværdi

$$V(\hat{Y}_{new}) = V(\boldsymbol{x}_{new}\hat{\boldsymbol{\beta}}) \tag{11}$$
$$= \sigma^2 \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T, \tag{12}$$

Prediction variance

$$V(Y_{new}) = V(\boldsymbol{x}_{new}\hat{\boldsymbol{\beta}} + \epsilon_{new}) \tag{13}$$
$$= \sigma^2(1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T). \tag{14}$$

in practice replace σ^2 with its estimate ($\hat{\sigma}^2$), and hence use quantiles of the appropriate t-distribution.

Oversigt

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

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 parameter estimater er signifikante

Model udvælgelse

- Der er ikke nogen 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
- Brug statistiske tests til at sammenligne modeller

Oversigt

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

Antagelser

- Lav et qq-plot (normal score plot) for at se om de ikke afviger fra at være normalfordelt
- Plot residualerne (e_i) mod de prædikterede (fittede) værdier ($\hat{y}_i$)
- Plot residualer mod de forklarende variabler

Uafhængigheds antagelsen (lidt ud over pensum)

- Vi antager det ofte uden yderligere undersøgelse
- Plot residualerne (e_i) som funktion af tiden (hvis meningsfyldt)
- Plot e_i mod e_{i-1}
- Tjek korrelationen mellem e_i mod e_{i-1}
- og en række andre metoder...

Oversigt

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

Brug modellen

$$Y_i = \beta_0 + \beta_1 wind + \beta_2 rad + \beta_3 temp + \beta_4 wind^2 + \beta_5 rad^2 + \beta_6 temp^2 + \epsilon_i$$

og brug back-ward selection til at finde den bedste model.

Oversigt

Konfidens- og prædiktionsintervaller i R

```
################################
## Konfidens- og prædiktionsintervaller for den kurvelineære model

fitWindSq <- lm(logOzone ~ temperature + wind + windSq + radiation, data=Air)

## 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)
```

Oversigt

Kollinearitet (Collinearity)

Det er problematisk hvis de forklarende variabler er stærkt korrelerede.

- De forklarende variable skal være lineært uafhængige.
- Tjek korrelationener mellem forklarende variable (ingen tæt på ± 1)
- Ingen korrelationer i parameter korrelationsmatricen tæt på en.
- Det er vigtigt hvordan man designer sit eksperiment!!

Kollinearitet (Colinearity)

Som eksempel se på

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

antag at $x_2 = a + bx_1$, så er

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_1 + \beta_2 (a + bx_1) + \epsilon_i \\ =& \beta_0 + \beta_2 a + (\beta_1 + \beta_2 b) x_1 + \epsilon_i. \end{aligned}$$

Dvs. 2 (ikke 3) parametre.

Kollinearitet (Collinearity)

If we have identified a collinarity problem,

- We should be carefull about parameter interpretation
- We should reduce the model by removing parameters

Other methods exist e.g. Principal Component Regression and Ridge regression.

Oversigt

Modellering af phytoplankton

Formuler en lineær model for phytoplankton i Skive fjord, estimer modellens parametre og foretag modelkontrol.

Outline