

# 02402: Introduktion til Statistik

## Forelæsning 9: Multipel lineær regression

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Agenda

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

## Eksempel: Ozonkoncentration

Vi har et sæt af sammenhængende målinger af: logaritmen af ozonkoncentration ( $\log(\text{ppb})$ ), 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

```

## Se info about data
?airquality
## Copy the data
Air <- airquality
## Remove rows with at least one NA value
Air <- na.omit(Air)

## Remove one outlier
Air <- Air[-which(Air$Ozone == 1), ]

## Check the empirical density
hist(Air$Ozone, probability=TRUE, xlab="Ozon", main="")

## Concentrations are positive and very skewed, let's
## log-transform right away:
## (although really one could wait and check residuals from models)
Air$logOzone <- log(Air$Ozone)
## Bedre epåf?
hist(Air$logOzone, probability=TRUE, xlab="log Ozone", main="")

## Make a time variable (R timeclass, se ?POSIXct)
Air$t <- ISOdate(1973, Air$Month, Air$Day)
## Keep only some of the columns
Air <- Air[,c(7,4,3,2,8)]
## New names of the columns
names(Air) <- c("logOzone", "temperature", "wind", "radiation", "t")

## What's in Air?
str(Air)
Air
head(Air)
tail(Air)

## Typically one would begin with a pairs plot
pairs(Air, panel = panel.smooth, main = "airquality data")

```

## Eksempel: Ozonkoncentration

- Lad os først se på sammenhængen mellem log ozon koncentrationen og temperaturen.
- Anvend 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

```
#####  
  
## See the relation between ozone and temperature  
plot(Air$temperature, Air$logOzone, xlab="Temperature", ylab="Ozon")  
  
## Correlation  
cor(Air$logOzone, Air$temperature)  
  
## Fit a simple linear regression model  
summary(lm(logOzone ~ temperature, data=Air))  
  
## Add a vector with random values, is there a significant linear relation?  
## ONLY for ILLUSTRATION purposes  
Air$noise <- rnorm(nrow(Air))  
plot(Air$logOzone, Air$noise, xlab="Noise", ylab="Ozon")  
cor(Air$logOzone, Air$noise)  
summary(lm(logOzone ~ noise, data=Air))
```

# Simple lineær regressionsmodel til de to andre

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

```
#####  
## With each of the other two independent variables  
  
## Simple linear regression model with the wind speed  
plot(Air$logOzone, Air$wind, xlab="logOzone", ylab="Wind speed")  
cor(Air$logOzone, Air$wind)  
summary(lm(logOzone ~ wind, data=Air))  
  
## Simple linear regression model with the radiation  
plot(Air$logOzone, Air$radiation, xlab="logOzone", ylab="Radiation")  
cor(Air$logOzone, Air$radiation)  
summary(lm(logOzone ~ radiation, data=Air))
```



# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression**
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# Multipel lineær regression

- $Y$  er den *afhængige variabel*
- Vi er interesseret i at modellere  $Y$ 's afhængighed af de *forklarende* eller *uafhængige* variable (explanatory eller independent variables)  $x_1, x_2, \dots, x_p$
- Vi modellerer en *lineær sammenhæng* mellem  $Y$  og  $x_1, x_2, \dots, x_p$ , ved en regressionsmodel på formen
$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$
- $Y_i$  og  $\varepsilon_i$  er stokastiske variable, og  $x_{j,i}$  er (deterministiske) variable

## 'Least squares'-estimator

- Koefficientestimatorerne findes ved at minimere:

$$RSS(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})]^2$$

- De "prædikterede" (= "fitted") værdier findes ved:

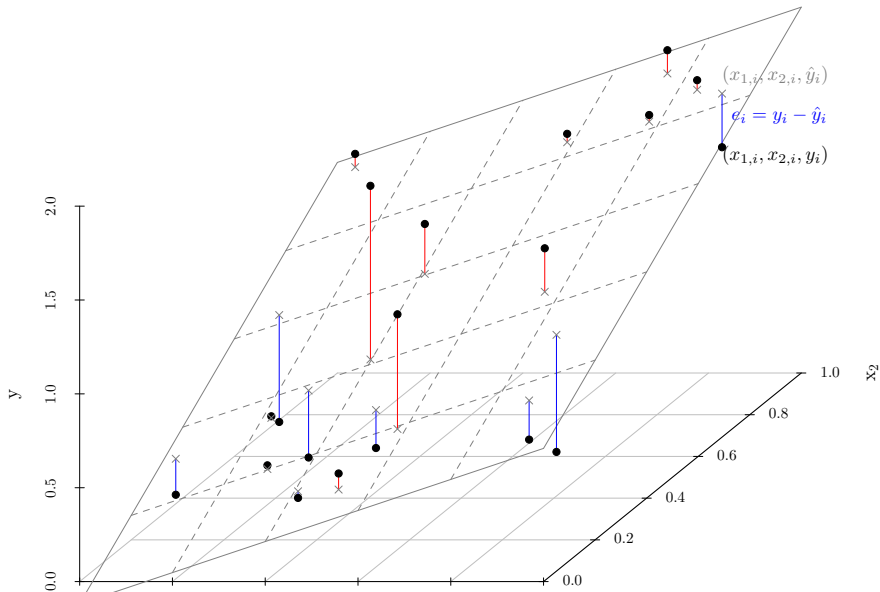
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p}$$

- Og residualerne er så:

$$e_i = y_i - \hat{y}_i$$

residual = observation – prædiktion

# 'Least squares'-estimer – Konceptet!



## MLR – vigtige resultater – uden eksplicitte formler

- Remark 6.6: Find  $\hat{\beta}_i$  og  $\hat{\sigma}_{\beta_i}$  fra R-outputtet (`summary(myfit)`)

## MLR – vigtige resultater – uden eksplicitte formler

- Remark 6.6: Find  $\hat{\beta}_i$  og  $\hat{\sigma}_{\beta_i}$  fra R-outputtet (`summary(myfit)`)
- Theorem 6.2: t-fordelingen kan bruges til inferens for modelparametre.

## MLR – vigtige resultater – uden eksplicitte formler

- Remark 6.6: Find  $\hat{\beta}_i$  og  $\hat{\sigma}_{\beta_i}$  fra R-outputtet (`summary(myfit)`)
- Theorem 6.2: t-fordelingen kan bruges til inferens for modelparametre.
- Metode 6.4 og 6.5: Hypotesetests og konfidensintervaller ud fra R-outputtet.

## MLR – vigtige resultater – uden eksplicitte formler

- Remark 6.6: Find  $\hat{\beta}_i$  og  $\hat{\sigma}_{\beta_i}$  fra R-outputtet (`summary(myfit)`)
- Theorem 6.2: t-fordelingen kan bruges til inferens for modelparametre.
- Metode 6.4 og 6.5: Hypotesetests og konfidensintervaller ud fra R-outputtet.
- Altsammen: **Samme som for simpel lineær regression!**



## MLR – vigtige resultater – uden eksplicitte formler

- Remark 6.6: Find  $\hat{\beta}_i$  og  $\hat{\sigma}_{\beta_i}$  fra R-outputtet (`summary(myfit)`)
- Theorem 6.2: t-fordelingen kan bruges til inferens for modelparametre.
- Metode 6.4 og 6.5: Hypotesetests og konfidensintervaller ud fra R-outputtet.
- Altsammen: **Samme som for simpel lineær regression!**
- (I Afsnit 6.6 af bogen: Matrix-baseret tilgang med eksplicitte formler. Ikke pensum i kursus 02402)

# Outputtet

```
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))

##
## Call:
## lm(formula = logOzone ~ temperature + wind + radiation, data = Air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.0203 -0.3150 -0.0094  0.3230  1.1223
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.261436   0.520496   0.50   0.62
## temperature  0.044457   0.005678   7.83 3.9e-12 ***
## wind         -0.069283   0.014514  -4.77 5.8e-06 ***
## radiation    0.002190   0.000516   4.25 4.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.467 on 106 degrees of freedom
## Multiple R-squared:  0.674, Adjusted R-squared:  0.664
## F-statistic: 72.9 on 3 and 106 DF,  p-value: <2e-16
```

- Læs estimer, usikkerheder osv. fra outputtet.

## Parameter interpretation in MLR (Remark 6.14)

Hvad er  $\hat{\beta}_i$  udtryk for?

- Den forventede ændring i  $y$  når  $x_i$  ændres én enhed.

## Parameter interpretation in MLR (Remark 6.14)

Hvad er  $\hat{\beta}_i$  udtryk for?

- Den forventede ændring i  $y$  når  $x_i$  ændres én enhed.
- Effekten af  $x_i$  givet de øvrige variable.
- Effekten af  $x_i$  korrigeret for de øvrige variables effekt.
- Effekten af  $x_i$  ”når de andre variable er uændret”.

## Parameter interpretation in MLR (Remark 6.14)

Hvad er  $\hat{\beta}_i$  udtryk for?

- Den forventede ændring i  $y$  når  $x_i$  ændres én enhed.
- Effekten af  $x_i$  givet de øvrige variable.
- Effekten af  $x_i$  korrigeret for de øvrige variables effekt.
- Effekten af  $x_i$  ”når de andre variable er uændret”.
- Afhænger af hvad der ellers i modellen!
- Generelt: IKKE en kausal effekt/interventionseffekt!

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion**
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# Udvid modellen (forward selection)

- Ikke inkluderet i kursusbogen
- Start med den *lineære regressionsmodel* med den mest signifikante forklarende variabel
- *Udvid modellen* med andre forklarende variabler (inputs) én ad gangen
- *Stop* når der ikke er flere signifikante udvidelser.

```
#####  
## Extend the model  
  
## Forward selection:  
## Add wind to the model  
summary(lm(logOzone ~ temperature + wind, data=Air))  
## Add radiation to the model  
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```

## Formindsk modellen (model reduction eller *backward selection*)

- *Beskrevet i kursusbogen, afsnit 6.5*
- Start med den fulde model.
- Fjern den mest insignifikante variabel.
- Stop når alle tilbageværende parameterestimer er signifikante.

```
#####
```

```
## Backward selection
```

```
## Fit the full model
```

```
summary(lm(logOzone ~ temperature + wind + radiation + noise, data=Air))
```

```
## Remove the most non-significant input, are all now significant?
```

```
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```



# Modelselektion

- Der er ikke nogen sikker metode til at finde den bedste model!
- Det kræver subjektive beslutninger at udvælge en model.
- Forskellige procedurer, enten forward eller backward selection (eller begge), afhænger af forholdene.
- Statistiske mål og tests til at sammenligne modeller
- Her i kurset er kun backward selection beskrevet

# Modelselektion

I rapporter: husk at beskrive og reflektere over hvad I gør! (gælder både statistik og andet)

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)**
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# Residualanalyse (modelkontrol)

- Modelkontrol: Analysér residualerne for at tjekke at antagelserne er opfyldt.
- $e_i \sim N(0, \sigma^2)$  er uafhængige og ensfordelte (i.i.d.)
- Samme som for den simple lineære model

# Antagelse om normalfordelte residualer

- Lav et qq-plot for at se om de ikke afviger fra at være normalfordelt

```
#####  
## Assumption of normal distributed residuals  
  
## Save the selected fit  
fitSel <- lm(logOzone ~ temperature + wind + radiation, data=Air)  
  
## qq-normalplot  
qqnorm(fitSel$residuals)  
qqline(fitSel$residuals)
```

## Antagelse om identisk fordelte residualer

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

```
#####
## Plot the residuals vs. predicted values

plot(fitSel$fitted.values, fitSel$residuals, xlab="Predicted values",
      ylab="Residuals")
```

- Det ser ud som om modellen godt kan forbedres...
- Plot residualer mod de forklarende variable:

```
#####
## Plot the residuals vs. the independent variables

par(mfrow=c(1,3))
plot(Air$temperature, fitSel$residuals, xlab="Temperature")
plot(Air$wind, fitSel$residuals, xlab="Wind speed")
plot(Air$radiation, fitSel$residuals, xlab="Radiation")
```

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet**
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

## Kurvelineær model

Hvis vi vil benytte en model af typen

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

så kan vi bruge en multipel lineær regressionsmodel

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

hvor

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

og bruge de samme metoder som for multipel lineær regression.



# Udvid ozonmodellen med passende kurvilinear regression

```
#####
## Extend the ozone model with appropriate curvilinear regression

## Make the squared wind speed
Air$windSq <- Air$wind^2
## Add it to the model
fitWindSq <- lm(logOzone ~ temperature + wind + windSq + radiation, data=Air)
summary(fitWindSq)

## Equivalently for the temperature
Air$temperature2 <- Air$temperature^2
## Add it
fitTemperatureSq <- lm(logOzone ~ temperature + temperature2 + wind + radiation, data=Air)
summary(fitTemperatureSq)

## Equivalently for the radiation
Air$radiation2 <- Air$radiation^2
## Add it
fitRadiationSq <- lm(logOzone ~ temperature + wind + radiation + radiation2, data=Air)
summary(fitRadiationSq)

## Which one was best?
## One could try to extend the model further
fitWindSqTemperaturSq <- lm(logOzone ~ temperature + temperature2 + wind + windSq + radiation, data=Air)
summary(fitWindSqTemperaturSq)

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

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller**
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# Konfidens- og prædiktionsintervaller, Method 6.9:

Som for simpel lineær regression (i princippet).

```
#####
## Confidence and prediction intervals for the curvilinear model

## Generate a new data.frame with constant temperature and radiation, but with varying wind speed
wind<-seq(1,20.3,by=0.1)
AirForPred <- data.frame(temperature=mean(Air$temperature), wind=wind,
                          windSq=wind^2, radiation=mean(Air$radiation))

## Calculate confidence and prediction intervals (actually bands)
CI <- predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95)
PI <- predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95)

## Plot them
plot(wind, CI[, "fit"], ylim=range(CI,PI), type="l",
      main=paste("At temperature =",format(mean(Air$temperature),digits=3),
                "and radiation =", format(mean(Air$radiation),digits=3)))
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("Prediction","95% confidence band","95% prediction band"), lty=c(1,2,2), col=1:3)
```

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet**
- 8 Den 'samlede' regressionsmetode

# Kollinearitet

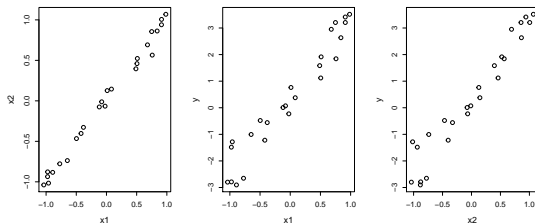
- Hvis to (eller flere) forklarende variable har en perfekt lineær sammenhæng, så kan vi ikke afgøre hvilken som er forklarende.
- Også et problem hvis sammenhængen er tæt på lineær.

# Kollinearitet

- Hvis to (eller flere) forklarende variable har en perfekt lineær sammenhæng, så kan vi ikke afgøre hvilken som er forklarende.
- Også et problem hvis sammenhængen er tæt på lineær.
- Med e.g. to meget korrelerede  $x$ -variable:
  - *Sammen* kan det være at ingen af dem har en "unik" effekt.
  - *Separat* kan de have en stor effekt

# Kollinearitet – eksempel

To meget korrelerede forklarende variable  $x_1$  og  $x_2$  og responsvariabel  $y$ .



## Kollinearitet – eksempel

x1 og x2: Hver for sig: vi så en tydelig effekt på y.  
Men *sammen*: ingen af dem er signifikante:

```
#####
L <- lm(y ~ x1 + x2)
summary(L)

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7951 -0.3723  0.0038  0.3546  1.2247
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.376     0.109     3.44  0.0023 **
## x1              0.709     1.535     0.46  0.6485
## x2              2.167     1.523     1.42  0.1688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.534 on 22 degrees of freedom
## Multiple R-squared:  0.941, Adjusted R-squared:  0.936
## F-statistic: 175 on 2 and 22 DF,  p-value: 3.05e-14
```



- Svært at separere effekter af kollinære variable
- Ingen nem løsning på kollinearitet
- Fornuftigt designet eksperiment kan hjælpe.

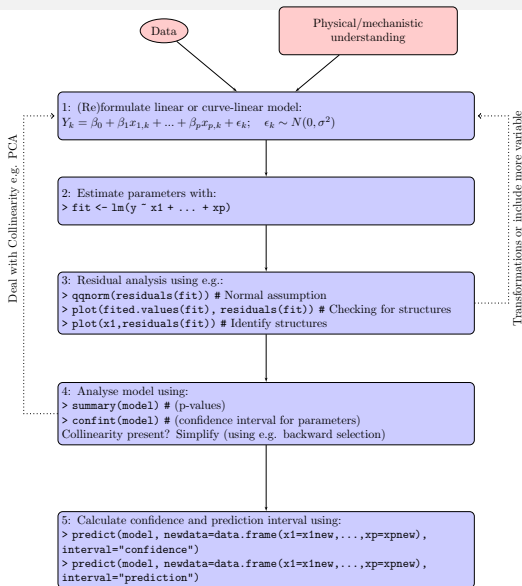
- Svært at separere effekter af kollinære variable
- Ingen nem løsning på kollinearitet
- Fornuftigt designet eksperiment kan hjælpe.

Det er vigtigt hvordan man designer sit eksperiment!

# Overview

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

# The overall regression method box 6.16



# Agenda

- 1 Opvarmning: simpel lineær regression
- 2 Multipel lineær regression
- 3 Modelselektion
- 4 Residualanalyse (modelkontrol)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode