Introduktion til Statistik

Forelæsning 9: Multipel lineær regression

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2021

Kapitel 6: Multipel lineær regressions analyse

Multipel lineær regressionsmodel

- Flere variabler: Y, x_1, x_2, $\ldots$
  (y afhængig/respons var. og x'er er forklarende/uafhængige var.)
- Mindstekvadraters rette plan *(et plan da der er $>$2 dimensioner)*

Inferens for en multipel lineær regressionmodel

- Statistisk model: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$
- Estimation af konfidensintervaller og tests for β'er
- Konfidensintervaller for modellen (middelplanet)
- Prædiktionsintervaller for nye punkter
- R^2 er andelen af den totale variationen som er forklaret af modellen

Model validering af antagelser ved residual analyse

- Normalfordeling? q-q plots af residualer
- Uafhængighed? Plot residualer mod prædikterede værdier $\hat{y}_i$ og inputs $x_{j,i}$

Chapter 6: Multiple linear Regression Analysis

Multipel lineær regressionsmodel

- Many quantitative variables: y, x_1, x_2, ...
  (y is the dependent/response var. and x's are explanatory/independent var.)
- Calculating least squares surface *(a plane surface since there are >2 dimensions)*

Inferences for a the multiple linear regression model

- Statistical model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_2 x_{p,i} + \varepsilon_i$
- Confidence interval estimation and test for the β's
- Confidence interval for the model (the mean surface)
- Prediction interval for new points
- R^2 expresses the proportion of the total variation explained by the linear fit

Model validation of assumptions with residual analysis

- Normal distribution? q-q plots of residuals
- Independence? Plot residuals against predicted values $\hat{y}_i$ and inputs $x_{j,i}$

Oversigt

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

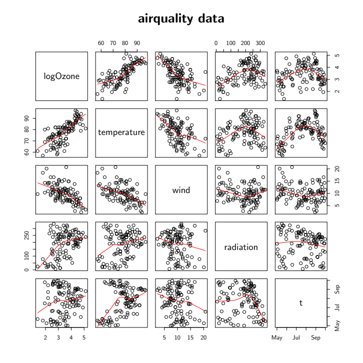
Eksempel: Ozon koncentration

Vi har givet et sæt af sammenhængende målinger af: 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: Ozon koncentration

```
## Se info om data
?airquality
## Indlæs data
Air <- airquality
## Fjern rækker hvor der er mindst en NA værdi
Air <- na.omit(Air)
## Fjern en outlier
Air <- Air[-which(Air$Ozone == 1), ]

## Se lige på empirisk tæthedsfunktion
hist(Air$Ozone, probability=TRUE, xlab="Ozon", main="")
## Koncentrationer er positive og meget højre-skæv fordeling, derfor log transformer
Air$logOzone <- log(Air$Ozone)
## Bedre epdf?
hist(Air$logOzone, probability=TRUE, xlab="log Ozon", main="")

## Lav en tid (R tidsklasse, se ?POSIXct)
Air$t <- ISOdate(1973, Air$Month, Air$Day)
## Behold kun nogle af kolonnerne
Air <- Air[ ,c(7,4,3,2,8)]
## Nye navne på kolonnerne
names(Air) <- c("logOzone","temperature","wind","radiation","t")

## Hvad er der i Air?
str(Air)
Air
head(Air)
tail(Air)

## Typisk vil man starte med et pairs plot
pairs(Air, panel = panel.smooth, main = "airquality data")
```



Eksempel: Ozonkoncentration

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

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

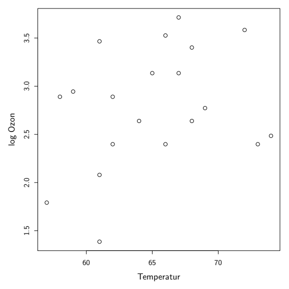
hvor

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

Fit simpel lineær regressions model i R

Antag at vi kun havde de første 20 datapunkter
Er der afhængighed?

```
## Start med at sige at vi har 20 datapunkter
Air20 <- Air[1:20, ]

## Se på sammenhængen mellem log(ozon) og temperatur
plot(Air20$temperature, Air20$logOzone, xlab="Temperatur", ylab="log Ozon")

## Korrelation
cor(Air20$logOzone, Air20$temperature)
```



Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation med 20 observationer
summary(lm(logOzone ~ temperature, data=Air20))

##
## Call:
## lm(formula = logOzone ~ temperature, data = Air20)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2476 -0.4560  0.0559  0.4154  0.8550
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3494     1.8483    0.19     0.85
## temperature   0.0375     0.0284    1.32     0.20
##
## Residual standard error: 0.6 on 18 degrees of freedom
## Multiple R-squared:  0.0883,Adjusted R-squared:  0.0377
## F-statistic: 1.74 on 1 and 18 DF,  p-value: 0.203
```

Er der signifikant korrelation mellem `logOzone` og `temperature` på 5% signifikansniveau?

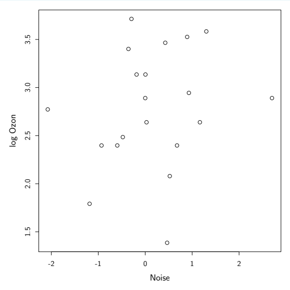
A: Ja B: Nej C: Ved ikke

Svar B: Nej, da p-værdien for $H_0 : \beta_1 = 0$ er $0.20 > 0.05$, dvs. H_0 accepteres

Tilføj en vektor med tilfældige værdier

Simuler 20 *uafhængige stokastiske variable* og tilføj til data.frame
Er der sammenhæng?

```
## Er der signifikant lineær sammenhæng (korrelation)?
Air20$noise <- rnorm(20)
plot(Air20$noise, Air20$logOzone, xlab="Noise", ylab="log Ozon")
cor(Air20$noise, Air20$logOzone)
```



Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation med 20 observationer
summary(lm(logOzone ~ noise, data=Air20))

##
## Call:
## lm(formula = logOzone ~ noise, data = Air20)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4344 -0.2721 -0.0303  0.4557  0.9819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.766      0.140   19.71  1.2e-13 ***
## noise          0.117      0.138    0.85     0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.62 on 18 degrees of freedom
## Multiple R-squared:  0.0382,Adjusted R-squared:  -0.0152
## F-statistic: 0.715 on 1 and 18 DF,  p-value: 0.409
```

Er der signifikant korrelation mellem `logOzone` og `noise` på 5% signifikansniveau?

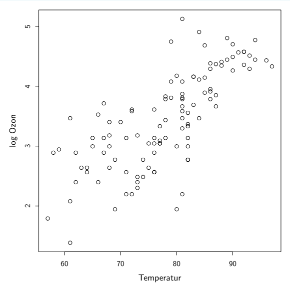
A: Ja B: Nej C: Ved ikke

Svar B: Nej, da p-værdien for $H_0 : \beta_1 = 0$ er $0.41 > 0.05$, dvs. H_0 accepteres

Fit simpel lineær regressions model i R

Nu tager vi alle observationer med.
Er der sammenhæng?

```
## Er der signifikant lineær sammenhæng (korrelation)?
plot(Air$temperature, Air$logOzone, xlab="Temperatur", ylab="log Ozon")
cor(Air$temperature, Air$logOzone)
```



Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation med ALLE 110 OBSERVATIONER
summary(lm(logOzone ~ temperature, data=Air))

##
## Call:
## lm(formula = logOzone ~ temperature, data = Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6303 -0.3331  0.0115  0.3455  1.4843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.49997    0.43485   -3.45    8e-04 ***
## temperature  0.06345    0.00554   11.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.54 on 108 degrees of freedom
## Multiple R-squared:  0.549,Adjusted R-squared:  0.544
## F-statistic:  131 on 1 and 108 DF,  p-value: <2e-16
```

Er der signifikant korrelation mellem `logOzone` og `temperature` på 5% signifikansniveau?

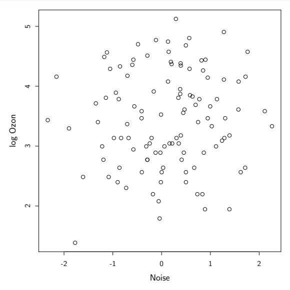
A: Ja B: Nej C: Ved ikke

Svar A: Ja, da p-værdien for $H_0: \beta_1 = 0$ er $2 \cdot 10^{-16} < 0.05$, dvs. H_0 forkastes

Tilføj en vektor med tilfældige værdier

Simuler 110 *uafhængige stokastiske variabler* og tilføj.
Er der sammenhæng?

```
## Tilføj en vektor med normalfordelte tilfældige værdier
## Er der signifikant lineær sammenhæng (korrelation)?
Air$noise <- rnorm(nrow(Air))
plot(Air$noise, Air$logOzone, xlab="Noise", ylab="log Ozon")
cor(Air$noise, Air$logOzone)
```



Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Test om der er signifikant korrelation med ALLE 110 OBSERVATIONER
summary(lm(logOzone ~ noise, data=Air))

##
## Call:
## lm(formula = logOzone ~ noise, data = Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9409 -0.5705 -0.0068  0.6247  1.6657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4396     0.0775   44.38   <2e-16 ***
## noise         0.0634     0.0820    0.77     0.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.81 on 108 degrees of freedom
## Multiple R-squared:  0.0055,Adjusted R-squared:  -0.0037
## F-statistic: 0.598 on 1 and 108 DF,  p-value: 0.441
```

Er der signifikant korrelation mellem `logOzone` og `noise` på 5% signifikansniveau?

A: Ja B: Nej C: Ved ikke

Svar B: Nej, da p-værdien for $H_0 : \beta_1 = 0$ er $0.44 > 0.05$, dvs. H_0 accepteres

Simpel lineær regressionsmodel til de to andre variabler

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

```
## Simpel lineær regressionsmodel med vindhastigheden
plot(Air$wind, Air$logOzone, xlab="Vindhastighed", ylab="log Ozon")
cor(Air$wind, Air$logOzone)
summary(lm(logOzone ~ wind, data=Air))

## Simpel lineær regressionsmodel med indstrålingen
plot(Air$radiation, Air$logOzone, ylab="log Ozon", xlab="Indstraaling")
cor(Air$radiation, Air$logOzone)
summary(lm(logOzone ~ 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

Vi har altså $i = 1, 2, \ldots, n$ observationer og tilsvarende Y_i og ε_i stokastiske 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

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

og trækkes fra

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

Bemærk, ofte bruges $\hat{\varepsilon}_i$ for residual istedet for e_i.

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) standardafvigelse er

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

Mindste kvadraters metode

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ 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 simpelthen gange nogle matricer sammen!

Udvid modellen (forward selection)

Forward selection (*ikke beskrevet i bogen*):

- 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, indstråling eller støj input til modellen
summary(lm(logOzone ~ temperature + wind, data=Air))
summary(lm(logOzone ~ temperature + radiation, data=Air))
summary(lm(logOzone ~ temperature + noise, data=Air))
## Tilføj indstråling eller støj input til modellen
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
summary(lm(logOzone ~ temperature + wind + noise, data=Air))
## Udvid yderligere med støj input?
summary(lm(logOzone ~ temperature + wind + radiation + noise, data=Air))
```

Formindsk modellen (model reduction eller backward selection)

Backward selection (*beskrevet i bogen, Method 6.16*):

- Start med den fulde model
- Fjern den mest insignifikante forklarende variabler
- Stop hvis alle prm. estimater er signifikante

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

Modeludvælgelse

Der er ikke noget sikker metode til automatisk at finde den bedste model!

- Det vil kræve subjektive beslutninger at udvælge en model
- Bedste procedure (forward eller backward): det afhænger af forholdene
- Statistiske tests og mål til at sammenligne modeller
- Her i kurset kun backward procedure inkluderet

Simulate an MLR model with 2 inputs

```
aspect3d(c(1,1,1))
axes3d(c('x--','y--','z--'))
mtext3d('x1', edge=c('x--'), line=2)
mtext3d('x2', edge=c('y--'), line=2)
mtext3d('y', edge=c('z--'), line=2)

## Estimate a plane
fit <- lm(y ~ x1 + x2)

## Make predictions for a grid to see the estimated plane
nplot <- 20
x1plot <- seq(min(x1),max(x1),len=nplot)
x2plot <- seq(min(x2),max(x2),len=nplot)
yprd <- outer(x1plot, x2plot, function(x1,x2){predict(fit, data.frame(x1=x1, x2=x2))})
## 'jet.colors' is "as in Matlab", alternatives see ?rainbow
jet.colors <- colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan",
"#7FFF7F", "yellow", "#FF7F00", "red", "#7F0000"))
## Use 100 different colors
colors <- jet.colors(100)
## Set the colors for z values
color <- colors[(yprd-min(yprd))/(max(yprd)-min(yprd))*100]
rgl.viewpoint(fov=40, theta=0, phi=-90)
## Make a surface with jet colors and grid
surface3d(x1plot, x2plot, yprd, color=color, alpha=0.5)
surface3d(x1plot, x2plot, yprd, front="lines", back="lines", alpha=0.5)
```

Spørgsmål om MLR estimat (socrative.com-ROOM:PBAC)

Hvordan ligger estimaterne af β_0, β_1 og β_2?

A: $\hat{\beta}_0 < 0$, $\hat{\beta}_1 < 0$ og $\hat{\beta}_2 < 0$
B: $\hat{\beta}_0 > 0$, $\hat{\beta}_1 > 0$ og $\hat{\beta}_2 > 0$
C: $\hat{\beta}_0 > 0$, $\hat{\beta}_1 > 0$ og $\hat{\beta}_2 < 0$
D: $\hat{\beta}_0 < 0$, $\hat{\beta}_1 > 0$ og $\hat{\beta}_2 > 0$
E: $\hat{\beta}_0 > 0$, $\hat{\beta}_1 < 0$ og $\hat{\beta}_2 > 0$

Svar C: $\hat{\beta}_0 > 0$, $\hat{\beta}_1 > 0$ og $\hat{\beta}_2 < 0$
Planet skærer y-aksen over 0
Planet går op når x_1 går op
Planet går ned når x_2 går op

Spørgsmål om modelreduktion (socrative.com-ROOM:PBAC)

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9195 -0.1555  0.0104  0.1465  0.6304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0528     0.2285   -0.23   0.8201
## x1           -0.7357     0.3034   -2.42   0.0275 *
## x2            0.2618     0.2937    0.89   0.3859
## x3            1.1817     0.3553    3.33   0.0043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.37 on 16 degrees of freedom
## Multiple R-squared:  0.507,Adjusted R-squared:  0.414
## F-statistic: 5.48 on 3 and 16 DF,  p-value: 0.00878
```

Skal modellen reduceres i backward selection step?

A: Nej B: Ja, x_1 skal væk C: Ja, x_2 skal væk D: Ja, x_3 skal væk

Svar C: Ja, x_2 skal væk, den er ikke signifikant forskellig fra 0 og mest insignifikant

Spørgsmål om σ på afvigelserne (socrative.com-ROOM:PBAC)

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9195 -0.1555  0.0104  0.1465  0.6304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0528     0.2285   -0.23   0.8201
## x1           -0.7357     0.3034   -2.42   0.0275 *
## x2            0.2618     0.2937    0.89   0.3859
## x3            1.1817     0.3553    3.33   0.0043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.37 on 16 degrees of freedom
## Multiple R-squared:  0.507,Adjusted R-squared:  0.414
## F-statistic: 5.48 on 3 and 16 DF,  p-value: 0.00878
```

Hvad er den estimerede standard deviation på afvigelserne $\hat{\sigma}$?

A: $\hat{\sigma} = 0.2285$ B: $\hat{\sigma} = 0.0104$ C: $\hat{\sigma} = 0.37$ D: Ved ikke

Svar C: $\hat{\sigma} = 0.37$

Residual analyse (model kontrol)

- Model kontrol: Analyser residualerne for at checke at forudsætningerne er opfyldt

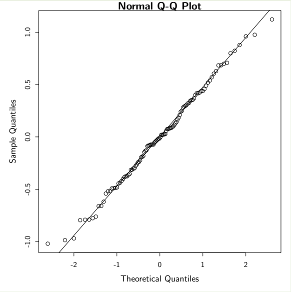
- $\varepsilon_i \sim N(0, \sigma^2)$ og er independent and identically distributed (i.i.d.)
  - Husk: ε_i er afvigelsen (en stokastisk variabel)
  - Husk: $e_i = \hat{\varepsilon}_i$ er residualet (realisationen eller observationen af afvigelsen)

- Samme som for simpel lineær model, dog også plot med residualer vs. inputs

Antagelse om normalfordelte residualer

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

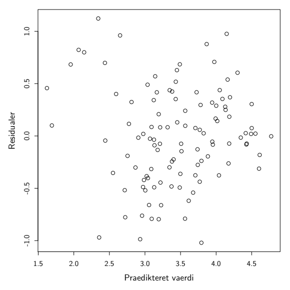
```
## Gem det udvalgte fit
fitSel <- lm(logOzone ~ temperature + wind + radiation,data=Air)

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



Antagelse om identisk distribution

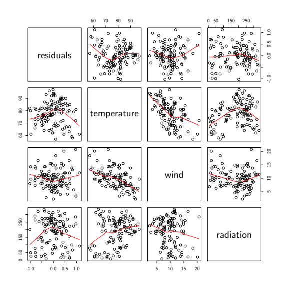
Plot residualerne (e_i) mod de prædikterede (fittede) værdier ($\hat{y}_i$)

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



Plot residualer mod de forklarende variabler

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



Kan måske forbedres med ikke-lineær sammenhæng temperature eller vindhastighed.

Kurvelineær (Curvilinear)

Hvis vi ønsker at estimere en model af typen

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

kan vi benytte multipel lineær regression i modellen

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

hvor

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

og benytte samme metoder som ved multipel lineær regression.

Udvid ozon modellen med passende kurvelineær regression

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

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

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

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

## Her kunne man prøve at udvide yderligere
fitWindSqTemperatureSq <- lm(logOzone ~ temperature + temperatureSq + wind + windSq + radiation, data=Air)
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 transformeres tilbage
CI <- exp(predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95))
PI <- exp(predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95))

## Plot them
Air$ozone <- exp(Air$logOzone)
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 indstr
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

```
## Lav en variabel, som er meget korreleret f.eks. endnu en vindmåling
set.seed(367)
Air$wind2 <- Air$wind + rnorm(nrow(Air), sd=1)
cor(Air$wind, Air$wind2)
plot(Air$wind, Air$wind2)
## Tilføj den til modellen
fitWind2 <- lm(logOzone ~ temperature + wind + wind2 + radiation, data=Air)
summary(fitWind2)

## Sammenlign med modellen med kun den ene
fitWind <- lm(logOzone ~ temperature + wind + radiation, data=Air)
summary(fitWind)
```

Få feedback fra TA om jeres projekt 1