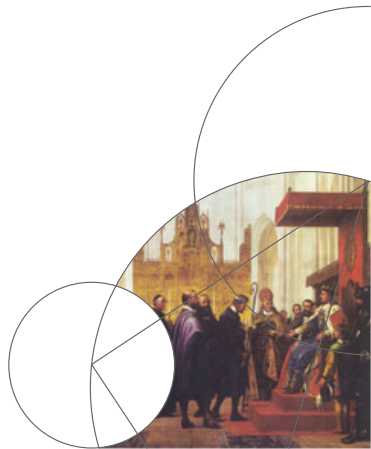




Det Natur- og Biovidenskabelige Fakultet

Multipel lineær regression

Anders Tolver
Institut for Matematiske Fag



Dagens program

Husk:

Upload besvarelse af den frivillige afleveringsopgave i Absalon!

Vi gennemgår lærebogens Kapitel 8.1

- Multipel lineær regression
- Begrebet (multi)kollinearitet
- Specialtilfælde: kvadratisk og kubisk regression (læses selv: slides 24-33 + R-program)

Opsummering på kursusuge 5 (videoer)

- Gennemgang af slides 27-34 fra mandag d. 4/10-2021 (allerede på Absalon)
- Gennemgang af Quiz 5 (omkring weekenden)
- Eventuelle hængepartier



Overblik

Vi skal have „udfyldt“ følgende skema over modeller (rækker) og statistiske begreber (søjler):

	Intro	Model	Est.+SE	KI	Test	Kontrol	Præd.
En stikprøve	✓	✓	✓	✓	✓	✓	✓
Ensidet ANOVA	✓	✓	✓	✓	✓	✓	✓
Lineær regr.	✓	✓	✓	✓	✓	✓	✓
To stikprøver	✓	✓	✓	✓	✓	✓	✓
Multipel regr.	nu	nu	nu	nu	nu	nu	nu
Tosidet ANOVA							



Multipel lineær regression



Eksempel 8.1: Volumen af kirsebærtræer

Data fra 31 kirsebærtræer, ligger som **trees** i *isdals*.

- Diameter i brysthøjde. Meget nem at måle
- Højde. Nogenlunde nem at måle
- Volumen. Kan kun måles efter fældning

NB: Variablen med diameter hedder `girth` (omkreds) i datasættet, men ifølge `?trees` indeholder den faktisk diameteren.



Eksempel 8.1: Volumen af kirsebærtræer

Data fra 31 kirsebærtræer, ligger som **trees** i *isdals*.

- Diameter i brysthøjde. Meget nem at måle
- Højde. Nogenlunde nem at måle
- Volumen. Kan kun måles efter fældning

NB: Variablen med diameter hedder `girth` (omkreds) i datasættet, men ifølge `?trees` indeholder den faktisk diameteren.

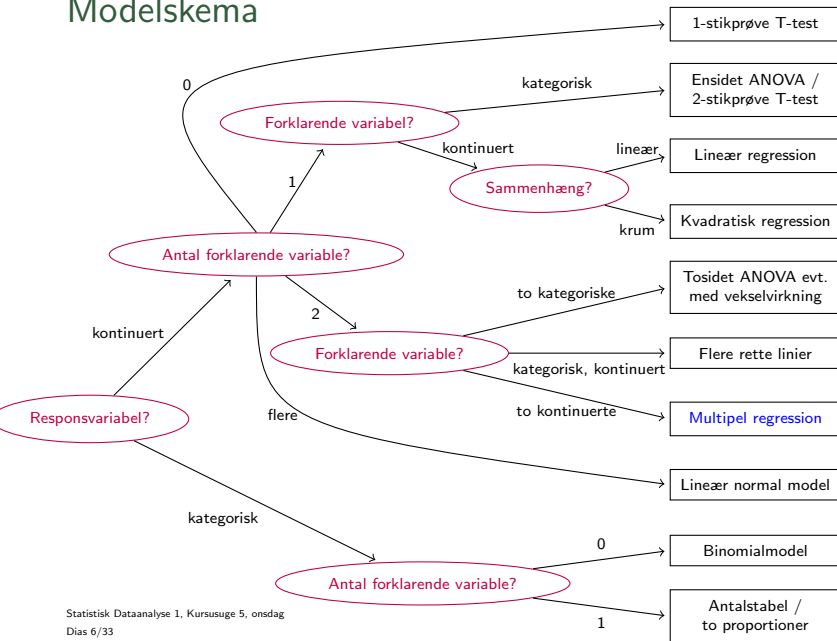
Spørgsmål:

- Bestem en **god prædiktionsmodel** for volume
- **Kan det betale sig også at måle højden?** Bidrager den faktisk med til at beskrive volumen (når vi har diameter)?

Respons? Forklarende variable? Hvor er vi i modelskemaet?

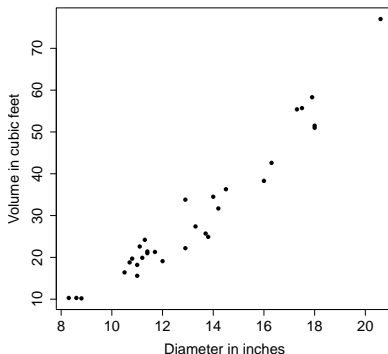
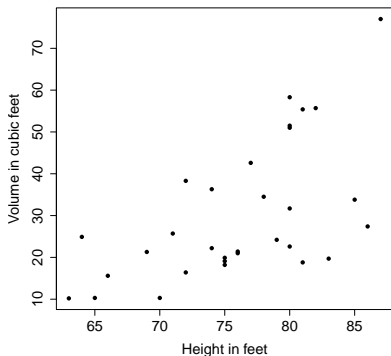


Modelskema



(Simpel) Lineær regression

Simpel lineær regression beskriver sammenhængen mellem responsvariabel og **én** kontinuert forklarende variabel:



Lineær regression

Regression af volumen på **højde**:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-87.12361	29.2731221	-2.976232	0.0058346689
## Height	1.54335	0.3838693	4.020509	0.0003783823



Lineær regression

Regression af volumen på **højde**:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-87.12361	29.2731221	-2.976232	0.0058346689
## Height	1.54335	0.3838693	4.020509	0.0003783823

Regression af volumen på **diameter**:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-36.943459	3.365145	-10.97827	7.621449e-12
## Girth	5.065856	0.247377	20.47829	8.644334e-19

Men hvad hvis **begge** variable har en betydning for volumen?



Multipel lineær regression

Multipel lineær regression: $d \geq 2$ kvantitative forkl. variable.

Statistisk model:

$$y_i = \alpha + \beta_1 \cdot x_{i1} + \cdots + \beta_d \cdot x_{id} + e_i$$

med iid. restled $e_i \sim N(0, \sigma^2)$ som sædvanlig.



Multipel lineær regression

Multipel lineær regression: $d \geq 2$ kvantitative forkl. variable.

Statistisk model:

$$y_i = \alpha + \beta_1 \cdot x_{i1} + \cdots + \beta_d \cdot x_{id} + e_i$$

med iid. restled $e_i \sim N(0, \sigma^2)$ som sædvanlig.

Når $d = 2$ er der tre middelværdiparametre:

- α skæring (intercept) med y -aksen når $x_{i1} = x_{i2} = 0$.
- β_1 og β_2 er **partielle hældninger**, dvs. ændring i y hvis en var. ændres med 1, og den anden forklarende var. "fastfryses".

Desuden er spredningen σ som sædvanlig en ukendt parameter.



Multipel lineær regression: Statistisk inferens

Vi **kan allerede det hele**: Estimation, modelkontrol, hypotesetest, konfidens- og prædiktionsintervaller fra uge 3–4.



Multipel lineær regression: Statistisk inferens

Vi **kan allerede det hele**: Estimation, modelkontrol, hypotesetest, konfidens- og prædiktionsintervaller fra uge 3–4.

R: Tilføj yderligere led til `lm`, med `+` imellem, fx:

```
multipel1 <- lm(Volume ~ Height + Girth, data=trees)
summary(multipel1)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-57.9876589	8.6382259	-6.712913	2.749507e-07
## Height	0.3392512	0.1301512	2.606594	1.449097e-02
## Girth	4.7081605	0.2642646	17.816084	8.223304e-17

Fortolkning af parameterestimer?



Er det en fornuftig model?

Er det en **fornuftig model**?

- Modelkontrol OK?
- Fra et mere "teoretisk" synspunkt? Modeller for træer?



Er det en fornuftig model?

Er det en **fornuftig model**?

- Modelkontrol OK?
- Fra et mere "teoretisk" synspunkt? Modeller for træer?

Naive modeller for træer:

- Træets form kan approksimeres med en **cylinder**
- Træets form kan approksimeres med en **kegle**



Transformation

De naive modeller:

- Cylinder med diameter d , højde h : volumen, $v = ?$
- Kegle med grundfladediameter d og højde h : vol. $v = \frac{\pi}{12} \cdot h \cdot d^2$



Transformation

De naive modeller:

- Cylinder med diameter d , højde h : volumen, $v = ?$
- Kegle med grundfladediameter d og højde h : vol. $v = \frac{\pi}{12} \cdot h \cdot d^2$

I begge tilfælde:

$$v = \text{konstant} \cdot h \cdot d^2$$

Træer er hverken cylindre eller kegler, men vi kan gøre modellen mere **fleksibel** ved at tillade andre potenser:

$$v = c \cdot h^{\beta_1} \cdot d^{\beta_2}$$



Transformation

De naive modeller:

- Cylinder med diameter d , højde h : volumen, $v = ?$
- Kegel med grundfladediameter d og højde h : vol. $v = \frac{\pi}{12} \cdot h \cdot d^2$

I begge tilfælde:

$$v = \text{konstant} \cdot h \cdot d^2$$

Træer er hverken cylindre eller kegler, men vi kan gøre modellen mere **fleksibel** ved at tillade andre potenser:

$$v = c \cdot h^{\beta_1} \cdot d^{\beta_2}$$

Efter log-transformation fås en **multipl lineær regression**:

$$\log v_i = \alpha + \beta_1 \cdot \log h_i + \beta_2 \cdot \log d_i + e_i$$



R

```
multipe12 <- lm(log(Volume) ~ log(Height) + log(Girth)
                , data = trees)
summary(multipe12)$coefficients

##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) -6.631617 0.79978973 -8.291701 5.057138e-09
## log(Height)  1.117123 0.20443706  5.464388 7.805278e-06
## log(Girth)   1.982650 0.07501061 26.431592 2.422550e-21

newData <- data.frame(Girth = 14, Height = 80)
predict(multipe12, newData, interval = "p")

##      fit      lwr      upr
## 1 3.495974 3.32548 3.666467
```

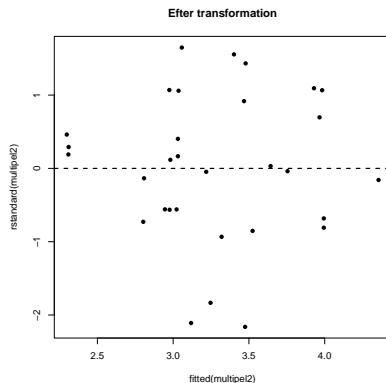
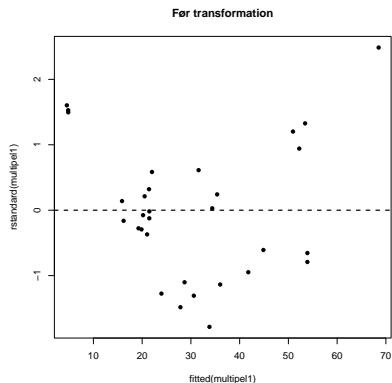


Spørgsmål

- Er modelantagelserne OK?
- Træ med diameter 14 og højde 80. Hvad er et fornuftigt bud på volumen? Prædiktionsinterval?
- Fortolkning af β_1 og β_2 ?
- Kan det betale sig også at måle højden? Bidrager den faktisk med til at beskrive volumen (når vi har diameter)?
- De naive modeller havde begge $\beta_1 = 1$, $\beta_2 = 2$. Passer det med data?



Residualplot for de to modeller



Ser bedst ud efter log-transformation.

Er sammenhængen som i de naive modeller?

Statistiske modeller:

- Generel model: $\log v_i = \alpha + \beta_1 \cdot \log h_i + \beta_2 \cdot \log d_i + e_i$
- Naive modeller: $\log v_i = \alpha + 1 \cdot \log h_i + 2 \cdot \log d_i + e_i$

De naive model er (som vi vidste) specialtilfælde af den generelle model. Svarer til **hypotesen**

$$H_0: \beta_1 = 1, \beta_2 = 2$$



Er sammenhængen som i de naive modeller?

Statistiske modeller:

- Generel model: $\log v_i = \alpha + \beta_1 \cdot \log h_i + \beta_2 \cdot \log d_i + e_i$
- Naive modeller: $\log v_i = \alpha + 1 \cdot \log h_i + 2 \cdot \log d_i + e_i$

De naive model er (som vi vidste) specialtilfælde af den generelle model. Svarer til **hypotesen**

$$H_0: \beta_1 = 1, \beta_2 = 2$$

- **Hver for sig** kan $H_0: \beta_1 = 1$ og $H_0: \beta_2 = 2$ testes med t -test
- **Hele hypotesen** kan testes med F -test (med 2 df i tælleren)
- F -testet giver $p = 0.85$, så H_0 accepteres. Potenserne fra de naive modeller er OK.



R: test for om $\beta_1 = 1, \beta_2 = 2$

```
naiv <- lm(log(Volume) ~ offset(1*log(Height) + 2*log(Girth))
           , data=trees)
anova(naiv, multipe12)

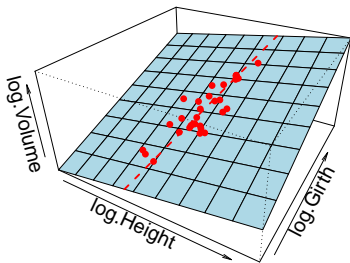
## Analysis of Variance Table
##
## Model 1: log(Volume) ~ offset(1 * log(Height) + 2 * log(Girth))
## Model 2: log(Volume) ~ log(Height) + log(Girth)
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1       30 0.18769
## 2       28 0.18546  2 0.0022224 0.1678 0.8464
```



Multikollinearitet i multipel lineær regression



Fortolkning og kollinearitet



- Model $y_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + e_i$
- β_1, β_2 : **partielle hældninger**, dvs. ændringen i y hvis andre variable fastfryses.
- Hvis x_1 og x_2 er **afhængige**, så er det svært at adskille effekten af dem. Dette kaldes **kollinearitet**.

Potentielle problemer ved multikollinearitet

Tegn på multikollinearitet:

- **Unaturlige estimer**, f.eks. forkert fortegn.
- Hverken β_1 eller β_2 er signifikante, men begge led ikke kan undværes på samme tid

Pas på med fortolkningerne.

Måske giver det slet ikke mening af tale om ændringen i en variabel, mens de andre fastholdes...



Eksempel: Timeløn vs. uddannelse, erfaring og alder

Data:

- Lille uddrag fra The Current Population Survey (CPS, USA, 1985)
- 52 observationer fra kvinder, som alle arbejder i professionskategorien "other".
- Respons: Timeløn (USD)
- Forklarende variable: samlet længde uddannelse, alder, erfaring (alle i år)



Eksempel: Timeløn vs. uddannelse, erfaring og alder

```
> summary(lm(wage ~ edu + exper + age, data=myData))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-15.4931	6.6457	-2.331	0.024 *
edu	0.7059	0.8524	0.828	0.412
exper	-0.6247	0.8723	-0.716	0.477
age	0.6775	0.7964	0.851	0.399



Eksempel: Timeløn vs. uddannelse, erfaring og alder

```
> summary(lm(wage ~ edu + exper + age, data=myData))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-15.4931	6.6457	-2.331	0.024 *
edu	0.7059	0.8524	0.828	0.412
exper	-0.6247	0.8723	-0.716	0.477
age	0.6775	0.7964	0.851	0.399

Spørgsmål:

- Hvad er fortolkningen af fortegnet for erfaring (exper)?
- Er der signifikant effekt af uddannelse (edu) hhv. erfaring (exper) hhv. alder (age)?



Eksempel: Timeløn vs. uddannelse, erfaring og alder

```
> summary(lm(wage ~ exper + edu, data=myData))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-11.85350	5.07118	-2.337	0.0235	*
exper	0.11552	0.06237	1.852	0.0700	.
edu	1.38007	0.31307	4.408	5.68e-05	***



Eksempel: Timeløn vs. uddannelse, erfaring og alder

```
> summary(lm(wage ~ exper + edu, data=myData))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-11.85350	5.07118	-2.337	0.0235	*
exper	0.11552	0.06237	1.852	0.0700	.
edu	1.38007	0.31307	4.408	5.68e-05	***

Spørgsmål:

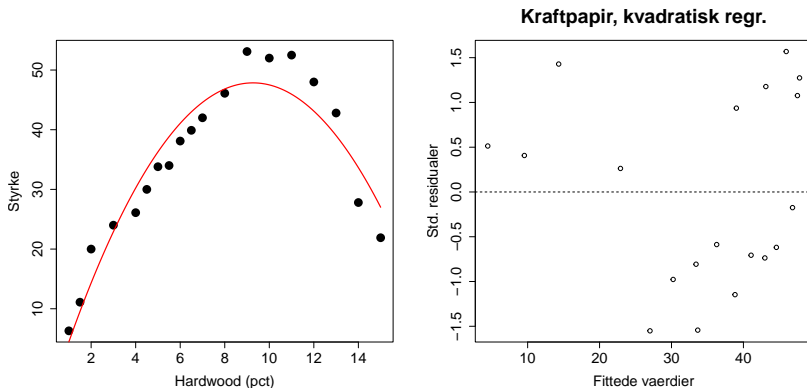
- Hvad skete der med fortegnet for erfaring?
- Er der signifikante effekter?
- Kan vi forklare "hvad der sker"?



Polynomiell regression



Eksempel 8.3: Kraftpapir (sidste uge)



- Kvadratisk regression: $str_i = \alpha + \beta_1 \cdot hw_i + \beta_2 \cdot hw_i^2 + e_i$
- Måske ikke helt tilfredse: Fanger ikke toppen, asymmetri

Polynomiell regression

Kvadratisk regression: $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + e_i$

Specialtilfælde af multipel lineær regression:

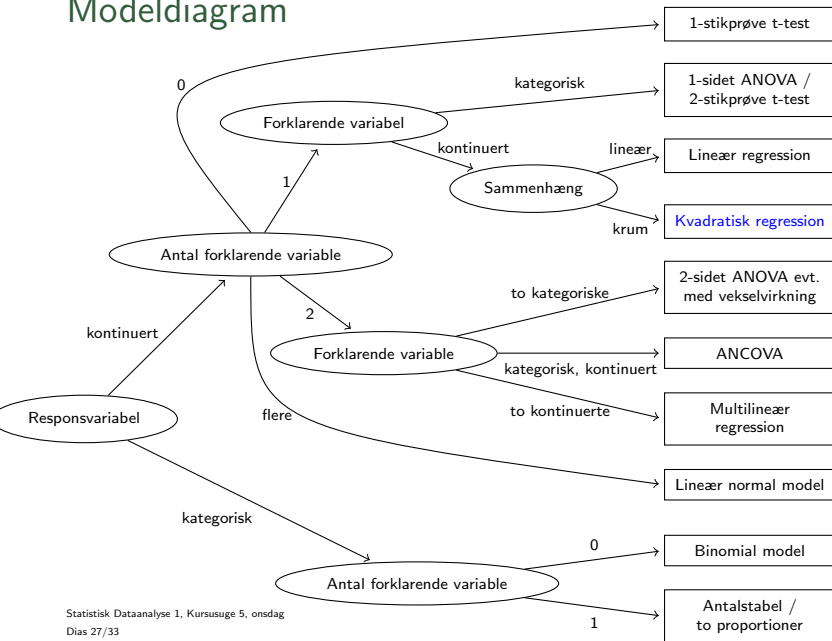
- De forklarende variable er potenser af samme variabel
- Kan ikke fortolke estimater som i multipel lineær regression.
Hvorfor ikke?

Check modeldiagram.

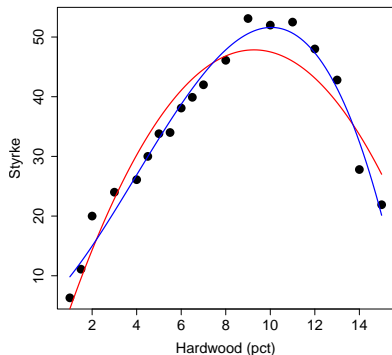
Kan udvide modellen med **flere potenser** → polynomiell regression



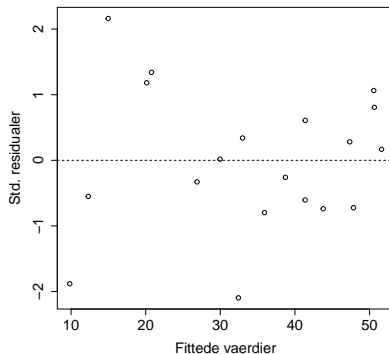
Modeldiagram



Eksempel 8.3: Kraftpapir



Kraftpapir, kubisk regr.



- Kubisk regression: $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + \beta_3 \cdot \text{hw}_i^3 + e_i$
- Residualplottet ser umiddelbart bedre ud

Hypotesetest

I sidste uge:

- Kvadratisk regression: $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + e_i$
- Hypotese, $H_0 : \beta_2 = 0$. Testet gav $T_{\text{obs}} = -10.3$, $p = 1.9 \cdot 10^{-8}$
- Konklusion: Kvadratisk model beskriver data bedre end lineær model



Hypotesetest

I sidste uge:

- Kvadratisk regression: $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + e_i$
- Hypotese, $H_0 : \beta_2 = 0$. Testet gav $T_{\text{obs}} = -10.3$, $p = 1.9 \cdot 10^{-8}$
- Konklusion: Kvadratisk model beskriver data bedre end lineær model

Tilsvarende:

- Kubisk regression: $\text{str}_i = \alpha + \beta_1 \cdot \text{hw}_i + \beta_2 \cdot \text{hw}_i^2 + \beta_3 \cdot \text{hw}_i^3 + e_i$
- Hypotese, $H_0 : \beta_3 = 0$. Testet giver $T_{\text{obs}} = 5.6$, $p = 4.7 \cdot 10^{-5}$
- Konklusion: Kubisk model beskriver data bedre end kvadratisk model



Konklusion

Kraftpapir:

- Den kubiske model beskriver data signifikant bedre end kvadratisk model
- Den kvadratiske model har dog **simplere** fortolkning (godt)
- Begge modeller har den vigtigste feature: der er en **optimal træmængde** der giver den største forventede styrke



R: kvadratisk regressionsmodel

```
kvadreg <- lm(strength ~ hardwood + I(hardwood^2)  
              , data = paperstr)  
summary(kvadreg)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-6.6741916	3.39970751	-1.963166	6.725203e-02
## hardwood	11.7640057	1.00278222	11.731366	2.854174e-09
## I(hardwood^2)	-0.6345492	0.06178832	-10.269727	1.894349e-08



R: kubisk regressionsmodel

```
cubicreg <- lm(strength ~ hardwood + I(hardwood^2)
               + I(hardwood^3), data = paperstr)
summary(cubicreg)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	5.6483950	2.954663227	1.911688	7.521268e-02
## hardwood	3.5784894	1.565854129	2.285327	3.726535e-02
## I(hardwood^2)	0.6536355	0.231329713	2.825558	1.278280e-02
## I(hardwood^3)	-0.0551876	0.009788835	-5.637811	4.721725e-05



Potentielle problemer med polynomiel regression

Vær **ekstra forsigtig med ekstrapolation** (prædiktion udover observationsområdet)

Pas på med at **"overfitte"**, dvs. tilpasse modellen **for godt**, således at resultatet ikke vil være reproducerbart.

- Kan tilpasse kurven fuldstændigt til data hvis vi bruger nok $n - 1$ potenser. Ikke reproducerbart
- Modellen skal fange egentlige features, men ikke tilfældige udsving.

Der findes andre metoder til kurvetilpasning (ikke StatDat1)

