

# HD Dataanalyse

*Note 9: Regression med ÉN  
forklarende variabel*

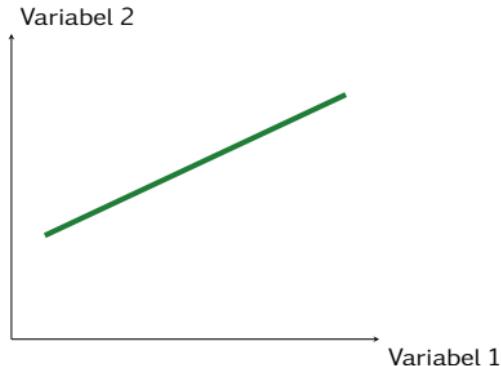
Copenhagen Business School

## EMNE I DETTE NOTESÆT

Model  
Estimation  
Forklaringsgrad  
Hypotesetest  
OPSUMMERING

Når vi betragter **to KVANTITATIVE variable** i et datamateriale, er vi interesseret i at vurdere, om der er en **sammenhæng mellem dem** eller ej.

Den simplest tænkelige sammenhæng mellem to kvantitative variable er en ret linje:



I denne note skal vi se på, hvordan man kan beregne den rette linje, der bedst beskriver sammenhængen mellem to kvantitative variable.

Dernæst skal vi se på, hvilke konklusioner vi ud fra den beregnede linje kan drage om sammenhængen mellem de to variable.

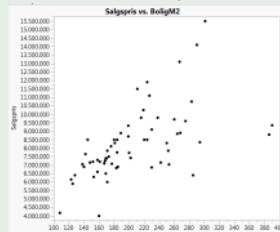
# EMNE I DETTE NOTESÆT

Model  
Estimation  
Forklæringsgrad  
Hypotesetest  
OPSUMMERING

## Eksempel: Boligpriser

Vi ser igen på datamaterialet med salgsinformation om boliger solgt i perioden 2014-2015. Salgsprisen på en bolig må formodes at afhænge af en række forskellige karakteristika så som beliggenhed, størrelse, alder etc.

Hvis vi eksempelvis sammenholder de solgte boligers pris med deres areal, vil vi regne med, at der er en sammenhæng. Tegner vi pris og boligareal op i en figur, så ser vi ikke overraskende, at der er en tendens til, at større boligareal hænger sammen med en højere salgspris. Vi har her for overskuelighedens skyld begrænset os til at se på villaer solgt på Frederiksberg.



Spørgsmålet er nu, hvordan sammenhængen mellem salgspris og boligareal mere præcist er? Umiddelbart er det mest enkle at antage, at sammenhængen mellem salgspris og areal kan beskrives ved en ret linje.

Selv om punkterne i figuren ikke ligger præcist på en ret linje, kan linjen måske alligevel godt give en grov sammenfatning af sammenhængen mellem salgspris og boligareal. I så fald er spørgsmålet, hvilken ret linje der bedst beskriver sammenhængen mellem salgspris og boligareal?

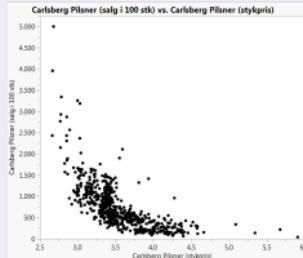
# EMNE I DETTE NOTESÆT

Model  
Estimation  
Forklæringsgrad  
Hypotesetest  
OPSUMMERING

## Eksempel: Ølsalg

Vi ser igen på datamaterialet med salgsinformation om forskellige ølmærker i udvalgte supermarkeder. Antallet af solgte øl må formodes at afhænge af prisen på øl, idet vi vil regne med, at der sælges flere øl, desto lavere prisen er.

Tegner vi eksempelvis det ugentlige antal solgte Carlsberg øl (33 cl glasflaske) op mod gennemsnitsprisen pr. øl, så ser vi ikke overraskende, at der er en tendens til, at lavere priser hænger sammen med et højere salg (i stk.).



Spørgsmålet er nu, hvordan sammenhængen mellem den omsatte mængde øl og stykprisen mere præcist er? Umiddelbart er det mest enkle at antage, at sammenhængen mellem den omsatte mængde og stykprisen kan beskrives ved en ret linje.

Selv om punkterne i figuren ikke ligger præcist på en ret linje, kan den måske alligevel godt give en grov sammenfatning af sammenhængen mellem mængde og pris. I så fald er spørgsmålet, hvilken ret linje der bedst beskriver sammenhængen mellem mængde og pris?

▶ JMP-video [Graph -> Graph Builder]

# INDHOLDSFORTEGNELSE

---

1 Model

2 Estimation

3 Forklaringsgrad

4 Hypotesetest

5 OPSUMMERING

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

## 1 Model

2 Estimation

3 Forklaringsgrad

4 Hypotesetest

5 OPSUMMERING

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

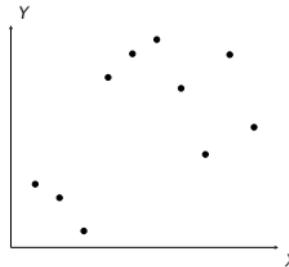
Når vi...

- betragter to **kvantitative** variable  $X$  og  $Y$  i et datamateriale
- og formoder at variablen  $Y$  **afhænger af** variablen  $X$

vil vi gerne undersøge, hvordan sammenhængen mellem variablene  $X$  og  $Y$  helt præcist er.

Variablen  $Y$  kaldes for **responsvariablen**, og variablen  $X$  kaldes for den **forklarende variabel**.

Ideen er, at vi gerne vil lave en model, der kan **forklare** værdien af  $Y$  ud fra værdien af  $X$ , eller med andre ord, forklare hvordan  $Y$  påvirkes af, dvs. **responderer** på, værdien af  $X$ .

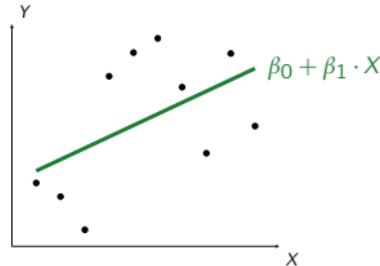


Sammenhængen mellem  $X$  og  $Y$  kan være vilkårligt kompliceret, men den simplest tænkelige **model** til beskrivelse af sammenhængen mellem  $X$  og  $Y$  er en ret linje. Hvis  $X$  og  $Y$  ligger *præcist* på en ret linje, kan vi skrive sammenhængen mellem dem som

$$Y = \beta_0 + \beta_1 \cdot X$$

hvor  $\beta_0$  og  $\beta_1$  er ukendte parametre, som vi kan estimere på baggrund af vores datamateriale.

$\beta_0$  angiver, hvor den rette linje skærer 2. aksen, mens  $\beta_1$  angiver hvor meget  $Y$  ændrer sig, hver gang  $X$  ændrer sig med 1.



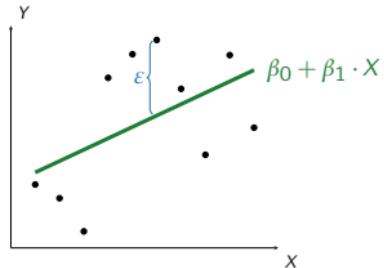
I praksis vil vi stort set aldrig se, at to variable  $X$  og  $Y$  ligger præcist på en ret linje. Derfor tillader vi en vis mængde tilfældig variation omkring den rette linje.

Vores model for sammenhængen mellem  $Y$  og  $X$  blive derfor

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

hvor  $\varepsilon$  kaldes for modellens **residual**.

Residualet beskriver den del af variationen i  $Y$ , som ikke kan beskrives ved hjælp af den rette linje. Ordet "residual" betyder "det, der er til overs", dvs. at  $\varepsilon$  beskriver den del af værdien af  $Y$ , der er til overs/ikke kan forklares af værdien af  $X$ .



## Vores model

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

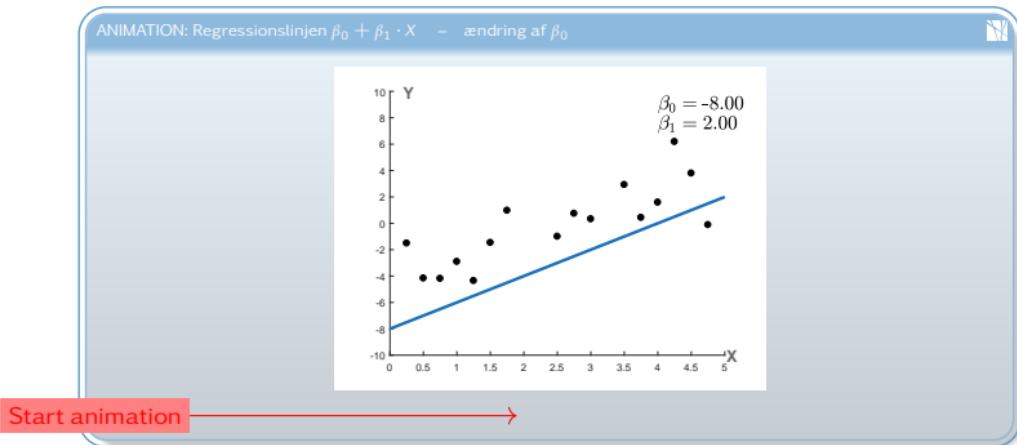
til forklaring af værdierne af variablen  $Y$  ud fra kendskab til værdierne af variablen  $X$  kalder vi for en **simpel lineær regressionsmodel**.

Logikken bag betegnelsen “simpel lineær regression” er følgende:

- Ordet **regression** henviser til, at modellen forsøger at forklare  $Y$  ud fra  $X$ , dvs. forsøger at “føre  $Y$  tilbage til  $X$ ”  
 (“regressere” betyder at “føre tilbage”)
- Ordet **lineær** henviser til, at modellen forklarer  $Y$  ud fra en lineær sammenhæng med  $X$  (dvs. en ret linje)
- Ordet **simpel** henviser til, at modellen forklarer  $Y$  ud fra kun én enkelt variabel  
(vi skal i Note 10 se, hvordan vi kan udvide modellen til at forklare  $Y$  ud fra mere end én variabel, og denne model kaldes så for en *multipel* lineær regressionsmodel)

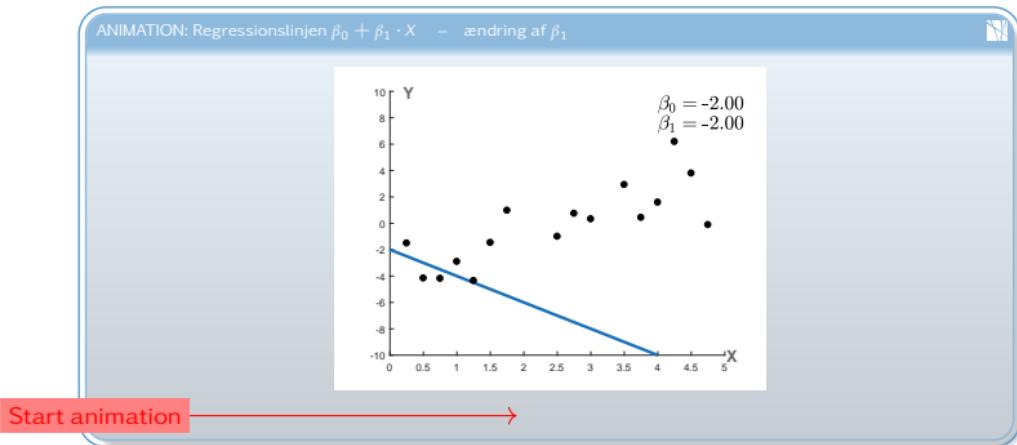
Nedenstående animation viser, hvordan regressionslinjen  $\beta_0 + \beta_1 \cdot X$  ændres, når parameteren  $\beta_0$  ændrer sig.

$\beta_0$  angiver, hvor den rette linje skærer 2. aksen.



Nedenstående animation viser, hvordan regressionslinjen  $\beta_0 + \beta_1 \cdot X$  ændres, når parameteren  $\beta_1$  ændrer sig.

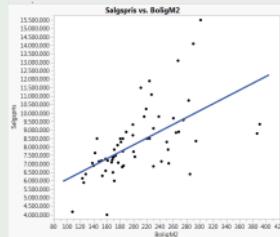
$\beta_1$  angiver hældningen på regressionslinjen.



## Eksempel: Boligpriser

Når vi ser på, hvordan salgsprisen på en bolig afhænger af boligens areal, bruger vi variablen salgspris som responsvariabel Y og variablen boligareal som forklarende variabel X.

Ved hjælp af JMP kan vi få indtegnet den rette linje  $\beta_0 + \beta_1 \cdot X$ , der bedst beskriver sammenhængen mellem Y og X:



Det er klart, at datamaterialet ikke passer perfekt med den indtegnede rette linje, omend linjen ser ud til at fange en vis del af sammenhængen mellem X og Y.

Det vi skal se på i resten af denne note er blandt andet...

- hvordan vi finder frem til den linje, der bedst beskriver sammenhængen mellem Y og X
- hvor præcist/usikkert bestemt den estimerede rette linje er
- hvilke konklusioner vi kan drage på baggrund af den estimerede rette linje

▶ JMP-video [Graph -> Graph Builder]

Model

Estimation

Forklaringsgrad

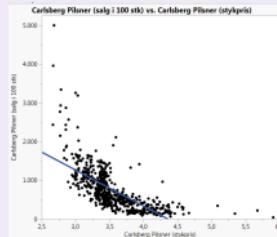
Hypotesetest

OPSUMMERING

## Eksempel: Ølsalg

Når vi ser på, hvordan omsætningen af øl afhænger af prisen på øl, bruger vi variablen omsætning som responsvariabel  $Y$  og variablen stykpris som forklarende variabel  $X$ .

Ved hjælp af JMP kan vi få indtegnet den rette linje  $\beta_0 + \beta_1 \cdot X$ , der bedst beskriver sammenhængen mellem  $Y$  og  $X$ :



Det er klart, at datamaterialet ikke passer perfekt med den indtegnede rette linje, omend linjen ser ud til at fange en vis del af sammenhængen mellem  $X$  og  $Y$ .

Afvigelserne mellem datapunkterne og den rette linje er dem, der i modellen beskrives af residaulaet  $\varepsilon$ .

▶ JMP-video [Graph -> Graph Builder]

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

## 1 Model

## 2 Estimation

## 3 Forklaringsgrad

## 4 Hypotesetest

## 5 OPSUMMERING

I vores simple lineære regressionsmodel

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

beskriver  $\varepsilon$  tilfældig/uforklarlig variation omkring den rette linje  $\beta_0 + \beta_1 \cdot X$ . Formelt set antager vi, at residualet  $\varepsilon$  er beskrevet ved en normalfordeling  $N(0, \sigma)$ .

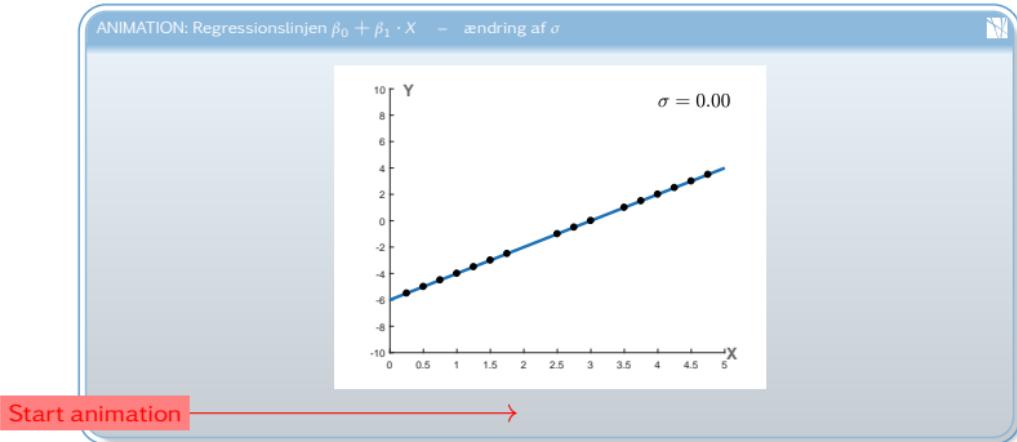
Residualet har en forventet værdi på  $E(\varepsilon) = 0$ , således at den rette linje  $\beta_0 + \beta_1 \cdot X$  beskriver vores forventning til værdien af  $Y$  givet vores kendskab til  $X$ , hvilket vi også skriver som

$$E(Y|X) = \beta_0 + \beta_1 \cdot X$$

Standardafvigelsen  $\sigma$  beskriver hvor meget/lidt vi regner med, at værdierne af  $Y$  vil afvige fra den rette linje  $\beta_0 + \beta_1 \cdot X$ .

Nedenstående animation viser, hvordan sammenhængen mellem regressionslinjen  $\beta_0 + \beta_1 \cdot X$  og data ændres, når parameteren  $\sigma$  ændrer sig.

$\sigma$  angiver, hvor meget datamaterialet varierer omkring/afviger fra den rette linje.



Når vi skal gætte på hvilken ret linje, der på baggrund af datamaterialet bedst beskriver sammenhængen mellem  $X$  og  $Y$ , så foregår det på følgende måde:

- For hver af datamaterialets  $n$  observationer  $i = 1, \dots, n$  (dvs. for hver række i vores JMP-fil) har vi en værdi  $X_i$  og en værdi  $Y_i$
- For hver af datamaterialets  $n$  observationer  $i = 1, \dots, n$  kan vi derfor beregne den kvadrerede afvigelse mellem variablen  $Y$  og modellens rette linje

$$\left( Y_i - (\beta_0 + \beta_1 \cdot X_i) \right)^2$$

- Som estimerede værdier for  $\beta_0$  og  $\beta_1$  vælger vi de værdier, som giver den mindste samlede afvigelse mellem datamateriale og modellens rette linje, dvs. de værdier som minimerer summen af alle de kvadrerede afvigelser

$$\sum_{i=1}^n \left( Y_i - (\beta_0 + \beta_1 \cdot X_i) \right)^2$$

- De estimerede værdier  $\hat{\beta}_0$  og  $\hat{\beta}_1$  angiver dermed den rette linje  $\hat{\beta}_0 + \hat{\beta}_1 \cdot X$  der bedst beskriver sammenhængen mellem  $X$  og  $Y$

## Eksempel: Boligpriser

Når vi estimerer den rette linje, der bedst beskriver sammenhængen mellem boligpris ( $Y$ ) og boligareal ( $X$ )

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

hvor  $\varepsilon$  er normalfordelt  $N(0, \sigma)$ , så finder vi at

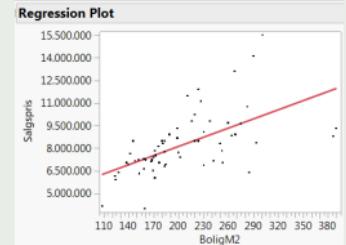
$$\hat{\beta}_0 = 4.069.442 \quad \hat{\beta}_1 = 20.277 \quad \hat{\sigma} = 1.678.393$$

Fortolkningen af de tre parametre er, at...

- for hver gang boligarealet ( $X$ ) stiger med  $1 m^2$ , så stiger den forventede salgspris med  $\hat{\beta}_1 = 20.277$  kr.
- for en bolig med et boligareal på  $0 m^2$ , så er den forventede salgspris  $\hat{\beta}_0 = 4.069.442$  kr.
- salgsprisen på en bolig vil med ca. 95% sandsynlighed variere med  $\pm 2 \cdot \hat{\sigma} = 3.356.786$  kr.

Det er klart, at fortolkningen af  $\hat{\beta}_0$  i dette eksempel bliver en smule aparte, fordi det er svært at forestille sig en bolig med et boligareal på  $X = 0$ .

Det er også klart, at den estimerede rette linje er meget usikkert bestemt (hvilket fremgår af den meget høje værdi for  $\hat{\sigma}$ ), fordi den alene er baseret på 62 observationer.



## Effect Summary

## Summary of Fit

RSquare	0.340119
RSquare Adj	0.329121
Root Mean Square Error	1678393
Mean of Response	8234466
Observations (or Sum Wgts)	62

## Analysis of Variance

Source	DF	Sum of Squares		F Ratio
		Mean Square	F Ratio	
Model	1	8.7117e+13	8.712e+13	30.9255
Error	60	1.6902e+14	2.817e+12	Prob > F
C. Total	61	2.5614e+14		<.0001*

## Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	40694419	878702.7	5.23	<.0001*
BoligM2	20277.306	3446.296	5.56	<.0001*

## Eksempel: Ølsalg

Når vi estimerer den rette linje, der bedst beskriver sammenhængen mellem omsætningen ( $Y$ ) og stykpris ( $X$ )

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

hvor  $\varepsilon$  er normalfordelt  $N(0, \sigma)$ , så finder vi at

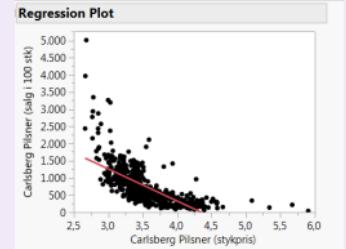
$$\hat{\beta}_0 = 4.052 \quad \hat{\beta}_1 = -930 \quad \hat{\sigma} = 400$$

Fortolkningen af de tre parametre er, at...

- for hver gang prisen ( $X$ ) stiger med 1 kr., så falder den forventede omsætning med  $\hat{\beta}_1 \cdot 100 = 93.000$  stk.
- hvis prisen er 0 kr, så er den forventede omsætning  $\hat{\beta}_0 \cdot 100 = 405.200$  stk.
- omsætningen vil med ca. 95% sandsynlighed variere med  $\pm 2 \cdot \hat{\sigma} \cdot 100 = 80.000$  stk.

Bemærk, at alle værdier her ganges med 100, fordi variablen  $Y$  er opgjort i 100 stk.

Bemærk også, at en prisstigning på 1 kr. er forholdsvis ekstrem. Ser vi i stedet på en prisstigning på eksempelvis 0,10 kr., så siger modellen, at den forventede omsætning vil falde med 0,1 ·  $\hat{\beta}_1 \cdot 100 = 9.300$  stk.

**Effect Summary****Summary of Fit****Analysis of Variance****Parameter Estimates** $\hat{\sigma}$  $\hat{\beta}_0$  $\hat{\beta}_1$

Formelt set er estimationen af parametrene  $\beta_0, \beta_1, \sigma$  baseret på følgende antagelser:

### Resultat [Simpel lineær regression – Antagelser]



Lad  $(X_1, Y_1), \dots, (X_n, Y_n)$  være  $n$  sammenhørende par af observationer fra et datamateriale.

Den lineære regressionsmodel for sammenhængen mellem  $X_i$ 'erne og  $Y_i$ 'erne er baseret på følgende antagelser:

- “Linearitet”:

Der eksisterer en lineær sammenhæng mellem responsvariablen og den forklarende variabel, dvs.  $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$  for  $i = 1, \dots, n$

- “Uafhængighed”:

Residualerne  $\varepsilon_1, \dots, \varepsilon_n$  er indbyrdes uafhængige

- “Konstant standardafvigelse”:

Alle residualerne  $\varepsilon_i$  har samme standardafvigelse  $\sigma$

- “Normalitet”:

Residualerne  $\varepsilon_1, \dots, \varepsilon_n$  er normalfordelte med middelværdi 0

Ud fra antagelserne bag den lineære regressionsmodel kan man vise, at de estimerede regressionsparametre  $\hat{\beta}_0$  og  $\hat{\beta}_1$  bliver normalfordelte.

Dermed kan vi...

- beregne konfidensintervaller for  $\beta_0$  og  $\beta_1$
- lave hypotesetest om  $\beta_0$  og  $\beta_1$

på præcis samme måde, som vi tidligere har gjort det for middelværdier i henholdsvis én og to grupper (Note 6, 7a, 7b).

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

## 1 Model

## 2 Estimation

## 3 Forklaringsgrad

## 4 Hypotesetest

## 5 OPSUMMERING

Når vi laver en model til at beskrive sammenhængen mellem to variable, er det naturligvis relevant at overveje, om det overhovedet er en god model. Mere præcist, overveje om modellen giver en tilstrækkeligt præcis beskrivelse af sammenhængen mellem de to variable.

Én måde at foretage den overvejelse på er ved at se på modellens såkaldte **forklaringsgrad**  $R^2$ . Forklaringsgraden er et tal mellem 0% og 100% og udtrykker, hvor stor en del af den samlede variation i variablen  $Y$ , som regressionsmodellen er i stand til at forklare.

Med "samlet variation i  $Y$ " mener vi variationen mellem de enkelte observationer af  $Y$  i datamaterialet. Det er den variation, vi mäter ved standardafvigelsen af  $Y$ . Standardafvigelsen mäter hvor meget  $Y$  varierer omkring sin gennemsnitlige værdi  $\bar{Y}$ .

(strent taget er det her variansen vi ser på, dvs. standardafvigelsen oploftet i 2. potens, men det er ikke vigtigt for at forstå fortolkningen af  $R^2$ )

Tanken bag beregningen af modellens forklaringsgrad  $R^2$  er at opdele variationen i  $Y$  i to dele:

$$\text{Samlet variation i } Y = \text{Forklaret variation i } Y + \text{Uforklaret variation i } Y$$

Her er...

- “*Forklaret variation i  $Y$* ” de forskelle på de enkelte observationer af  $Y$ , der skyldes forskelle i  $X$  (og som dermed kan forklares af regressionsmodellen)
- “*Uforklaret variation i  $Y$* ” de forskelle på de enkelte observationer af  $Y$ , der IKKE skyldes forskelle i  $X$  (og som dermed IKKE kan forklares af regressionsmodellen)

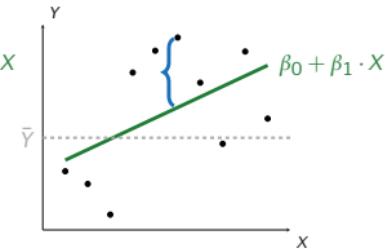
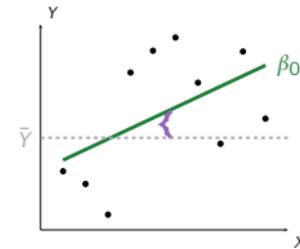
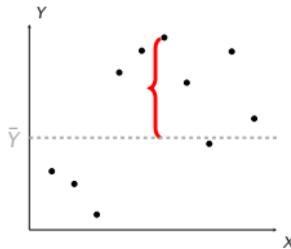
“*Forklaret variation i  $Y$* ” er den variation i  $Y$ , der skyldes ændringer i regressionslinjen, når  $X$  ændres. Dvs. den variation der skyldes regressionslinjens hældning  $\beta_1$ .  
“*Uforklaret variation i  $Y$* ” er den variation i  $Y$ , der skyldes standardafvigelsen  $\sigma$ .

Forklaringsgraden  $R^2$  beskriver dermed forholdet mellem regressionslinjens hældning  $\beta_1$  og standardafvigelsen  $\sigma$ : Hvis der er stor variation i  $Y$ , kan det enten skyldes en stejl hældning  $\beta_1$  på regressionslinjen eller en stor standardafvigelse  $\sigma$ .

Vi kan illustre beregningen af forklaringsgraden  $R^2$  ved at tegne figurer af hver af de tre forskellige typer af variation, vi ser på:

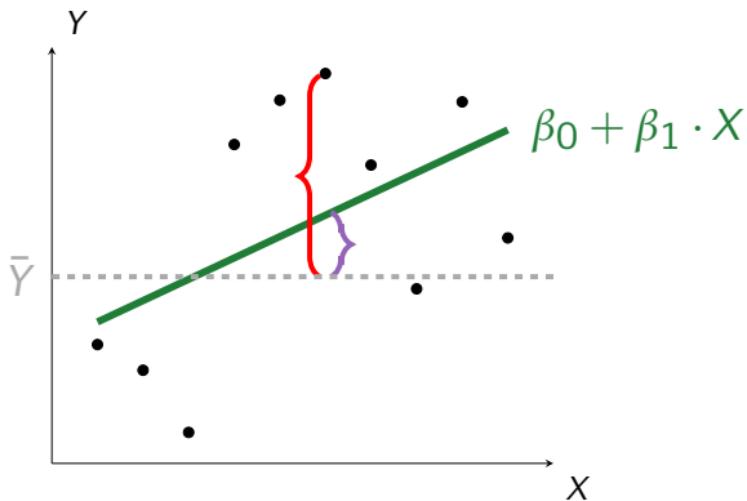
- “*Samlet variation i Y*”
- “*Forklaret variation i Y*” (variation i Y forklaret af regressionsmodellen)
- “*Uforklaret variation i Y*” (variation i Y IKKE forklaret af regressionsmodellen)

Illustration af de tre typer af variation for én enkelt observation:



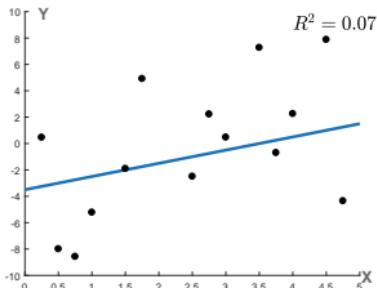
Forklaringsgraden  $R^2$ , der også kaldes for **determinationskoefficienten**, fremkommer ved at se på den **forklarede variation** for alle observationer og sammenholde den med den **samlede variation** for alle observationer.

$$R^2 = \frac{\text{"Forklaret variation i } Y\text{"}}{\text{"Samlet variation i } Y\text{""}}$$



Nedenstående animation viser, hvordan forklaringsgraden  $R^2$  ændres i takt med at sammenhængen mellem regressionslinjen  $\beta_0 + \beta_1 \cdot X$  og datamaterialet ændres.

Forklaringsgraden  $R^2$  beskriver, hvor stor en del af variationen i variablen  $Y$ , der beskrives af regressionslinjen  $\beta_0 + \beta_1 \cdot X$ .

ANIMATION: Ændring af forklaringsgraden  $R^2$ 

Start animation

## Eksempel: Boligpriser

Hvis vi estimerer den rette linje, der bedst beskriver sammenhængen mellem boligpris ( $Y$ ) og boligareal ( $X$ )

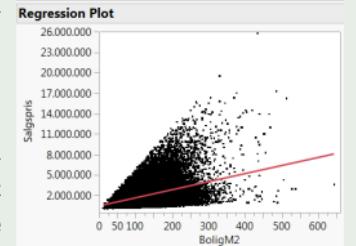
$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

baseret på hele datamaterialets 77.625 observationer, så får vi en relativt lav forklaringsgrad på  $R^2 = 0,1764$ . Intuitivt er det ikke overraskende, idet vi beder modellen forklare sammenhængen mellem boligpris og boligareal for boliger i alle dele af landet.

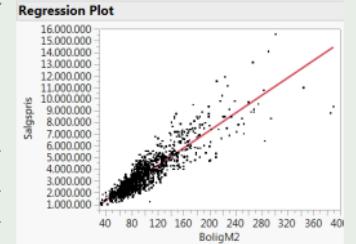
Hvis vi i stedet estimerer den rette linje, der bedst beskriver sammenhængen mellem boligpris ( $Y$ ) og boligareal ( $X$ )

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

udelukkende baseret på datamaterialets 1.381 boliger i Frederiksberg kommune, så får vi en relativt høj forklaringsgrad på  $R^2 = 0,7947$ . Selv om vi ikke tager højde for forskelle i boligerne's alder, type (villa/rækkehus/ejerlejlighed) m.m., så er det alligevel en langt mere ensartet gruppe af boliger, vi nu ser på. Dermed er det også langt lettere for modellen at give en god beskrivelse af datamaterialet.



Summary of Fit	
RSquare	0.176427
RSquare Adj	0.176416
Root Mean Square Error	1238204
Mean of Response	1911910
Observations (or Sum Wgts)	77625



Summary of Fit	
RSquare	0.794734
RSquare Adj	0.794585
Root Mean Square Error	791457.4
Mean of Response	3422709
Observations (or Sum Wgts)	1381

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

## 1 Model

## 2 Estimation

## 3 Forklaringsgrad

## 4 Hypotesetest

## 5 OPSUMMERING

I den simple lineære regressionsmodel

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

kan man teste flere forskellige statistiske hypoteser.

Den eneste form for hypoteser, vi vil se på i dette kursus, er hypoteserne af formen

$$H_0 : \beta_0 = b_0 \quad \text{og} \quad H_a : \beta_0 \neq b_0$$

(Nulhypotese: "linjens skæring  $\beta_0$  er lig  $b_0$ ")

og

$$H_0 : \beta_1 = b_1 \quad \text{og} \quad H_a : \beta_1 \neq b_1$$

(Nulhypotese: "linjens hældning  $\beta_1$  er lig  $b_1$ ")

Af særlig interesse er tilfældet  $H_0 : \beta_1 = 0$ ,  $H_a : \beta_1 \neq 0$ , fordi det svarer til at undersøge nulhypotesen om, at variablen  $X$  ikke har nogen betydning for variablen  $Y$ .

## Eksempel: Boligpriser

Vi estimerer den rette linje, der bedst beskriver sammenhængen mellem boligpris ( $Y$ ) og boligareal ( $X$ )

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

udelukkende baseret på datamaterialets 1.381 boliger i Frederiksberg kommune.

Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	
Intercept	-192926	53870,67	-3,58	0,0004*	
BoligM2	37524,175	513,5435	73,07	<.0001*	

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_0 : \beta_1 = 0 \end{cases}$$

Hvis vi tester nulhypotesen  $H_0 : \beta_0 = 0$ , så får vi en P-værdi på 0,04%, og forkaster derfor nulhypotesen (ved  $\alpha = 5\%$ ). Hvis vi tester nulhypotesen  $H_0 : \beta_1 = 0$ , får vi en P-værdi på under 0,01% og forkaster derfor ligeledes også her nulhypotesen.

Test af nulhypoteser af om  $\beta_0$  eller  $\beta_1$  kan være lig med en værdi (hhv.  $b_0$  eller  $b_1$ ), der ikke er 0, kan ikke laves i JMP ved at se på P-værdi. Her er man i stedet nødt til at se på det relevante konfidensinterval.

Hvis vi eksempelvis tester nulhypotesen  $H_0 : \beta_1 = 37.000$  og bruger  $\alpha = 5\%$  som signifikansniveau, så kan vi ikke forkaste nulhypotesen, netop hvis 37.000 er en del af  $1 - \alpha = 95\%$ -konfidensintervallet for  $\beta_1$ .

Parameter Estimates						
Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	-192926	53870,67	-3,58	0,0004*	-298603,3	-87248,63
BoligM2	37524,175	513,5435	73,07	<.0001*	36516,764	38531,586

Eftersom 95%-konfidensintervallet for  $\beta_1$  er [36.517; 38.532], så kan vi ikke forkaste nulhypotesen. Med andre ord, så er der ikke på baggrund af datamaterialet belæg for at afvise en påstand om, at salgsprisen på en bolig på Frederiksberg stiger med 37.000 kr. for hver ekstra kvadratmeter boligareal.

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

## 1 Model

## 2 Estimation

## 3 Forklaringsgrad

## 4 Hypotesetest

## 5 OPSUMMERING

## Kort opsummering af dette notesæt:

### Simpel lineær regression

En simpel lineær regressionsmodel beskriver en lineær sammenhæng mellem en forklarende variabel  $X$  og en responsvariabel  $Y$

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

hvor  $\varepsilon \sim N(0, \sigma)$  beskriver tilfældig variation omkring den rette linje  $\beta_0 + \beta_1 \cdot X$ .

Antagelser bag simpel lineær regression:

$(X_1, Y_1), \dots, (X_n, Y_n)$  er sammenhørende par af observationer som opfylder:

- “Linearitet”:  
Der eksisterer en lineær sammenhæng mellem responsvariablen og den forklarende variabel, dvs.  
 $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$  for  $i = 1, \dots, n$
- “Uafhængighed”:  
Residualerne  $\varepsilon_1, \dots, \varepsilon_n$  er indbyrdes uafhængige
- “Konstant standardafvigelse”:  
Alle residualerne  $\varepsilon_i$  har samme standardafvigelse  $\sigma$
- “Normalitet”:  
Residualerne  $\varepsilon_1, \dots, \varepsilon_n$  er normalfordelte med middelværdi 0

Forklарingsgraden  $R^2$  måler hvor godt den simple lineære regressionsmodel passer til datamaterialet. Forklарingsgraden antager altid en værdi mellem 0 og 1, og jo højere værdier, desto bedre passer modellen til datamaterialet.

# INDEKS

Model

Estimation

Forklaringsgrad

Hypotesetest

OPSUMMERING

Antagelser, simpel lineær regression	s. 19	Forklaringsgrad	s. 22
Determinationskoefficient	s. 25	Residual	s. 7
Forklarende variabel	s. 5	Responsvariabel	s. 5

## Nye funktionaliteter i dette notesæt:

- *Analyze -> Fit Model:*
  - Estimation af simpel lineær regression

## JMP-videoer:

- s. 2: ► [Graph -> Graph Builder] (lineær sammenhæng – boligpriser)
- s. 3: ► [Graph -> Graph Builder] (lineær sammenhæng – ølsalg)
- s. 17: ► [Analyze -> Fit Model] (estimation af lineær regression – boligpriser)
- s. 18: ► [Analyze -> Fit Model] (estimation af lineær regression – ølsalg)
- s. 27: ► [Analyze -> Fit Model] (forklaringsgrad  $R^2$ )