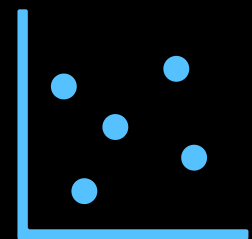




ISTx1003: Statistisk læring og data science

Forelesning 2: Introduksjon og multippel lineær regresjon

Forelesningen blir tatt opp

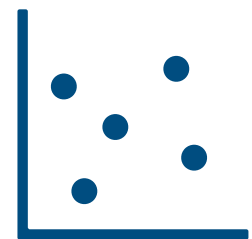




ISTx1003: Statistisk læring og data science

Forelesning 2: Introduksjon og multippel lineær
regresjon

Forelesningen blir tatt opp



Takk til

Mette Langaas (pensumvideoer, kompendium, slides og inspirasjon)
Stefanie Muff (slides og inspirasjon) Ingeborg Hem Sørmoen (slides og inspirasjon)

Mine slides bygger på materiale fra både Stefanie, Mette og Ingeborg, som underviste dette kurset i 2021-2023

Plan

Se Wiki sida for
utfyllende info

Forelesninger (små endringer kan forekomme)

- Mandag 21. oktober 15:15-16:00: Introduksjon
- Tirsdag 22. oktober 14:15-16:00: Oppgave 1 (multippel lineær regresjon)
- Mandag 28. oktober 15:15-16:00 Oppgave 2 (klassifikasjon) og oppgave 3 (klyngeanalyse)
- ??? Tirsdag 29. oktober 14:15-16:00 Oppgave 3 fors. (?), Q&A og praktisk informasjon
- Alle de 4 forelesningene vil være hel-digitale
- Alle forelesningene tas opp, tentative slides legges ut før forelesningen og slides med notater legges ut etter forelesningen
- Forelesningene vil i all hovedsak utdype prosjektoppgavene. Pensum dekkes av kompendier og tilhørende videoer, og blir ikke repetert her.

Plan

Se Wiki Sida for
utfyllende info

Pensumvideoer

For å få fullt utbytte av forelesningene bør dere ha sett følgende:

- Enkel lineær regresjon (uke 9) innen mandag 21. oktober (i dag)
- Multippel lineær regresjon innen tirsdag 22. oktober (i morgen)
- Klassifikasjon innen 28. oktober
- Klyngeanalyse innen 28. oktober

Plan

Se Wiki Sida for
utfyllende info

Veiledning

- Digitalt:
 - Tirsdag 5/11 og 12/11, 14:15-16:00 Vi øker mengden digital veiledning ved behov.
 - Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00
- Fysisk:
 - Trondheim: Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00 i S5 (SBII)
 - Gjøvik: Kun digitalt
 - Ålesund: Digital. I tillegg gruppene kan møte Siebe til veiledning, avtal tid med han. Se blackboardsiden ISTA: Campus Ålesund for mer informasjon.
- Forum:
 - Åpent 24/7, besvares jevnlig av fagteamet

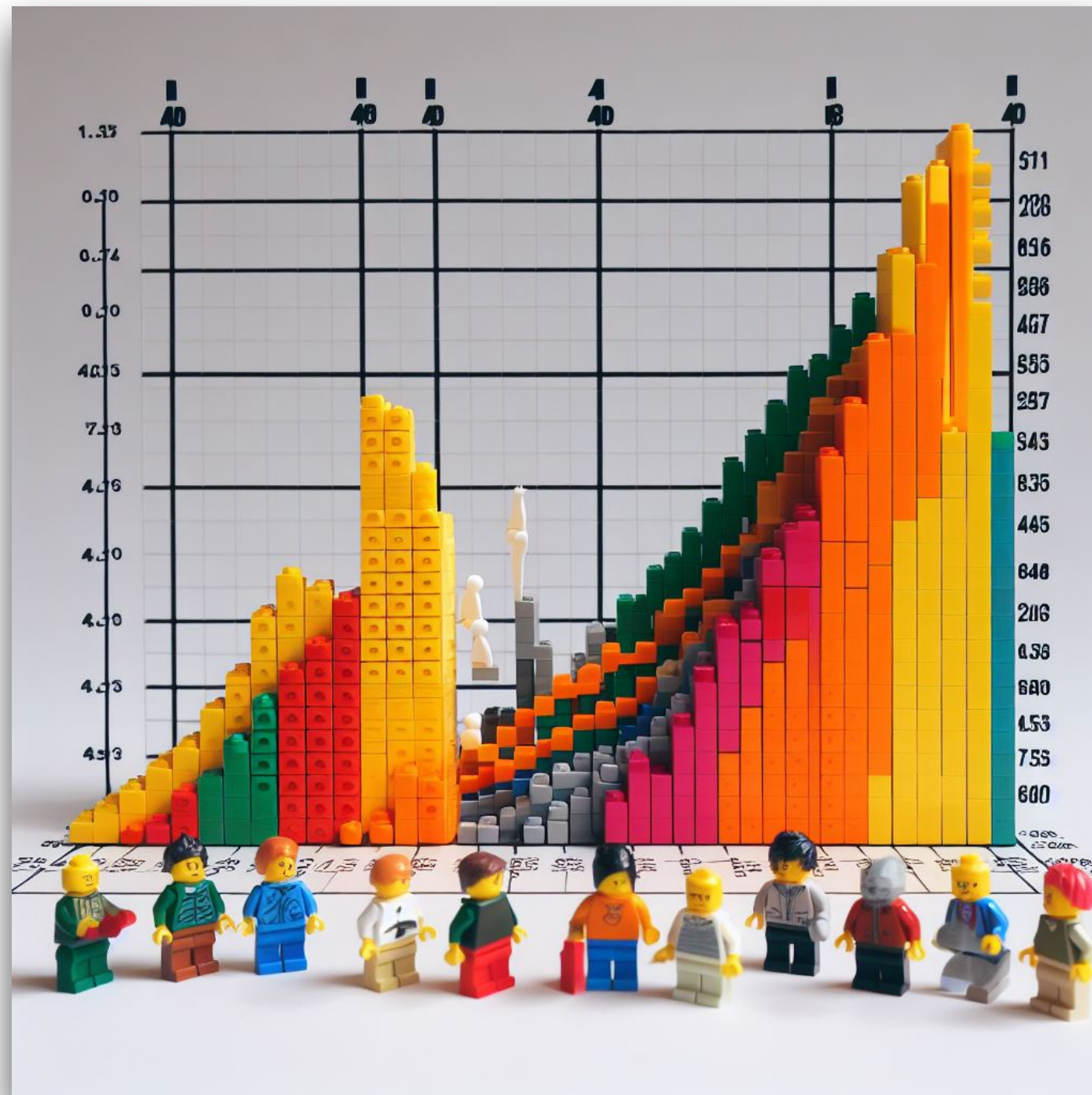
**18. november 12:00:
Leverer prosjekt i Inspira**

Plan

I dag, 22. oktober

1. Diverse informasjon
2. Regresjon
 1. Laeringsmål

Opppgave 1: Regresjon



Microsoft bing prompt: "Multiple linear regression visualized using LEGO"

Læringsmål
Introduksjon
Enkel lineær regresjon
Multippel lineær regresjon (MLR)
Noen råd til slutt
Referanser

Multippel lineær regresjon: Interaksjonseffekter

Ingeborg Hem Sørmoen, med hjelp fra Kenneth Aase

Dette er et kort notat om interaksjoner i multippel lineær regresjon (MLR). Se også læringsmaterialet om regresjon.

Superkort oppsummering

En interaksjonseffekt/samspilleffekt kan tenkes på som en ny forklaringsvariabel, bestående av produktet av to andre forklaringsvariabler. Legg til denne i modellen din, og du har en interaksjonseffekt. Merk at det ikke er alle interaksjonseffekter som gir mening, så det er viktig å undersøke betydningen av modellen din før du setter i gang!

I Python kodes dette som: $y \sim x + \text{cat} + x:\text{cat}$ der y er respons og x og cat er forklaringsvariabler. $x:\text{cat}$ lager interaksjonseffekten. Dette kan også skrives slik: $y \sim x*\text{cat}$, det blir det samme.

Endre referansekategori: $+ x*C(\text{cat}, \text{Treatment}("nyrefcat"))$.

Introduksjon

I dette eksempelet bruker vi et subset av datasettet som benyttes i Oppgave 1 i prosjektet, nemlig LEGO-datasettet (hentet fra Peterson and Ziegler (2021)). Dette er et datasett bestående av en kontinuerlig responsvariabel (pris), en kontinuerlig forklaringsvariabel (antall brikker), og en kategorisk forklaringsvariabel (tema). Variablene er som følger:

Theme	Tema settet hører til
Pieces	Antall brikker i settet
Price	Pris

Vi vil se hvordan pris kan beskrives av antall brikker i LEGO-settet. De øverste 6 radene i datasettet ser slik ut:

	Theme	Price	Pieces
1	Tema 2	9.99	47
2	Tema 3	129.99	123
3	Tema 3	19.99	15
4	Tema 2	99.99	1218
5	Tema 3	19.99	20
6	Tema 1	19.99	121

Hvis vi plotter dette datasettet får vi følgende figur:

1

https://www.math.ntnu.no/emner/IST100x/ISTx1003/notat_interaksjoner.pdf

Eksempel: Leiepriser

IST[A/G/T]1003: Statistisk læring og data science

Regresjon

Mette Langaas IMF/NTNU

19 October, 2020

Læringsmål

Etter du har lest dette kompendiet, sett videoene som er laget og deltatt på zoom-forelesningene skal du

- kunne forstå hva regresjon går ut på og
- kjenne igjen situasjoner der regresjon vil være en aktuell metode å bruke
- kjenne til modellen for multippel lineær regresjon, og kunne tolke utvalgte deler av en generell utskrift fra modelltilpasning
- forstå hvordan vi utfører multippel lineær regresjon i Python
- kunne svare godt på regresjonsoppgaven i den tellende prosjektoppgaven

Introduksjon

Videoressurser

Videoer som diskuterer innholdet av dette kompendiet:

For enkel lineær regresjon finnes 4 videoer i fellesdelen. Disse er

- [Del 1](#)
- [Del 2](#)
- [Del 3](#)
- [Del 4](#)

Det er i tillegg laget en video for utvidelsen fra enkel til multippel lineær regresjon

- [Multippel lineær regresjon: introduksjon](#)
- [Multippel lineær regresjon: analyse av et datasett](#)

Hvorfor regresjon?

Det er to hovedgrunner til at vi vil utføre en regresjonsanalyse:

1. Vi vil lage en modell for å hjelpe oss med å **forstå sammenhengen** mellom en *respons* og en eller flere *forklaringsvariable*.
2. Vi vil lage en modell for å **predikere** en *respons* fra en eller flere *forklaringsvariable*.

Lineær regresjon er en veiledet og parametrisk metode, og er byggesteinen for veiledet maskinlæring.

<https://www.math.ntnu.no/emner/IST100x/ISTx1003/Regresjon.html>

Multippel lineær regresjon: introduksjon

Mette Langaas, Institutt for matematiske fag, NTNU

Multippel lineær regresjon: analyse av et datasett

Mette Langaas, Institutt for matematiske fag, NTNU

Regresjon

Læringsmål

Etter du har lest kompendiet, sett videoene som er laget og deltatt på Zoom-forelesningene skal du:

- kunne forstå hva regresjon går ut på og
- kjenne igjen situasjoner der regresjon vil være en aktuell metode å bruke
- kjenne til modellen for multippel lineær regresjon, og kunne tolke utvalgte deler av en generell utskrift fra modelltilpasning
- forstå hvordan vi utfører multippel lineær regresjon i Python
- **kunne svare godt på [oppgave 1](#) i den tellende prosjektoppgaven**

Regresjon

1. Vi vil lage en modell for å hjelpe oss med å forstå sammenhengen mellom en respons og en eller flere forklaringsvariabler.
2. Vi vil lage en modell for å predikere en respons fra en eller flere forklaringsvariabler.

Regresjon kan brukes til veldig mye! Et godt sted å starte en analyse som man tenker skal bli mer avansert.

Pris på varer Helsedata Aksjemarked
Materialstyrke
Analysere lønn Løpedata

Oppgave 1 - Multippel lineær regresjon

Superkort oppsummering

- Rapporten skal være på **maksimalt 5 sider** (referanser og kode kommer i tillegg).
- Det legges vekt på at rapporten skal være sammenhengende.
- 'oppg_1.ipynb' inneholder kode for å renske dataene.
- Problemstilling bestemmes **først**.
- Det anbefales å sjekke med faglærer (Sara) at problemstillingen er god før dere starter.

Obs! Interaksjonsnotatet, eksempelkode i oppg_1.ipynb og jeg bruker bare deler av datasettet. Dere skal bruke hele datasettet!

Vi skal bruke et LEGO-datasett, der hver observasjon består av:

Set_Name	Navn p°a settet
Theme	Tema settet hører til
Pieces	Antall brikker i settet
Price	Pris (i \$)
Pages	Antall sider i instruksjonsmanualen
Unique_Pieces	Antall unike brikker

I denne oppgaven skal dere selv definere en problemstilling (innenfor visse rammer, se punkt 1 under) som dere skal prøve å besvare ved å bruke LEGO-datasettet og multippel lineær regresjon. Oppgavens pipeline er som følger:

1. Definere problemstilling. Problemstillingen skal formuleres på en måte slik at dere må lage en ny kategorisk variabel som grupperer de ulike temaene. Eksempler på problemstillinger er:
 - (a) Er LEGO for gutter dyrere enn LEGO for jenter?
 - (b) Betaler vi ekstra for LEGO som assosieres med et varemerke?
 - (c) Varierer prisen på LEGO i ulike aldersgrupper?
 - (d) Har/hadde medlemmene på vår gruppe dyr eller billig LEGO (som barn)?

Eksempler på tilhørende grupperinger kan da være henholdsvis:

- (a) kjønn (lego for gutter, jenter, kjønnsnøytralt)
- (b) varemerke (ja, nei, ingen informasjon) (Disney er for eksempel et eget varemerke)
- (c) alder (baby/småbarn, skolebarn, eldre, uvisst)
- (d) hva gruppemedlemmene eier/eide som barn (ja, nei, usikker)

Den nye kategoriske variabelen skal ha maksimalt 4 ulike kategorier (helst akkurat 3), og kun fantasien setter grenser.

2. Finne (det har vi gjort for dere) og laste ned data. Og deretter rense dataene (det har vi også gjort for dere).
3. Pre-prosessere dataene så dere kan svare på problemstillingen (altså: gruppere temaene). Husk å beskrive hvilke temaer som havner i hvilken kategori **og hvorfor!** Hele datasettet skal brukes.
4. Formulere og begrunne modell(er) og hypotese(r) for å svare på problemstillingen. Husk å forklare hva som er respons, hva som er/kan være forklaringsvariabler, **og hvorfor!**
5. Tilpasse modell(er) og evaluere den/dem.
6. Diskutere om problemstillingen kunne besvares med datasettet dere endte med (hvorfor/hvorfor ikke?). Hva ble konklusjonen? (Husk at det ikke er lov til å endre problemstilling etter dere har undersøkt dataene).

Rapporten dere skal levere skal struktureres i samme rekkefølge som pipelinen for oppgaven. I tillegg til at rapporten inneholder relevante analyser og faglig innhold, legges det også vekt på at den er sammenhengende. Husk å forklare teori, metoder og begreper, referere til kilder dere bruker, og rapportere, tolke og diskutere relevante modeller, tall og resultater. Lærebok, kompendier og videoer er helt greit å referere til. Rapporten skal inneholde figurer (husk relevante navn på akser). Koden dere har brukt skal IKKE inkluderes underveis i rapporten, men til slutt som et appendix (eller la koden være tilgjengelig på nett).

Rapporten skal være på **maksimalt 5 sider** (ikke inkludert referanser og kode), skrevet med Times New Roman skriftstørrelse 12, Calibri (body) skriftstørrelse 11, eller lignende. **Kun de første 5 sidene rettes!!!**

Merk: Problemstillingen skal defineres **først**, og ikke endres i ettertid når dere ser resultatene. En god rapport er ikke ekvivalent med spennende resultater – å svare at “vi har ikke grunnlag for å svare på problemstillingen” (med analyser og begrunnelser) er også et godt svar. Vi anbefaler at dere sjekker med faglærer (Sara) at problemstillingen deres er god før dere begynner.

Rettemal oppgave 1

I rettingen vurderes det i hvor stor grad:

- Rapporten er sammenhengende og helheltsinntrykket er godt
- Problemstillingen passer til MLR-rammeverket og er innenfor oppgavens retningslinjer
- Rapporten har en god beskrivelse av den nye kategorien og alle data som brukes
- Rapporten svarer på den valgte problemstillingen
- Rapporten inneholder relevant teori
- Rapporten inneholder relevante analyser
- Rapporten inneholder relevante og beskrivende figurer
- Rapporten inneholder relevante og gode tolkninger
- Rapporten inneholder en relevant og god diskusjon av hvorvidt problemstillingen kunne besvares
- Rapporten inneholder referanser

Deretter settes det en score for hele rapporten.

Skriv rapporten i f.eks. Word eller Latex, og lagre den som en pdf før innlevering.

1) Problemstilling og gruppering

Problemstilling: “Betaler vi ekstra for LEGO som assosieres med et varemerke?”

Gruppering: varemerke (ja, nei, uklart)

Obs: Pass på å stille et spørsmål som faktisk kan besvares ved hjelp av datasettet. Det må være nok observasjoner i hver gruppe!

Helt ok å velge ett av forslagene i oppgaveteksten (men kanskje litt kjedelig?).

Anbefales å spørre faglærer (Sara) om problemstillingen er god før dere starter (f.eks. via epost).

2) Finne, laste ned og rensse data

Vi har funnet dataene for dere, de finnes under “Supplementary materials” [her](#).

Datasets and Stories

Building a Multiple Linear Regression Model With LEGO Brick Data

Anna D. Peterson & Laura Ziegler 

Pages 297-303 | Published online: 31 Jul 2021

 Cite this article  <https://doi.org/10.1080/26939169.2021.1946450>

Supplemental material

Supplemental Material

 Download Zip (412 KB)

Last ned dataene og last dem så opp i huben eller legg dem der du kjører notatboka. I notatboka ligger det kode for å rensse dataene, så det slipper dere å skrive selv.

3) Preprosessere dataene

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

Mitt krympede datasett har 11 temaer, fordelt på 356 LEGO-sett (observasjoner):

Tema	Varemerke	Ikke varemerke	Uklart
Ideas			
Juniors			
Batman			
Jurassic World			
NINJAGO			
Spider-Man			
Marvel			
THE LEGO MOVIE 2			
Harry Potter			
Creator 3-in-1			
City			
SUM			

3) Preprosessere dataene

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

Mitt krympede datasett har 11 temaer, fordelt på 356 LEGO-sett (observasjoner):

Tema	Varemerke	Ikke varemerke	Uklart
Ideas		X	
Juniors		X	
Batman			X
Jurassic World	X		
NINJAGO	X		
Spider-Man			X
Marvel	X		
THE LEGO MOVIE 2			X
Harry Potter	X		
Creator 3-in-1		X	
City		X	
SUM	170	144	42

Husk å beskrive hvorfor dere tar valgene dere tar!

4) Formulere og begrunne modell(er) og hypotese(r)

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

Respons: Pris (343 observasjoner)

Forklaringsvariable: Antall brikker (kontinuerlig), tema (kategorisk), antall sider i manualen (kontinuerlig).

Jeg **velger** å bruke kun disse forkl.var.

For å undersøke om vi betaler mer for varemerke-assosiert LEGO, vil jeg tilpasse følgende modeller:

- A. Pris beskrevet av antall brikker
- B. Pris beskrevet av antall brikker og antall sider i manualen
- C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe
 - A. Egen modell for hver gruppe
 - B. Kun eget skjæringspunkt
 - C. Eget skjæringspunkt og stigningstall

4) Formulere og begrunne modell(er) og hypotese(r)

Modell A: Pris beskrevet av antall brikker

Y_i = Pris for observasjonen i

x_i = Antall brikker i observasjon i

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

4) Formulere og begrunne modell(er) og hypotese(r)

Modell B: Pris beskrevet av antall brikker og antall sider i manualen

Y_i = Pris for observasjonen i

x_{1i} = Antall brikker i observasjon i

x_{2i} = Antall sider i observasjon i

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

4) Formulere og begrunne modell(er) og hypotese(r)

Modell C.A: Egen modell for hver gruppe

y_i^{Cat} = Pris for observasjonen i i gruppe Cat

x_i^{Cat} = Antall brikker i observasjon i i gruppe Cat

$$y_i^{\text{Cat}} = \beta_0^{\text{Cat}} + \beta_1^{\text{Cat}} x_i^{\text{Cat}} + \epsilon_i^{\text{Cat}}$$

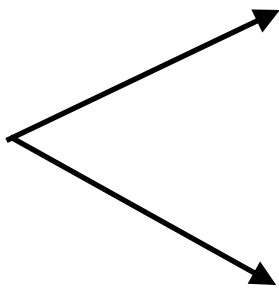
4) Formulere og begrunne modell(er) og hypotese(r)

Modell C.B: Kun eget skjæringspunkt

Y_i = Pris for observasjonen i

x_i = Antall brikker i observasjon i

z_i = Kategori for observasjonen i



$z_{1i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 2} \\ 0 & \text{ellers} \end{cases}$

$z_{2i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 3} \\ 0 & \text{ellers} \end{cases}$

$$Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 x_i + \epsilon_i$$

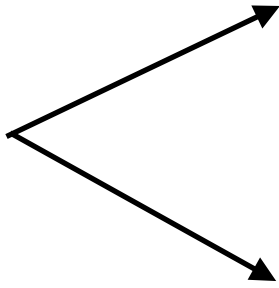
4) Formulere og begrunne modell(er) og hypotese(r)

Modell C.C: Eget skjæringspunkt og stigningstall

Y_i = Pris for observasjonen i

x_i = Antall brikker i observasjon i

z_i = Kategori for observasjonen i



$z_{1i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 2} \\ 0 & \text{ellers} \end{cases}$

$z_{2i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 3} \\ 0 & \text{ellers} \end{cases}$

$$Y_i = \beta_0 + \beta_1 z_{1i} \beta_1 + \beta_2 z_{3i} + (\beta_4 z_{1i} \beta_5 + \beta_6 z_{3i}) x_i + \epsilon_i$$

5) Tilpasse og evaluere modell(er)

Regresjonsanalyse i 5 steg:

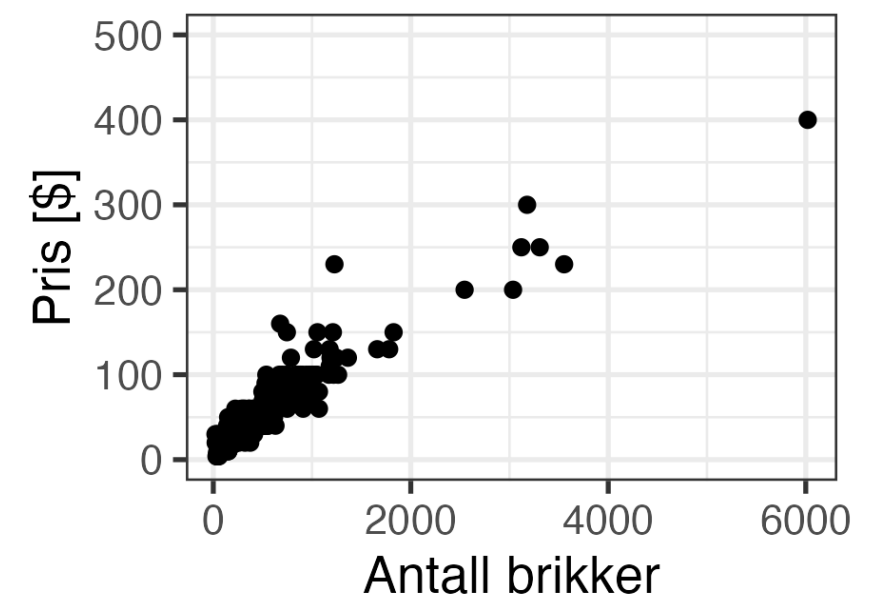
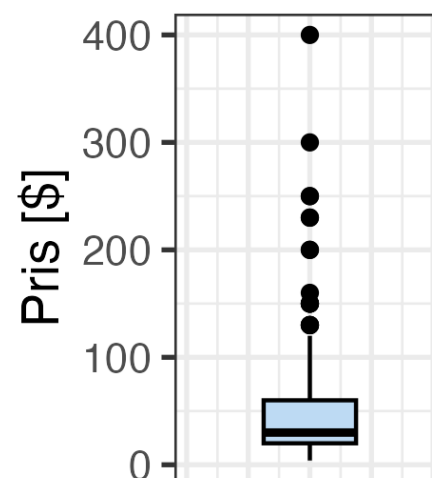
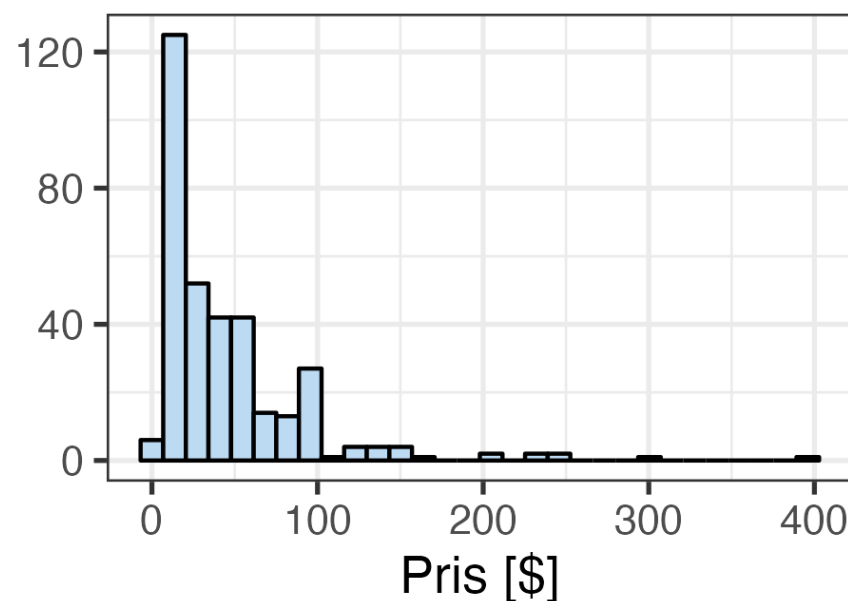
1. Bli kjent med dataene ved å se på oppsummeringsmål og ulike typer plott
2. Spesifisere en matematisk modell
3. Initialisere og tilpasse modellen
4. Presentere resultater fra den tilpassede modellen
5. Evaluere om modellen passer til dataene

Jeg viser ikke alt for alle modellene her. I rapporten holder det å presentere relevante modeller, og heller beskrive resten uten å rapportere alle resultater. Men, det må være tydelig at dere har forstått pensum!

5) Tilpasse og evaluere modell(er)

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

1. Bli kjent med dataene



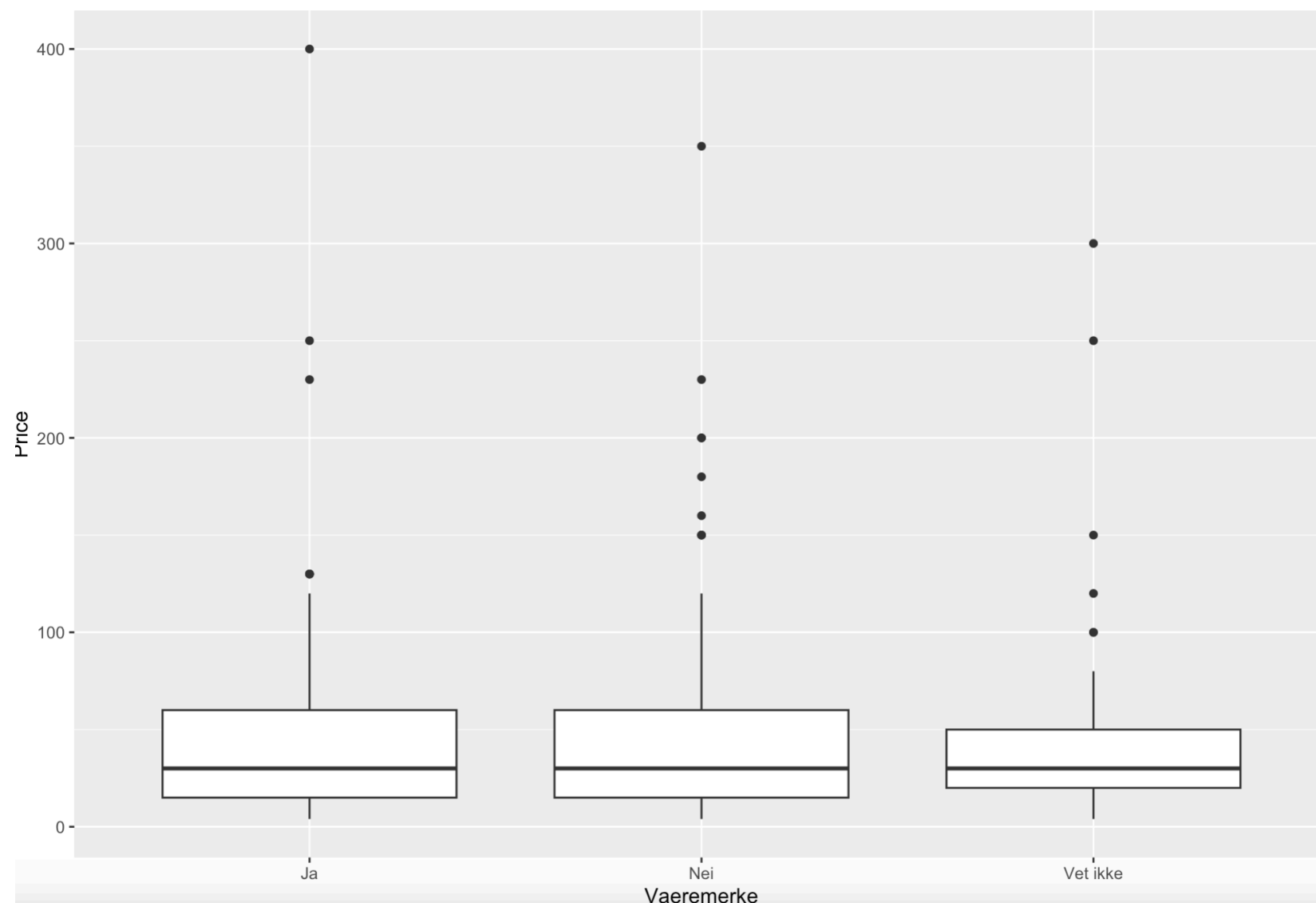
min: 3.99, max: 399.99, mean: 46.4477, median: 29.99

Obs! Det er ikke selve responsen som skal være normalfordelt, det er residualene!

5) Tilpasse og evaluere modell(er)

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

1. Bli kjent med dataene



5) Tilpasse og evaluere modell(er)

A. Pris beskrevet av antall brikker

2. Matematisk modell

$$Y_i = \beta_0 + \beta_1 x_{1,i} + e_i$$

$$y \sim x1$$

3. Initialisere og tilpasse modellen

Tilpasse betyr å estimere de ukjente modellparametrene i (for eksempel) Python

$$\beta_0, \beta_1 \rightarrow \hat{\beta}_0, \hat{\beta}_1$$

$$\sigma^2 \rightarrow S^2$$

5) Tilpasse og evaluere modell(er)

A. Pris beskrevet av antall brikker

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

4. Resultater

```
modA = smf.ols('Price ~ Pieces', data = df)
modA.fit().summary()
```

OLS Regression Results						
Dep. Variable:	Price			R-squared:	0.873	
Model:	OLS			Adj. R-squared:	0.873	
Method:	Least Squares			F-statistic:	2345.	
Date:	Fri, 13 Oct 2023			Prob (F-statistic):	6.78e-155	
Time:	10:05:13			Log-Likelihood:	-1450.6	
No. Observations:	343			AIC:	2905.	
Df Residuals:	341			BIC:	2913.	
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	13.2558	1.131	11.720	0.000	11.031	15.480
Pieces	0.0756	0.002	48.427	0.000	0.073	0.079
Omnibus:	215.964	Durbin-Watson:		1.325		
Prob(Omnibus):	0.000	Jarque-Bera (JB):		3092.865		
Skew:	2.355	Prob(JB):		0.00		
Kurtosis:	16.937	Cond. No.		910.		

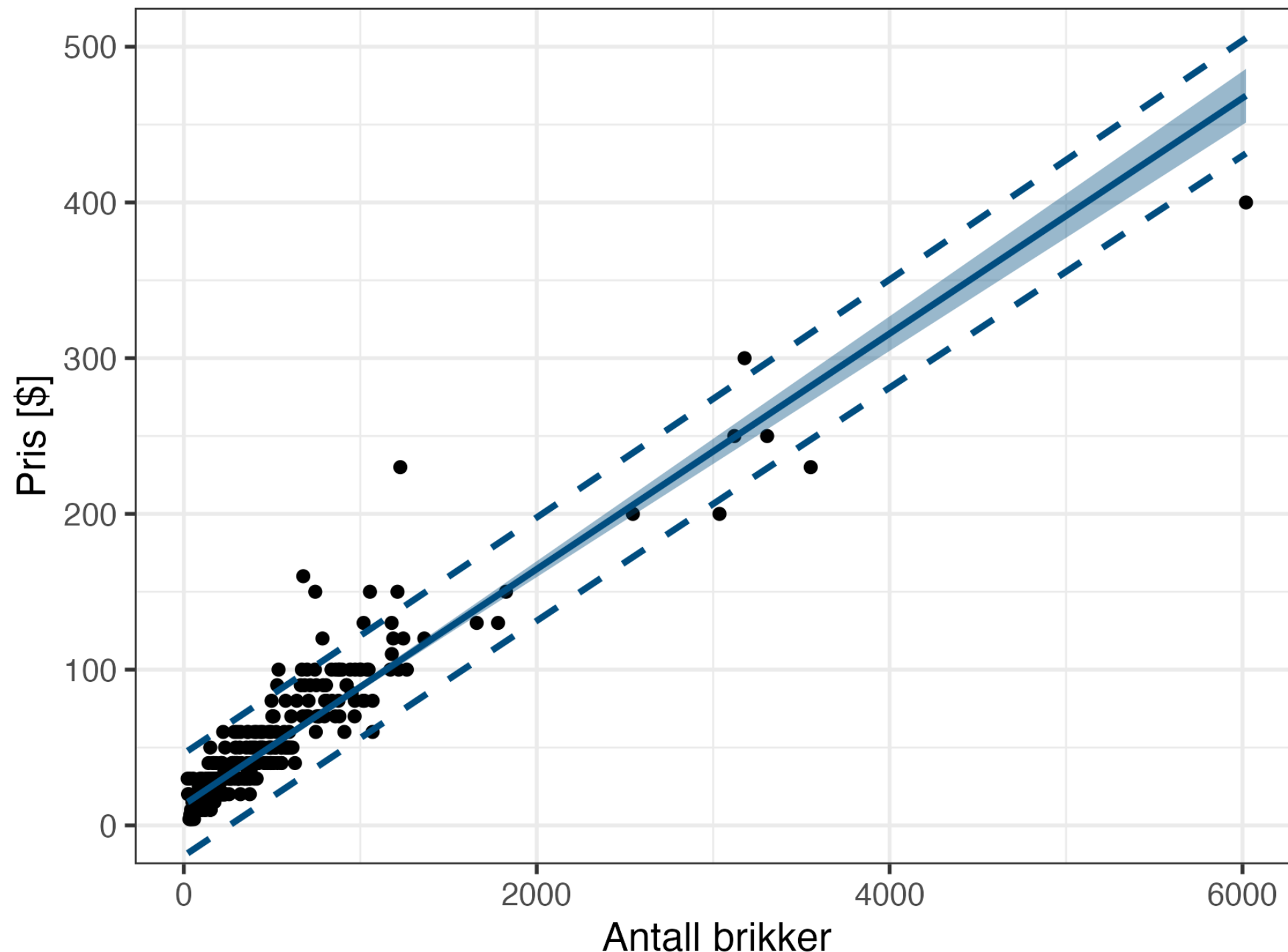
Obs! p -verdi kan ikke være 0.000, det betyr egentlig at $p < 0.0005$

5) Tilpasse og evaluere modell(er)

A. Pris beskrevet av antall brikker

4. Resultater

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

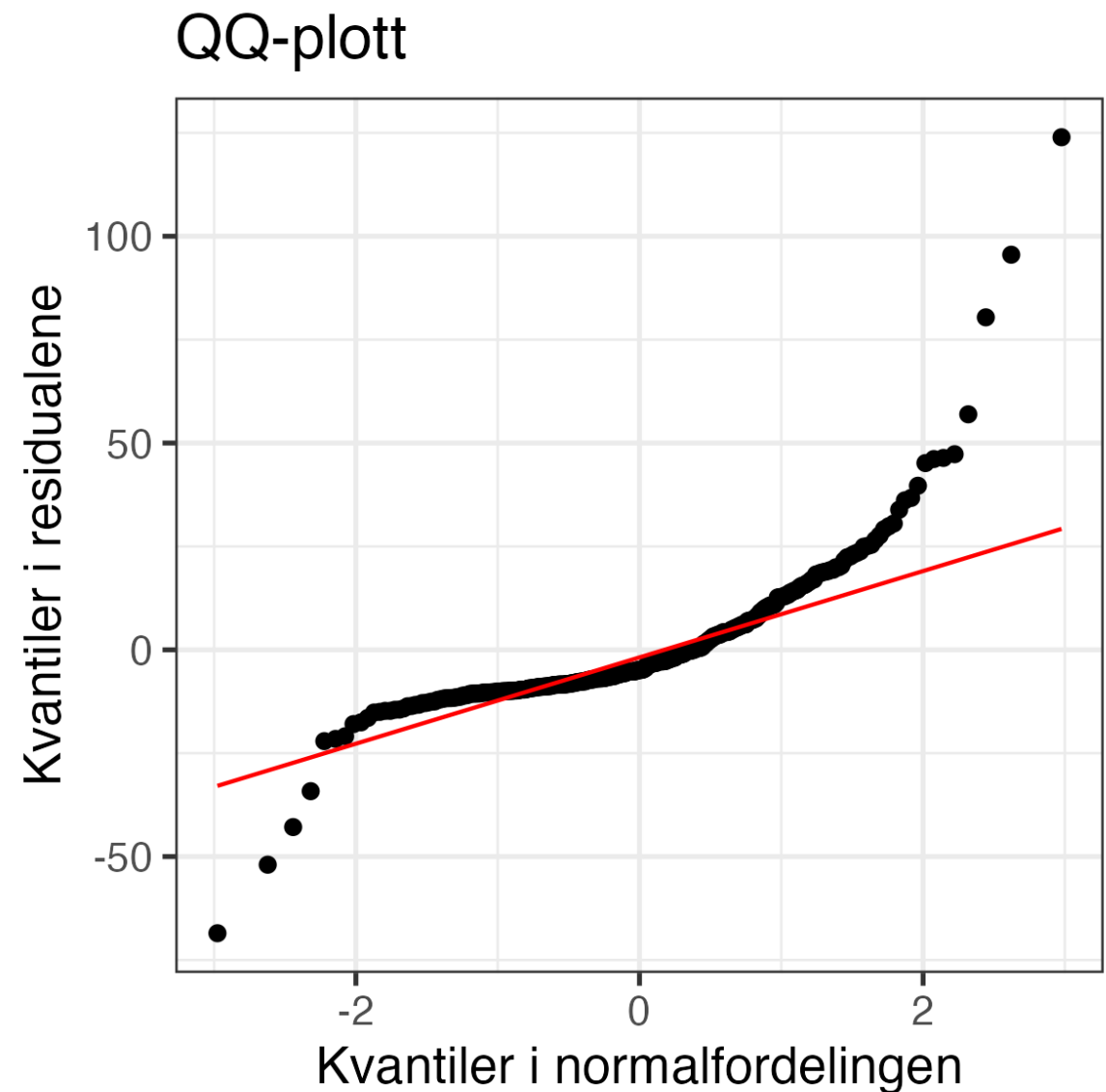
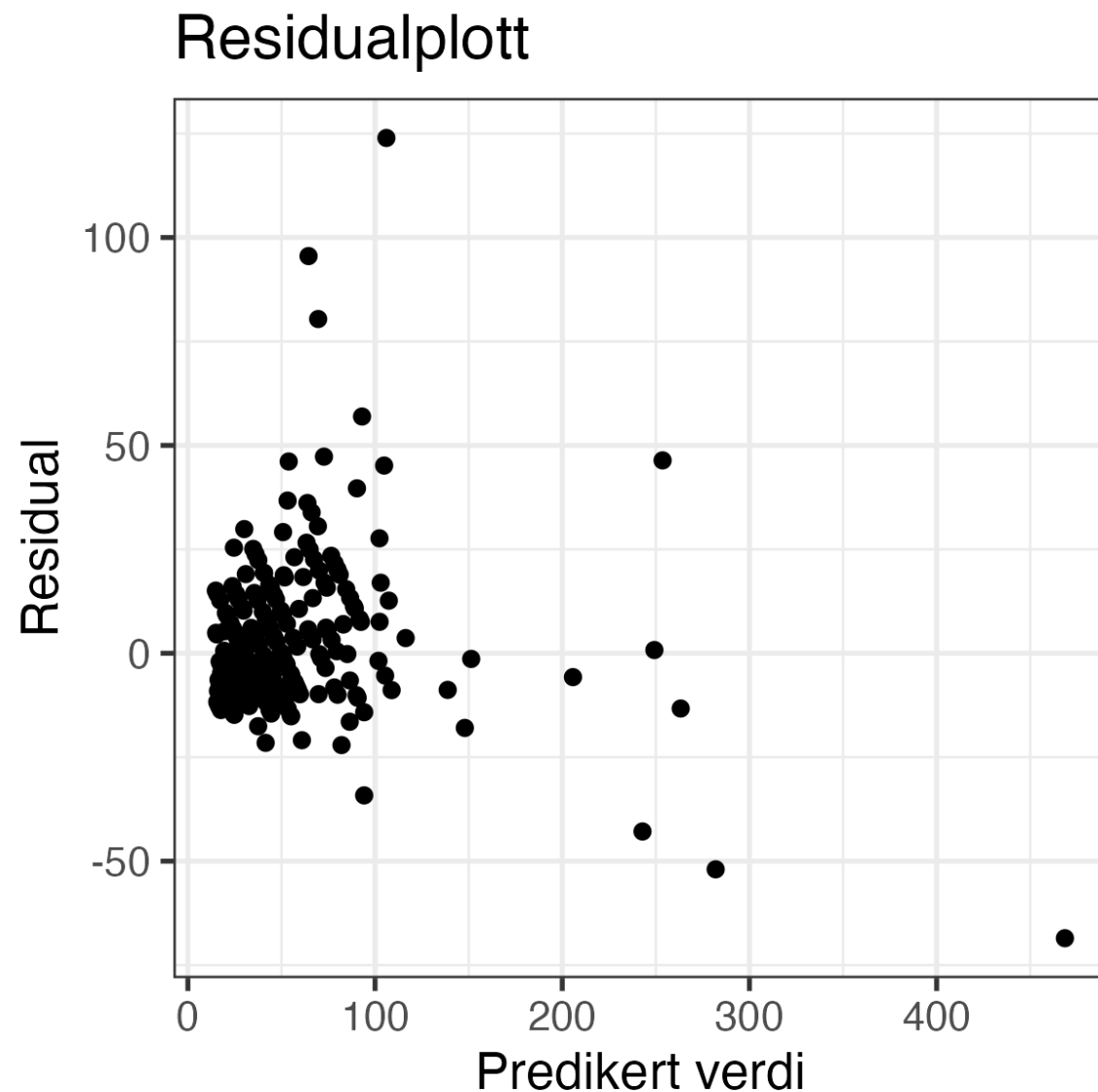


5) Tilpasse og evaluere modell(er)

A. Pris beskrevet av antall brikker

5. Passer modellen til dataene?

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!



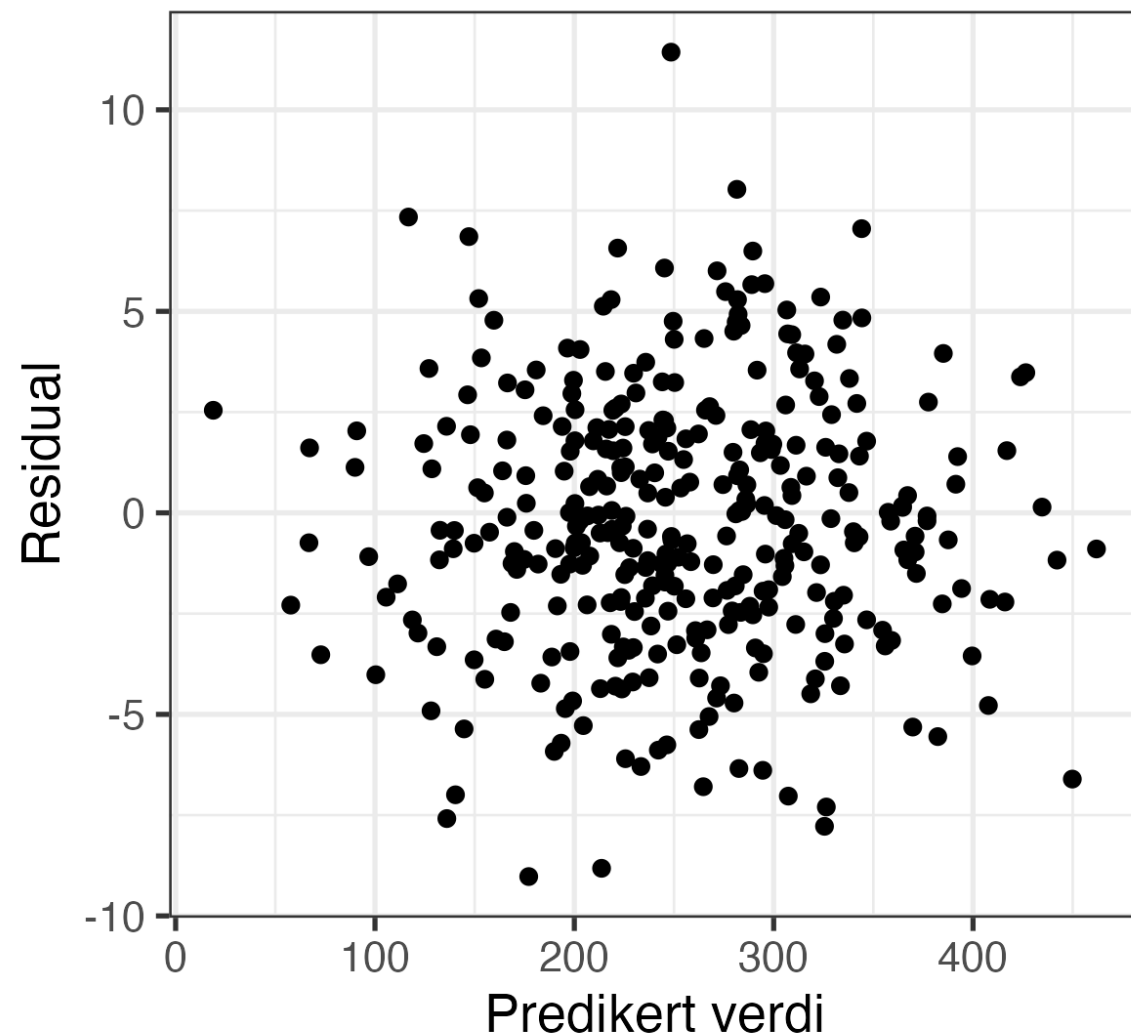
R^2 : 0.873

5) Tilpasse og evaluere modell(er)

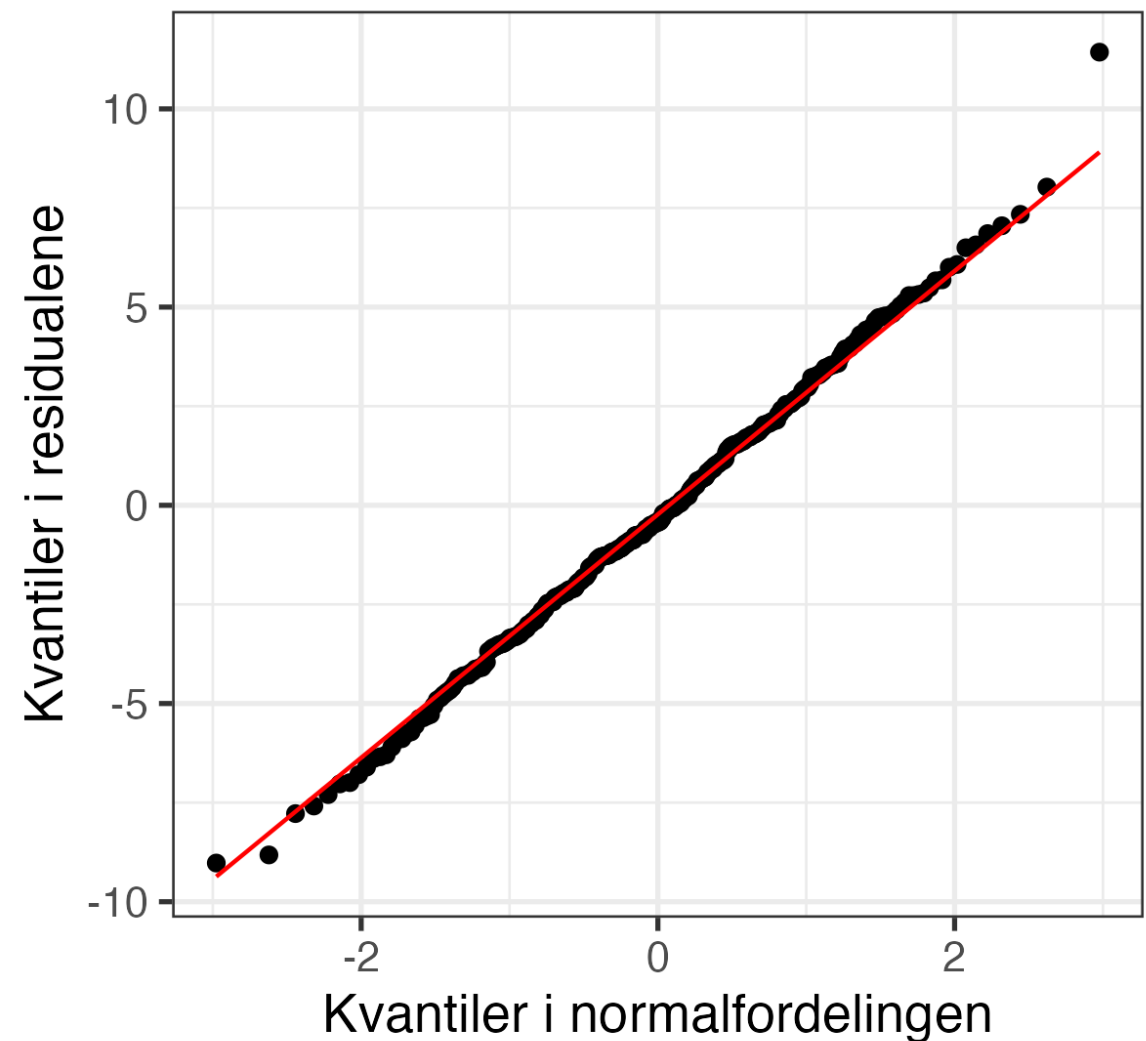
Når alt er “bra”

5. Passer modellen til dataene?

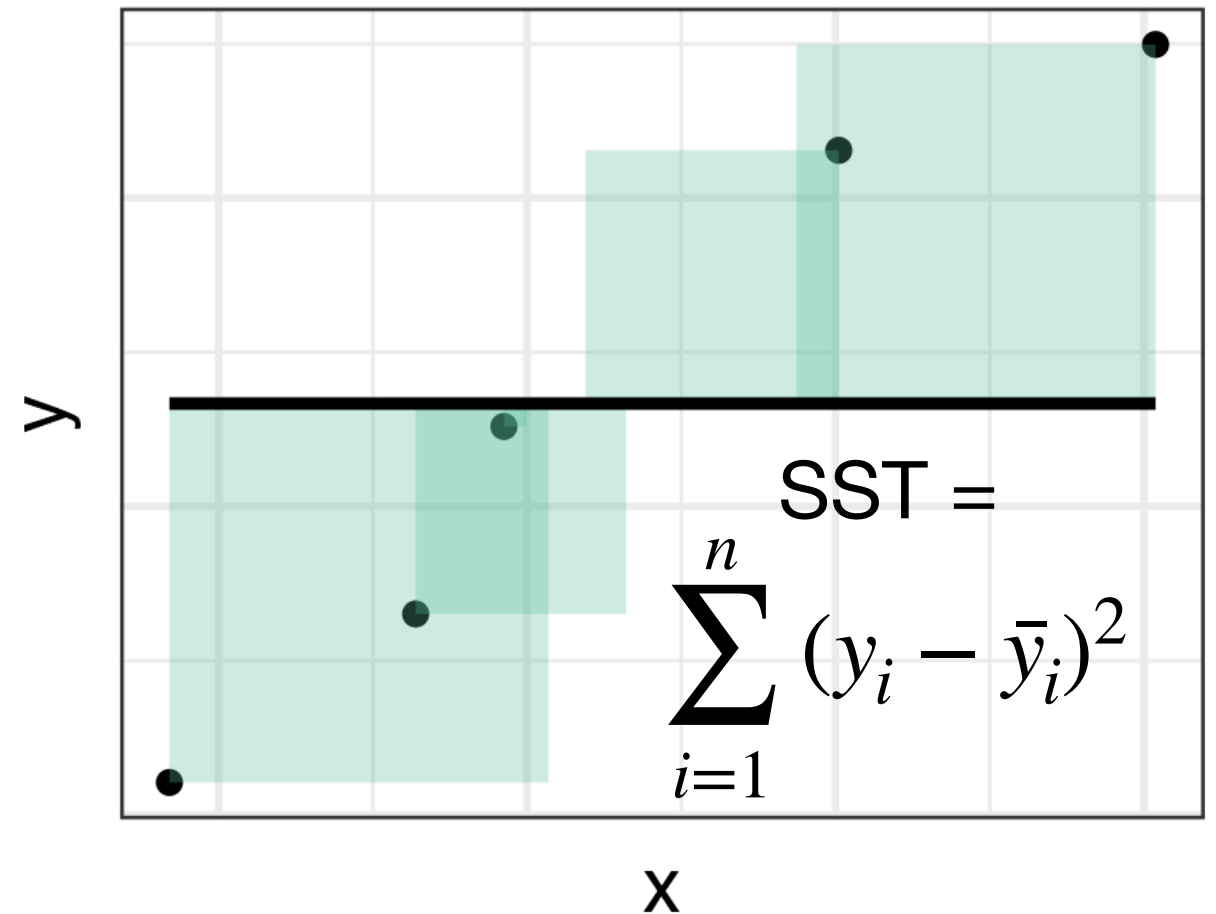
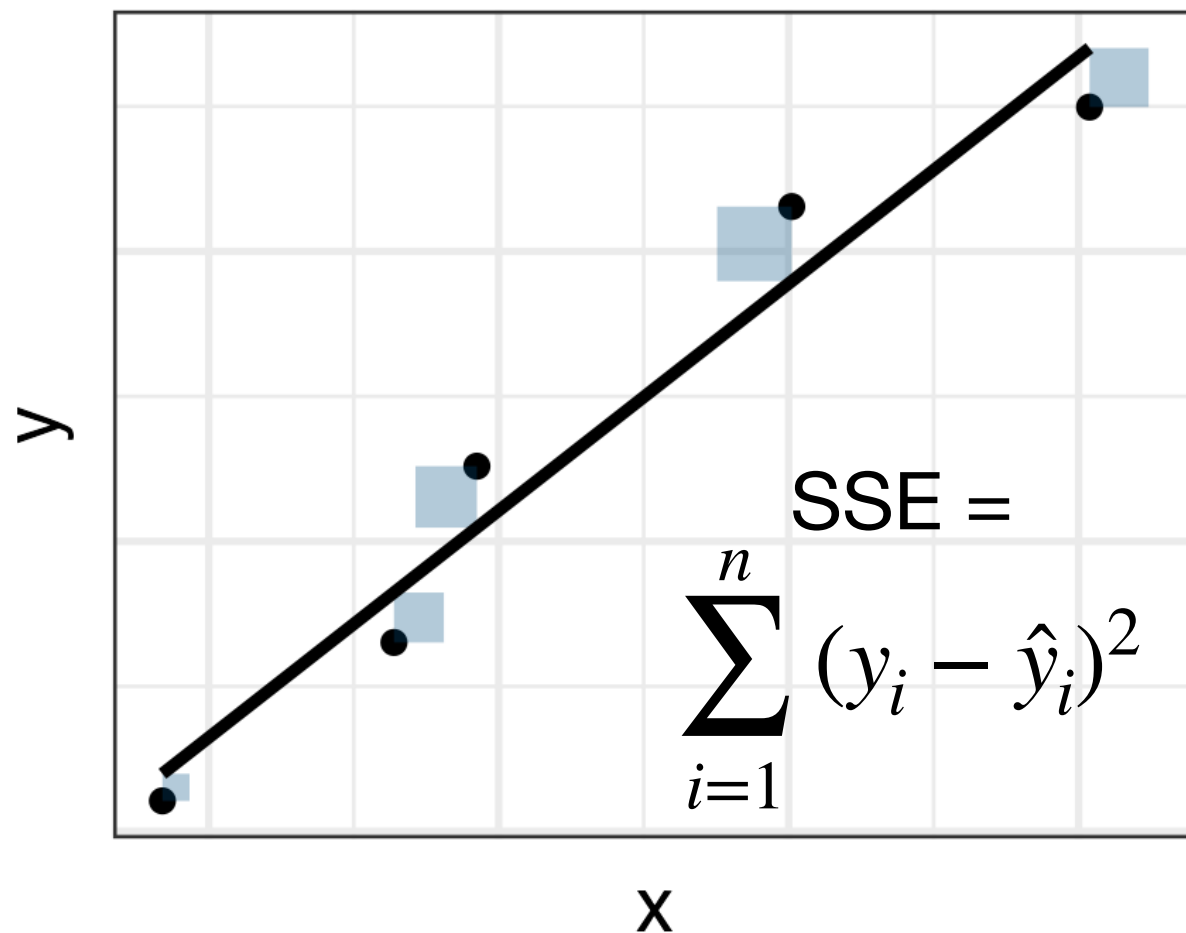
Residualplott, simulerte data



QQ-plott, simulerte data



$$R^2$$



$$R^2 = \frac{SST - SSE}{SST} = \text{“andel av varians i dataene som er forklart av modellen”}$$

Justert R^2 tar hensyn til at vi legger til parametere

5) Tilpasse og evaluere modell(er)

Modell B: Pris beskrevet av antall brikker og antall sider i manualen

Y_i = Pris for observasjonen i

x_{1i} = Antall brikker i observasjon i

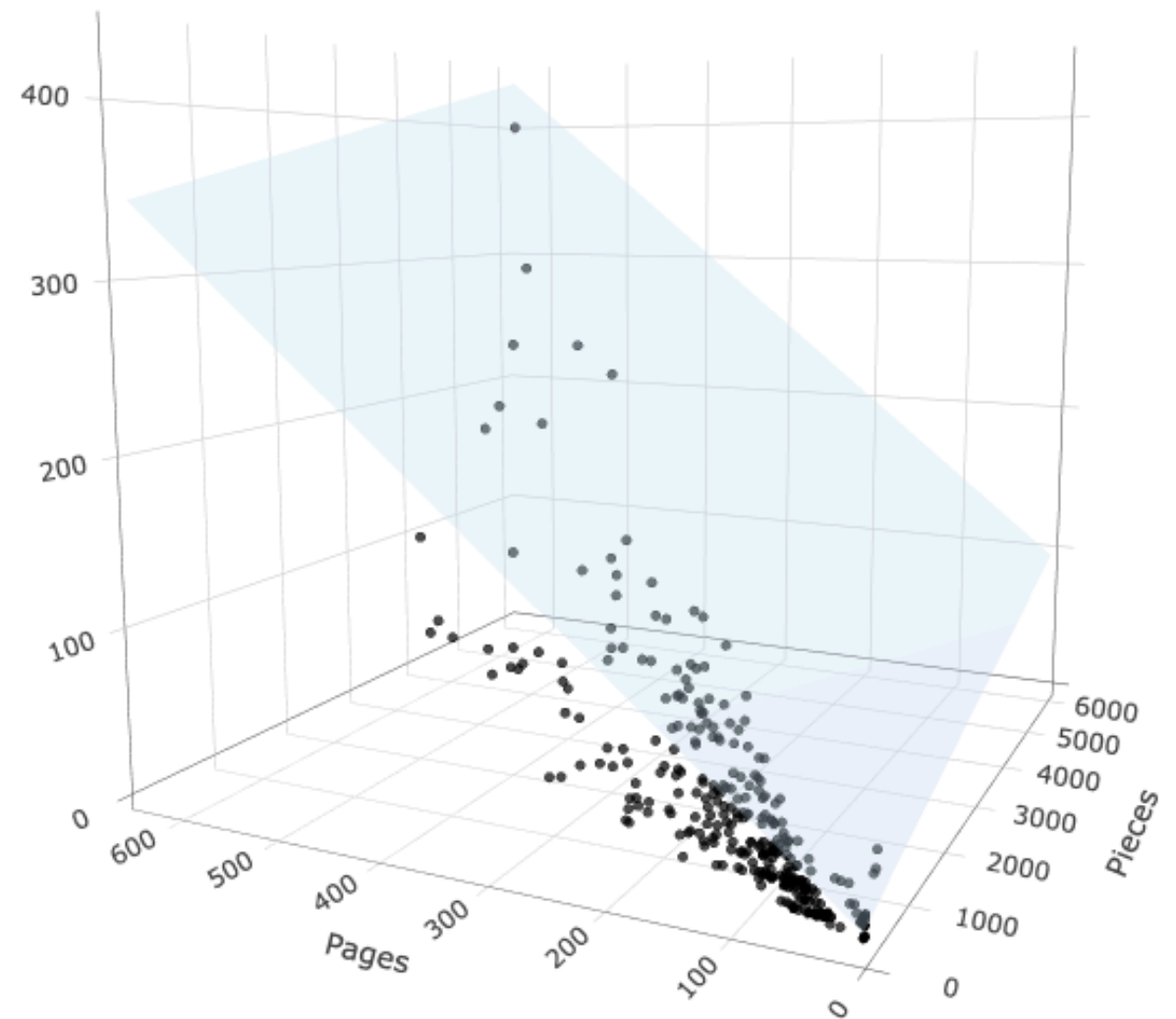
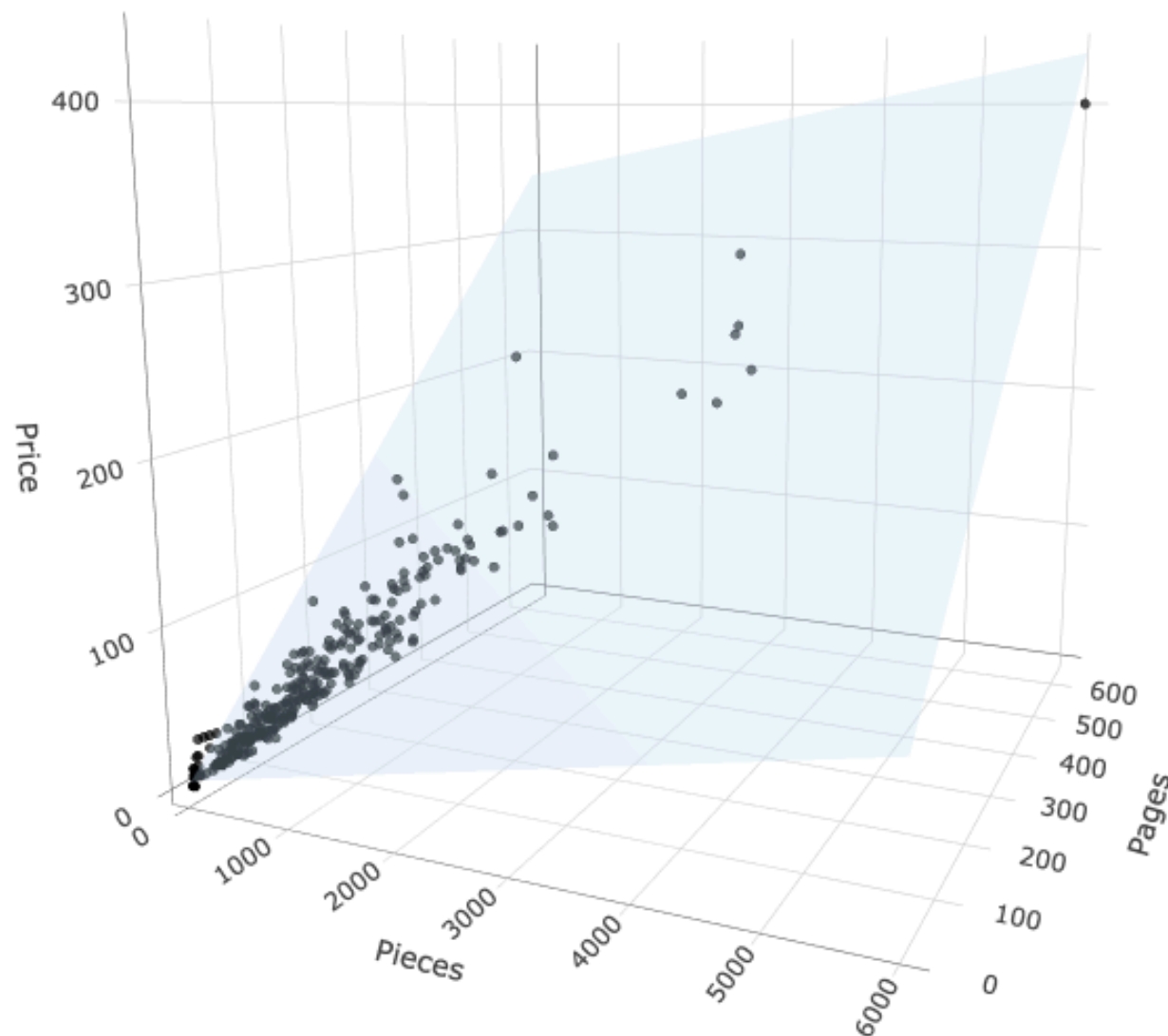
x_{2i} = Antall sider i observasjon i

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

5) Tilpasse og evaluere modell(er)

B. Pris beskrevet av antall brikker og antall sider i manualen

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!



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

$$y \sim x1 + x2$$

5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

C.A. En modell per gruppe (3x enkel LR)

C.B. Felles modell med eget skjæringspunkt for hver gruppe

C.C. Felles modell med eget skjæringspunkt og stigningstall for hver gruppe (*interaksjonseffekt*)

Hvis vi tror at, eller vil undersøke om, pris per LEGO-kloss varierer mellom de grupperte temaene, kan det være gunstig å inkludere en *interaksjonseffekt* i modellen.

Det er et eget notat om interaksjonseffekter, lenke finnes på Wiki sida (samme datasett brukes der, men en annen gruppering!)

5) Tilpasse og evaluere modell(er)

Modell C.A: Egen modell for hver gruppe

y_i^{Cat} = Pris for observasjonen i i gruppe Cat

x_i^{Cat} = Antall brikker i observasjon i i gruppe Cat

$$y_i^{\text{Cat}} = \beta_0^{\text{Cat}} + \beta_1^{\text{Cat}} x_i^{\text{Cat}} + \epsilon_i^{\text{Cat}}$$

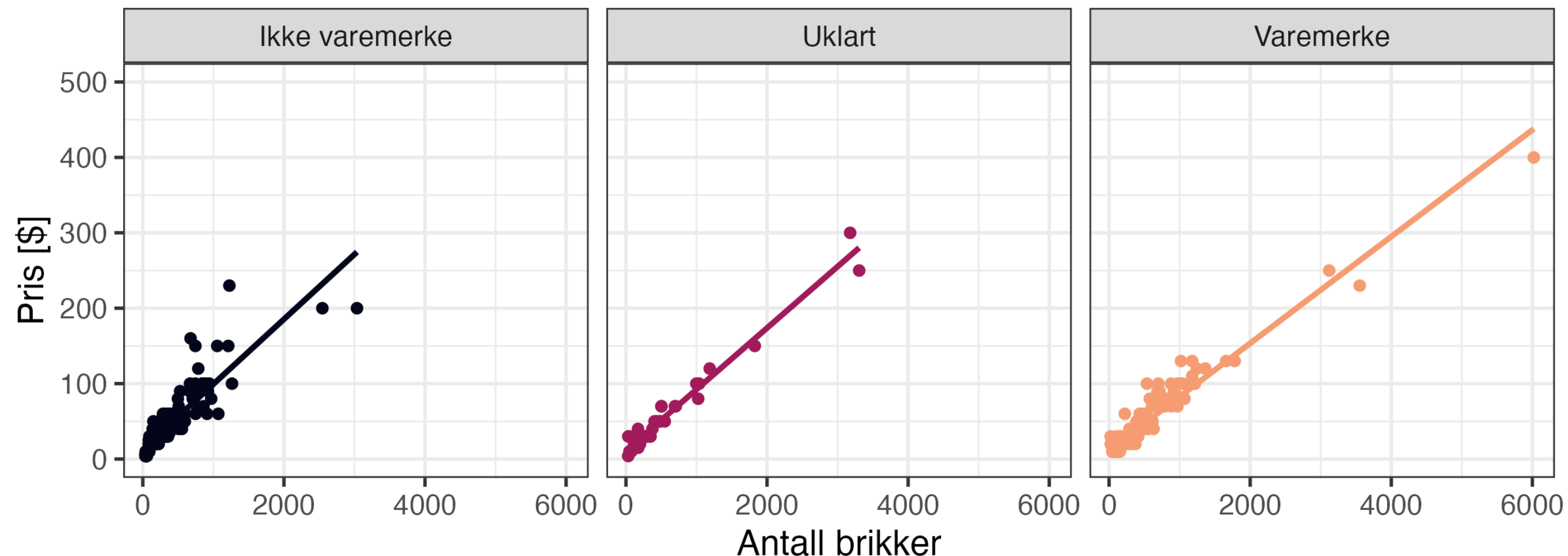
5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

C.A. En modell per gruppe (3x enkel LR)

$Price \sim Pieces$



Totalt $3 \times 3 = 9$ parametere brukt

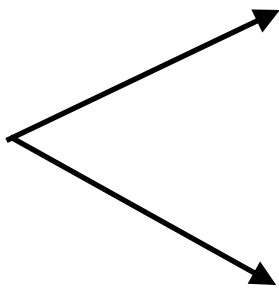
5) Tilpasse og evaluere modell(er)

Modell C.B: Kun eget skjæringspunkt

Y_i = Pris for observasjonen i

x_i = Antall brikker i observasjon i

z_i = Kategori for observasjonen i



$z_{1i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 2} \\ 0 & \text{ellers} \end{cases}$

$z_{2i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 3} \\ 0 & \text{ellers} \end{cases}$

$$Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 x_i + \epsilon_i$$

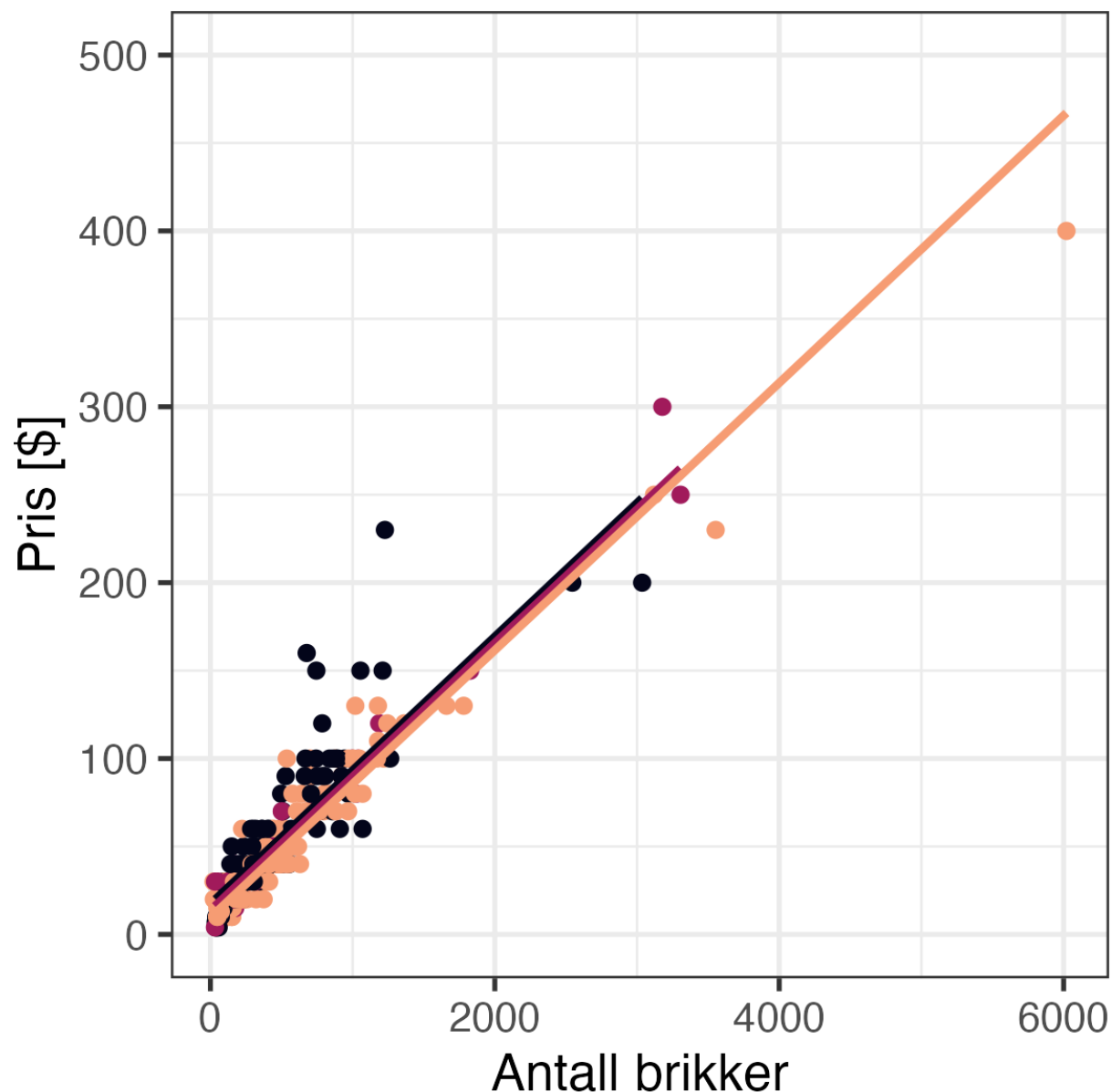
5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

```
Price ~ Pieces + cat
```

C.B. Felles modell med eget skjæringspunkt for hver gruppe



	coef
Intercept	17.2250
cat[T.Uklart]	-3.6094
cat[T.Varemerke]	-7.6550
Pieces	0.0760

Totalt 5 parametere brukt

Justert R^2 fra modell A: 0.873

Justert R^2 fra modell C.B: 0.878

5) Tilpasse og evaluere modell(er)

Modell C.C: Eget skjæringspunkt og stigningstall

Y_i = Pris for observasjonen i

x_i = Antall brikker i observasjon i

z_i = Kategori for observasjonen i

$$z_{1i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 2} \\ 0 & \text{ellers} \end{cases}$$

$$z_{2i} = \begin{cases} 1 & \text{hvis } z_i = \text{Cat 3} \\ 0 & \text{ellers} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 z_{1i} \beta_1 + \beta_2 z_{3i} + (\beta_4 z_{1i} \beta_5 + \beta_6 z_{3i}) x_i + \epsilon_i$$

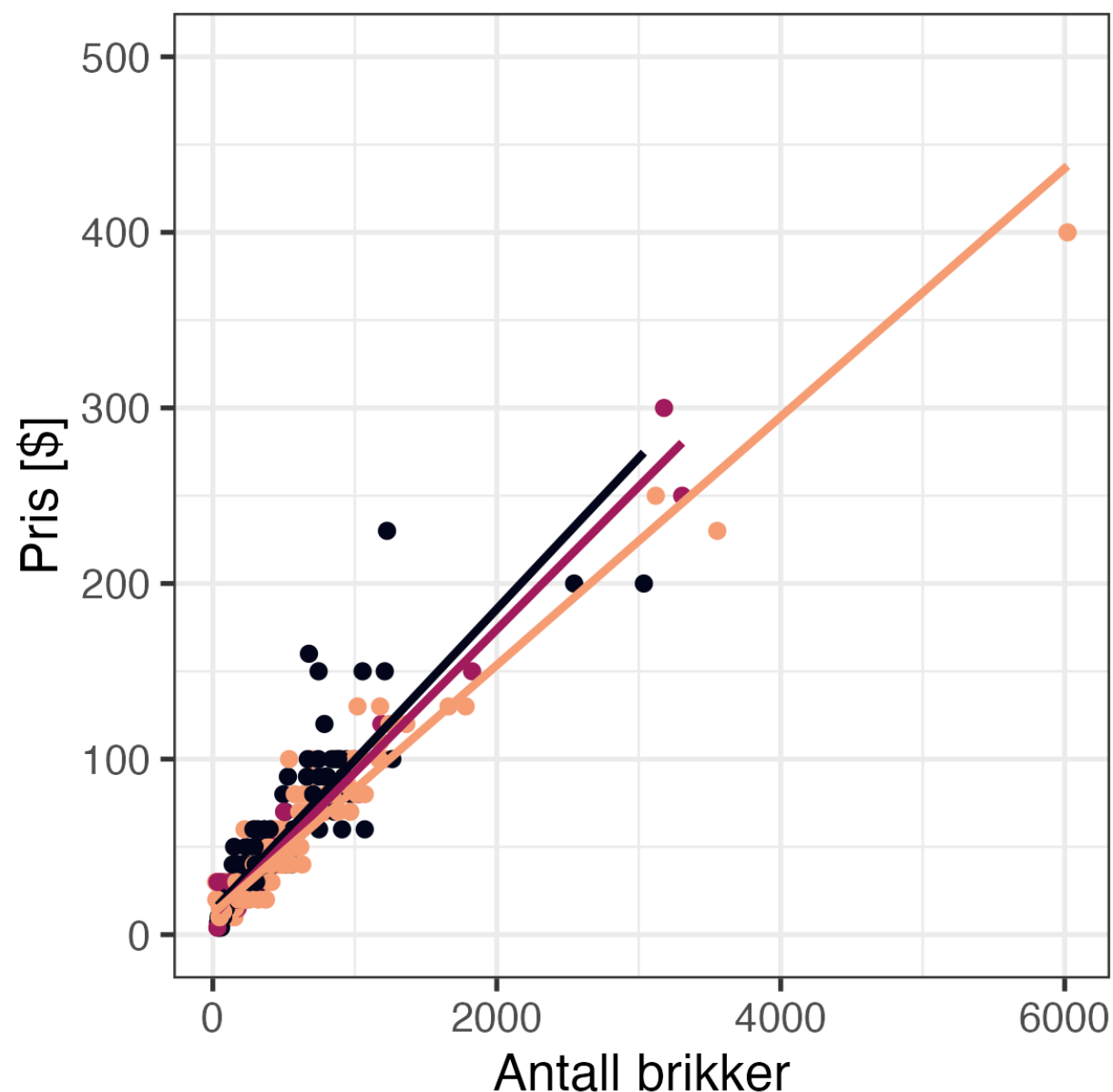
5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

```
Price ~ Pieces * cat
```

C.C. Felles modell med eget skjæringspunkt og stigningstall for hver gruppe



Totalt 7 parametere brukt

Justert R^2 fra modell A: 0.873

Justert R^2 fra modell C.B: 0.878

Justert R^2 fra modell C.C: 0.884

5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

```
Price ~ Pieces * cat
```

C.C. Felles modell med eget skjæringspunkt og stigningstall for hver gruppe

	coef	std err	t	P> t	[0.025	0.975]
Intercept	13.3555	1.859	7.184	0.000	9.699	17.012
cat[T.Uklart]	-2.5100	3.520	-0.713	0.476	-9.434	4.414
cat[T.Varemerke]	-1.3406	2.408	-0.557	0.578	-6.077	3.396
Pieces	0.0860	0.003	26.418	0.000	0.080	0.092
Pieces:cat[T.Uklart]	-0.0046	0.005	-0.965	0.335	-0.014	0.005
Pieces:cat[T.Varemerke]	-0.0153	0.004	-4.038	0.000	-0.023	-0.008

Referansekategori: “Ikke varemerke”

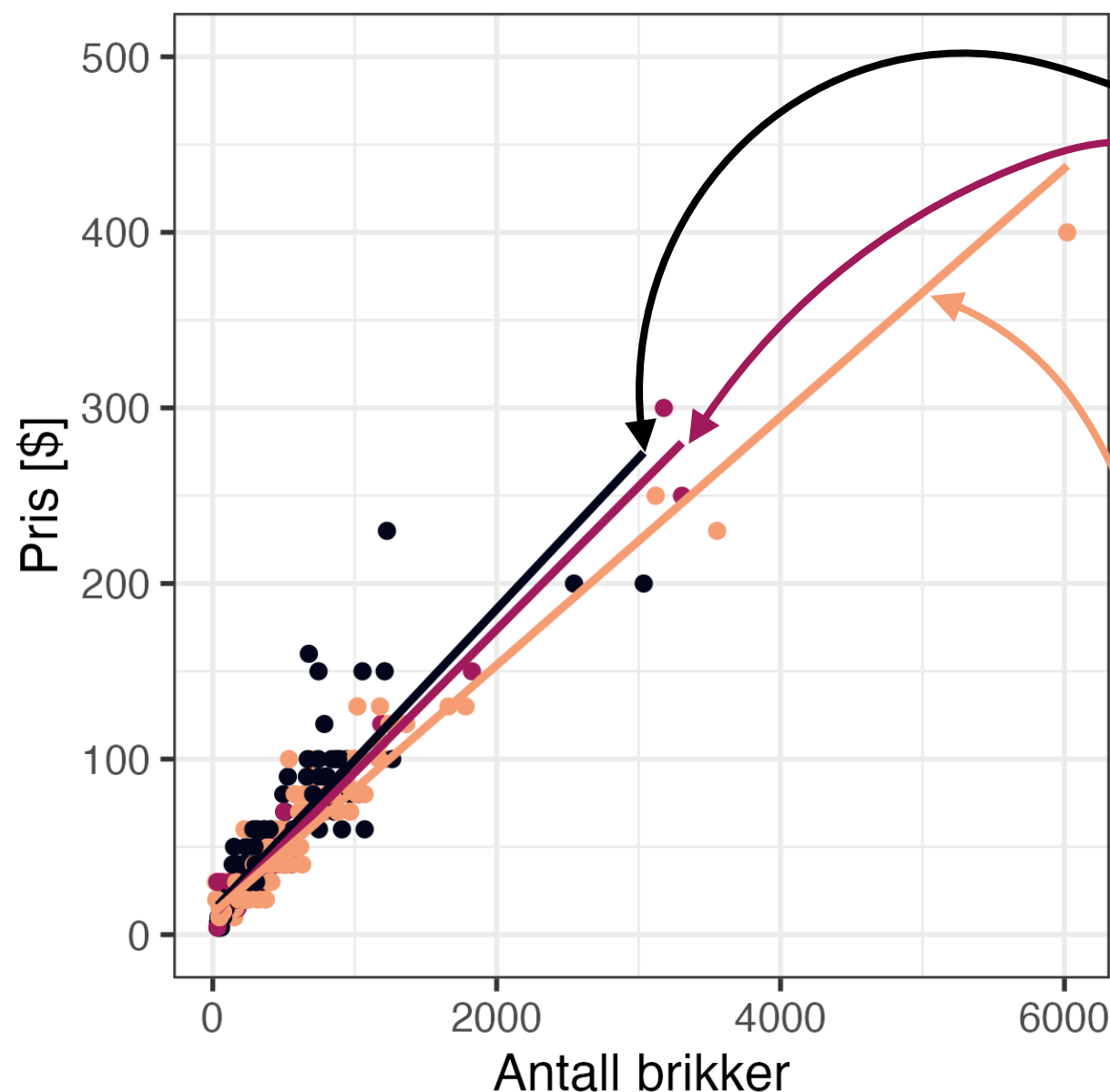
5) Tilpasse og evaluere modell(er)

C. Pris beskrevet av antall brikker, med egne parametere for hver gruppe

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

```
Price ~ Pieces * cat
```

C.C. Felles modell med eget skjæringspunkt og stigningstall for hver gruppe



$$\hat{\beta}_0 + \hat{\beta}_3 x = 13.3 + 0.086 \cdot x$$

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_3 + \hat{\beta}_4) \cdot x \\ = 10.85 + 0.0814 \cdot x \end{aligned}$$

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_3 + \hat{\beta}_5) \cdot x \\ = 12.01 + 0.0707 \cdot x \end{aligned}$$

Oppgave 1

OBS! Jeg bruker bare et subsett av datasettet, dere skal bruke hele!

6) Ble problemstillingen besvart? Konklusjon?

Ble problemstillingen besvart? Hvorfor/hvorfor ikke?

Hva ble konklusjonen?

“Betaler vi ekstra for LEGO som assosieres med et varemerke?”

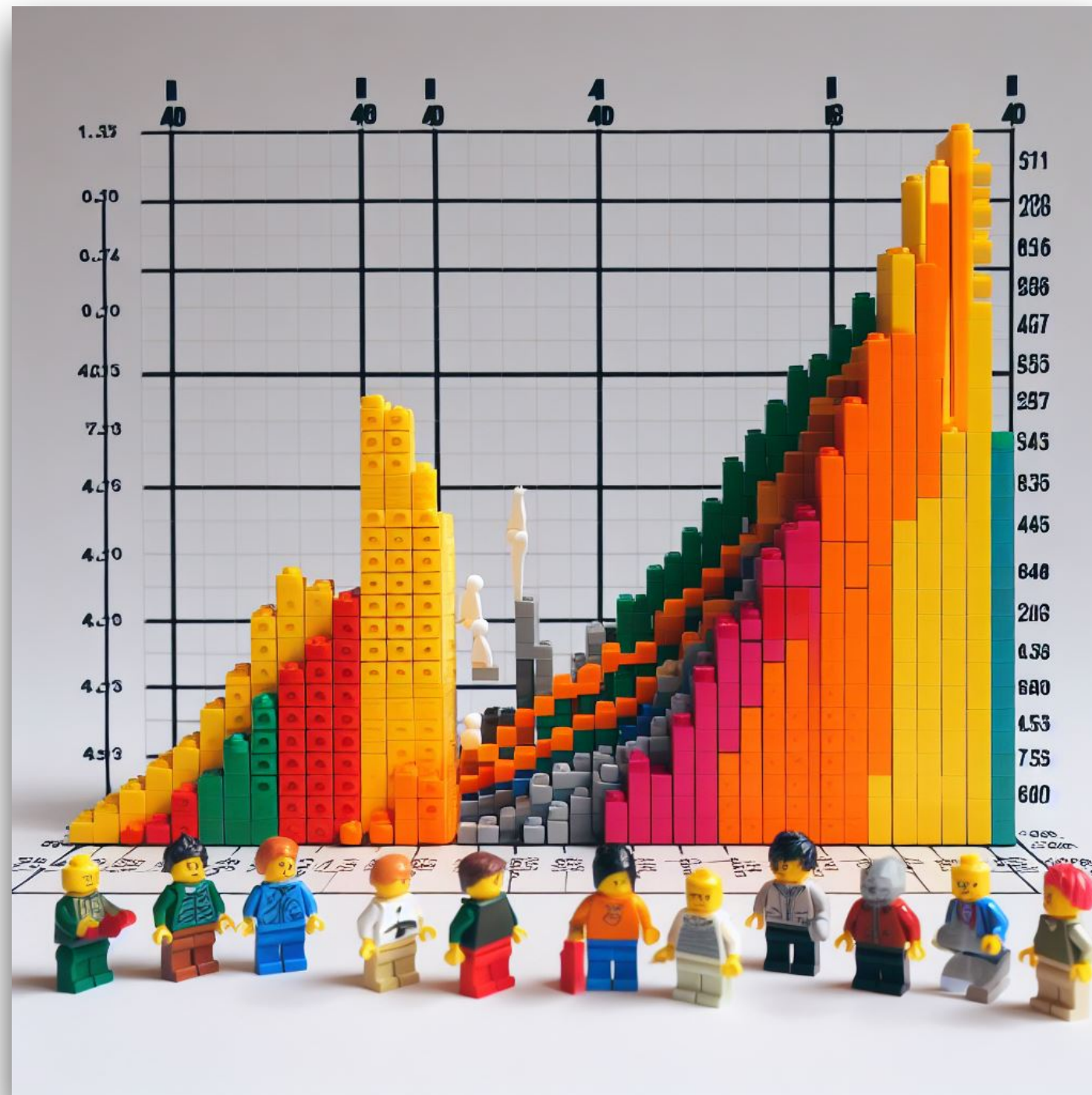
Obs! Det er ikke lov å endre problemstilling bare fordi konklusjonen ble kjedelig. Spennende problemstilling og resultat fører ikke nødvendigvis til en bedre karakter.

Her: LEGO assosiert med varemerker koster signifikant mindre enn LEGO som ikke er assosiert med varemerker. Vi kunne svare på problemstillingen.

Uklart-kategorien kan vi ikke skille fra “ikke varemerke”. Hvis vi vil teste om den er forskjellig fra “varemerke” kan vi endre referansekategori:

```
Price ~ Pieces * C(cat, Treatment("Ikke varemerke"))
```

Opppgave 1: Regresjon



Microsoft bing prompt: "Multiple linear regression visualized using LEGO"

Hva skjer nå? (tirsdag)

Ny forelesning mandag 15:15 (digitalt, samme link).

Fysisk og digital veiledning på torsdag 14:15-16:00.

Neste uke: Klassifikasjon og klyngeanalyse. Se videoer før forelesning for størst utbytte!

OBS! Neste uke vil jeg også gå gjennom spørsmål om prosjektet i plenum, og er avhengig av at dere stiller spørsmål så jeg har noe å svare på!