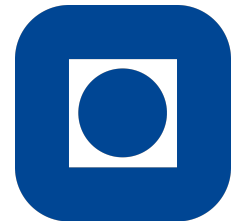


Versjon fra zoom-foredlesningen 28.10.2020.

ISTx1003: Statistisk læring og data science

Zoom-forelesning 2: Multippel lineær regresjon

Mette Langaas, IMF/NTNU. 28.10.2020



NTNU

Plan: Multippel lineær regresjon

1. Læringsmål, pensum og læringsressurser
2. Datasett: Diabetesprogresjon
3. Fra enkel til multippel regresjon (oppsett og resultater, tolke estimerte koeffisienter, prediksjon, KI og hypotesetest)
4. Velge modell med justert R^2
5. Datasett: Leieindeks i München
6. Dummyvariabelkoding av kategoriske kovariater
7. Passer modellen - og hva hvis ikke?
8. Har vi dekket pensum?
9. Etter timen

(Inkludert 15 minutters pause)

Læringsmål: regresjon

- 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 besvare oppgave 1 av prosjektoppgaven på en god måte!**

Pensum er definert som

"svarene på det du blir spurt om på prosjektoppgaven"

og de kan du finne ved å bruke læringsressursene.

Læringsressurser regresjon

Kompendium: Regresjon pdf html

Videoer:

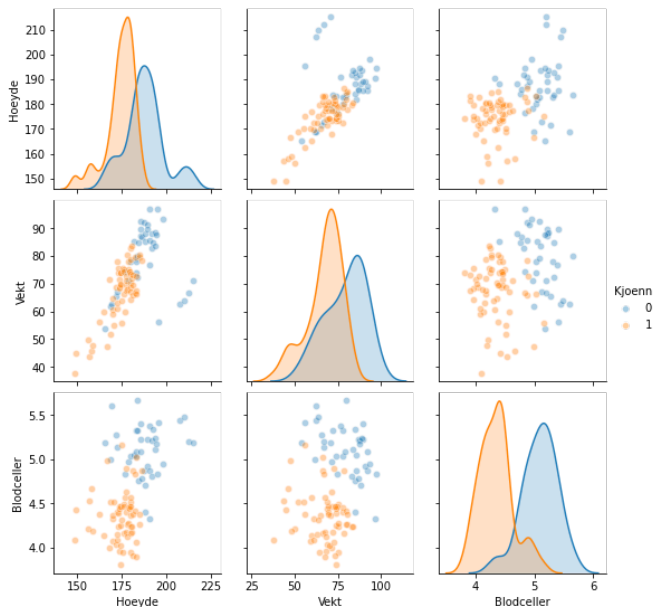
1. Multippel lineær regresjon: introduksjon (14:07 min)

2. Multippel lineær regresjon: analyse av et datasett (15:20 min)

Zoom: slides med notater, to notatbøker (Jupyter og html-versjon)

Regresjon

Veiledet læring - vi kjenner responsen



Prosjektet oppgave 1:

Hva er sammenhengen mellom
antall røde blodceller pr liter blod
- og høyde, vekt, kjønn og
idrettsgren - til idrettsutøvere?

Q1.1 - Q1.8

tolke ulike plott av x og y

estimert regresjon (linje/hyperplan)

regresjonskoeffisienter

velg mellom to modeller

R^2 justert R^2

tolke estimerte koeffisienter

forskjell enkel og multippel

predikert verdi

konfidensintervall

residual

hypotesetest og p-verdi

modellantagelser

hvilke

hvordan
evaluere

kontinuerlig kovariat

residualplott

kategorisk kovariat:
dummyvariabelkoding

QQ-plott

Hvor mye skal man svare på hvert spørsmål: 3-4 setninger (minimum 1 og maksimum 5) **passer modellen?**

Pensum: hva er spurt om i Oppgave 1

Diabetes



Hvilke faktorer kan bidra til forverring av sykdomsforløp?

Respons: prognoseskår *Kontinuerlig, lav verdi er bra*

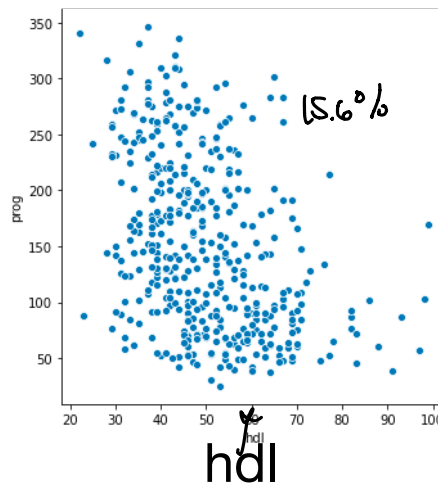
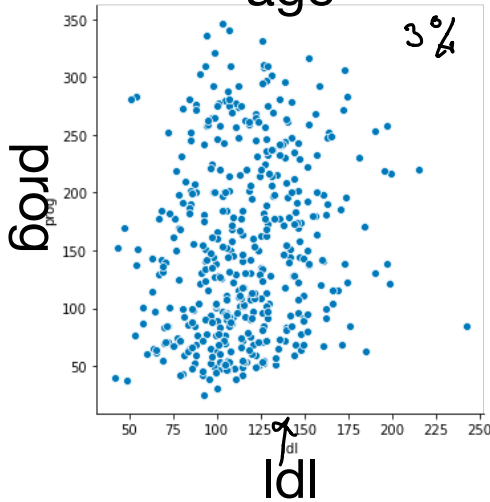
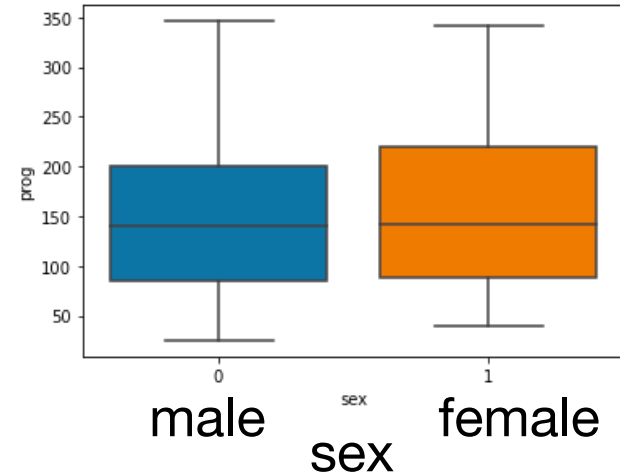
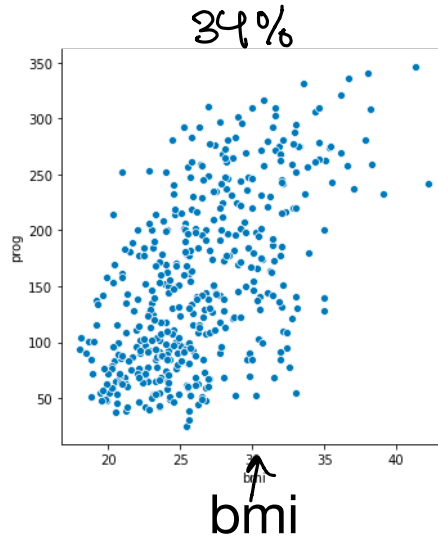
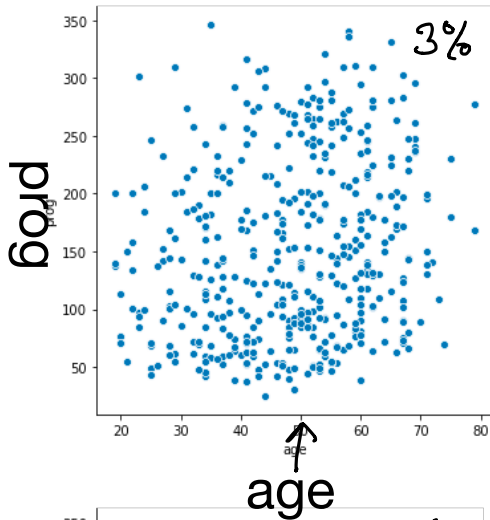
Forklaringsvariabler:

- kjønn
- alder
- kroppsmasseindeks (BMI)
- LDL kolesterolmåling
- HDL kolesterolmåling

Diabetes

tolke ulike plott av x og y (Q1.5)

R^2



Poll: Hvis vi lager en enkel lineær regresjon mellom prognose og hver av age, bmi, ldl og hdl, hvilken regresjon vil ha høyest R^2 ?

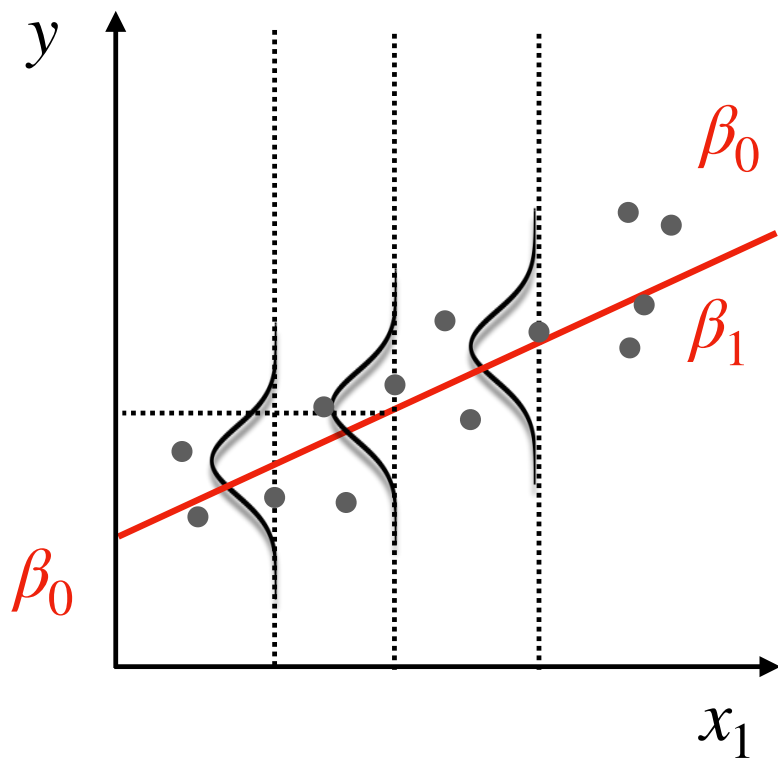
||
bmi

Enkel lineær regresjonsmodell

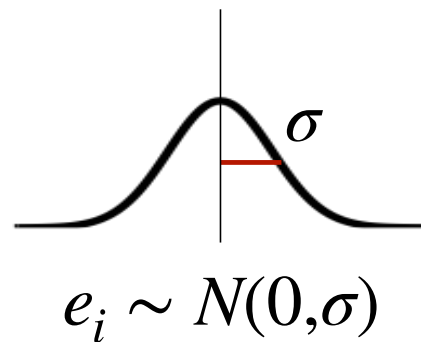
uavhengige
observasjonspar

$$(x_{1i}, y_i)$$

$$i = 1, \dots, n$$

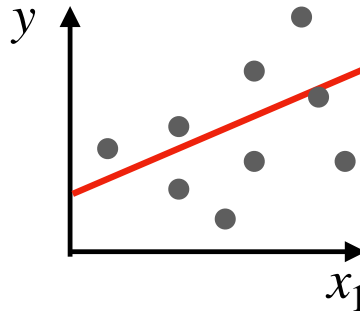


$$Y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

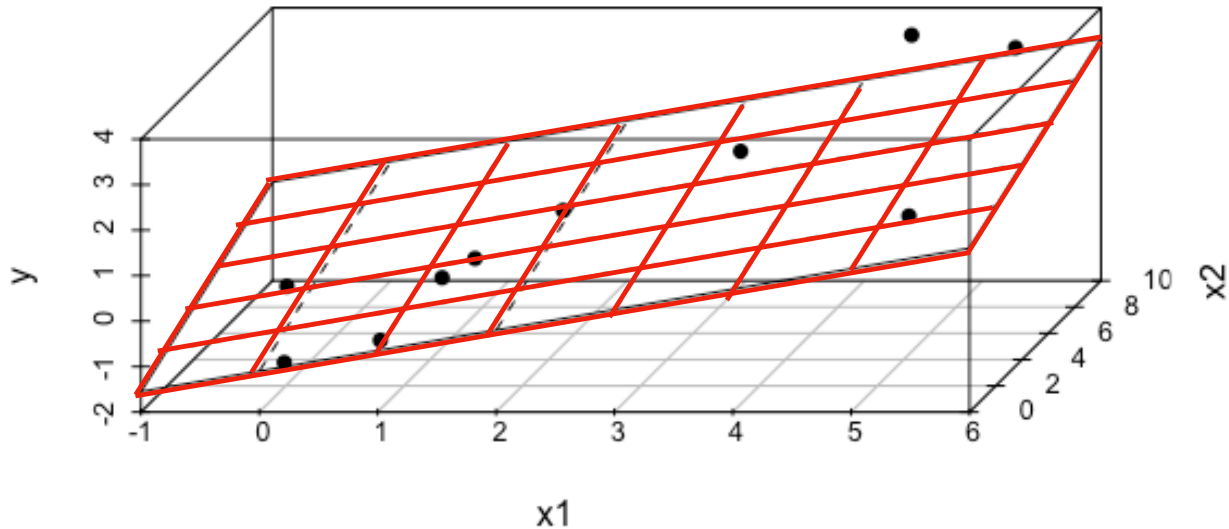


Fra enkel til ...

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



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



Multippel lineær regresjonsmodell

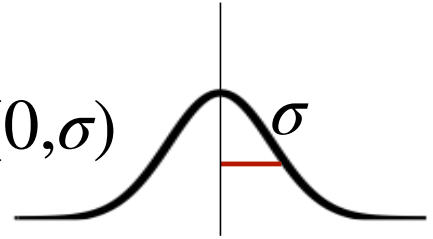
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

uavhengige
observasjonsenheter

$(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$

$i = 1, \dots, n$

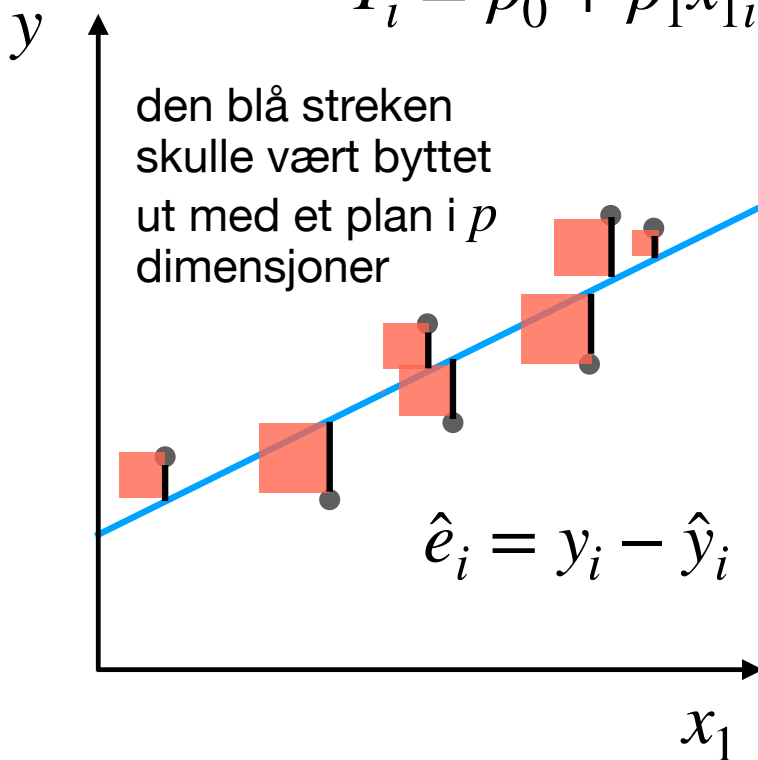
feilledd $e_i \sim N(0, \sigma)$



Parameterestimering

Multipel lineær regresjon

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$$

Finn $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$
ved å minimere

$$\text{SSE} = \sum_{i=1}^n \hat{e}_i^2$$

Diabetes: hva betyr alt dette?

Regressjonsfeilen 39.8%
av variabiliteten i dataen

OLS Regression Results

Dep. Variable:	prog	R-squared:	0.398
Model:	OLS	Adj. R-squared:	0.391
Method:	Least Squares	F-statistic:	57.66
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	5.50e-46
Time:	19:16:01	Log-Likelihood:	-2435.0
No. Observations:	442	AIC:	4882.
Df Residuals:	436	BIC:	4907.
Df Model:	5		
Covariance Type:	nonrobust		

$n - p - 1 = 442 - 6$
 $p = \text{antall kovariater}$

	coef	std err	t	P> t	[0.025	0.975]
0 Intercept	-15.1287	28.518	-0.531	0.596	-71.178	40.920
1 age	0.5883	0.229	2.570	0.011	0.138	1.038
2 sex	-17.2501	6.311	-2.733	0.007	-29.654	-4.846
3 bmi	8.5057	0.722	11.778	0.000	7.086	9.925
4 ldl	-0.0225	0.100	-0.225	0.822	-0.219	0.174
5 hdl	-1.5055	0.258	-5.835	0.000	-2.013	-0.998

β_j

$\hat{\beta}_j$

$SE(\hat{\beta}_j)$

$\frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$

p-verdi
 $H_0: \beta_j = 0$
 $H_1: \beta_j \neq 0$

95% konf. int
 β_j

02235

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-15.1287.	28.518	-0.531	0.596	-71.178	40.920
age	0.5883	0.229	2.570	0.011	0.138	1.038
sex	-17.2501	6.311	-2.733	0.007	-29.654	-4.846
bmi	8.5057	0.722	11.778	0.000	7.086	9.925
ldl	-0.0225	0.100	-0.225	0.822	-0.219	0.174
hdl	-1.5055	0.258	-5.835	0.000	-2.013	-0.998

PREDIKSJON

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{age} \cdot X_{age} + \hat{\beta}_{sex} \cdot X_{sex} + \hat{\beta}_{bmi} \cdot X_{bmi} + \hat{\beta}_{ldl} \cdot X_{ldl} + \hat{\beta}_{hdl} \cdot X_{hdl}$$

$$= -15.1 + 0.59 \cdot X_{age}^{50} - 17.25 \cdot X_{sex}^1 + 8.51 \cdot X_{bmi}^{30} - 0.02 \cdot X_{ldl}^{100} - 1.5 \cdot X_{hdl}^{40}$$

Prediker verdi
for prog.skår
for en pasient
med:

$$\hat{y} = 190.1$$

- age=50
- sex=female
- bmi=30
- ldl=100
- hdl=40

→ Hva betyr $\hat{\beta}_0$?

$\hat{y}$ for en person med
0 for alle kovariater

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-15.1287	28.518	-0.531	0.596	-71.178	40.920
age	0.5883	0.229	2.570	0.011	0.138	1.038
sex	-17.2501	6.311	-2.733	0.007	-29.654	-4.846
bmi	8.5057	0.722	11.778	0.000	7.086	9.925
ldl	-0.0225	0.100	-0.225	0.822	-0.219	0.174
hdl	-1.5055	0.258	-5.835	0.000	-2.013	-0.998

Hva betyr $\hat{\beta}_{hdl} = -1.5$?

nytt, variabel med
for enkel
bin regr.

Vi ser på to personer som er like gamle, samme kjønn, samme BMI og samme verdi for ldl.

La person A ha en enhet hdl høyere enn person B.

Da vil i gjennomsnitt person A ha en prognosestør som er 1.5 lavere enn person B.

Hva betyr $\hat{\beta}_{sex} = -17.25$

mann 0
kvinne 1

Kvinne har i gjennomsnitt en prognosestør som er 17.25 lavere enn menn, hvis de har samme BMI, alder, ldl og hdl.

	coe	std err	t	P> t	[0.025	0.975]
Intercept	-15.1287	28.518	-0.531	0.596	-71.178	40.920
age	0.5883	0.229	2.570	0.011	0.138	1.038
sex	-17.2501	6.311	-2.733	0.007	-29.654	-4.846
bmi	8.5057	0.722	11.778	0.000	7.086	9.925
ldl	-0.0225	0.100	-0.225	0.822	-0.219	0.174
hdl	-1.5055	0.258	-5.835	0.000	-2.013	-0.998

← 95% konfidensintervall

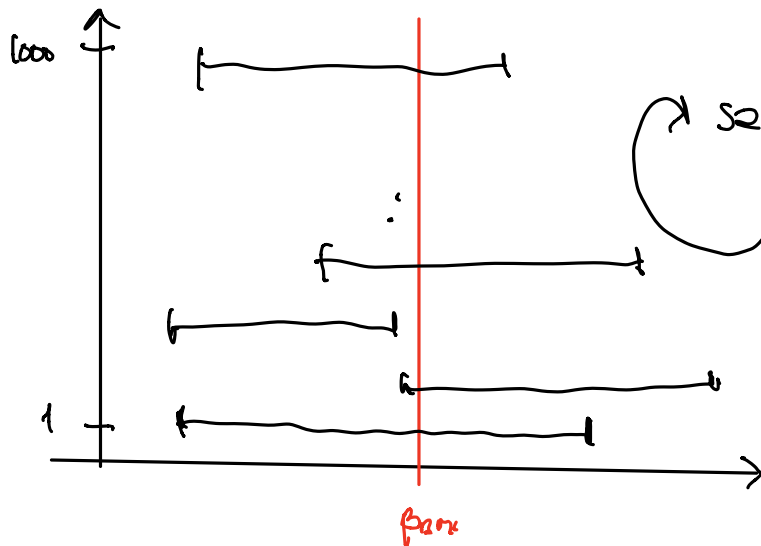
↑
 β_j
↑
børke gjeitt

Intervall i har stor tilfod til at inneholder den samme β_j

$\hat{\beta}_{bmi} = 8.5 \rightarrow$ børke gjeitt for sann algent β_{bmi}

$$\left[\hat{\beta}_j \pm t_{\frac{\alpha}{2}, df} \cdot SE(\hat{\beta}_j) \right]$$

1.96



Smale inn data $\rightarrow$ lage KI for β_{bmi}
gjenta loo

I det lange løp vil 95% av disse intervallene dekke den sanne β_{bmi} .

Husk at alle andre variabler er med i modellen.

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-15.1287	28.518	-0.531	0.596	-71.178	40.920
age	0.5883	0.229	2.570	0.011	0.138	1.038
sex	-17.2501	6.311	-2.733	0.007	-29.654	-4.846
bmi	8.5057	0.722	11.778	0.000	7.086	9.925
ldl	-0.0225	0.100	-0.225	0.822	-0.219	0.174
hdl	-1.5055	0.258	-5.835	0.000	-2.013	-0.998

Hypotesetest

$H_0: \beta_j = 0$ vs. $H_1: \beta_j \neq 0$

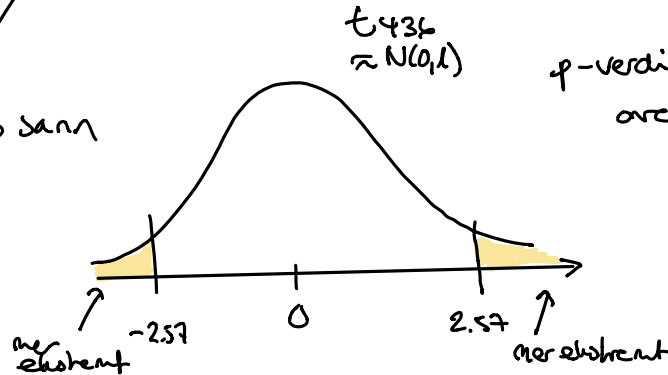
I en modell der alle de andre kovariaterne er med.

$$\hat{\beta}_{age} = 0.59$$

$$t_{age} = \frac{\hat{\beta}_{age} - 0}{SE(\hat{\beta}_{age})}$$

2.57

H_0 sann



p-verdi = sum av de to hale-arealene.

$$p\text{-verdi} = 0.011$$

p-verdi: (uformelt): "Sannsynligheten for det vi har observert eller noe mer ekstremt dersom H_0 er sann"

Vi forkaster H_0 når p-verdi ≤ 0.05 : her: forkast H_0 for $\beta_{age} = 0$

OLS Regression Results

Dep. Variable:	prog	R-squared:	0.398
Model:	OLS	Adj. R-squared:	0.391
Method:	Least Squares	F-statistic:	57.66
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	5.50e-46
Time:	19:16:01	Log-Likelihood:	-2435.0
No. Observations:	442	AIC:	4882.
Df Residuals:	436	BIC:	4907.
Df Model:	5		
Covariance Type:	nonrobust		

$$R^2 = \frac{SST - SSE}{SST}$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$p+1$

$$= 1 - \frac{SSE}{SST}$$

→ vil alltid like eller bli uforenlig
hår nge kovarians legges til

justert R^2 :

$$1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

← brukes til modellvalg

Diabetes - med enda flere kovariater

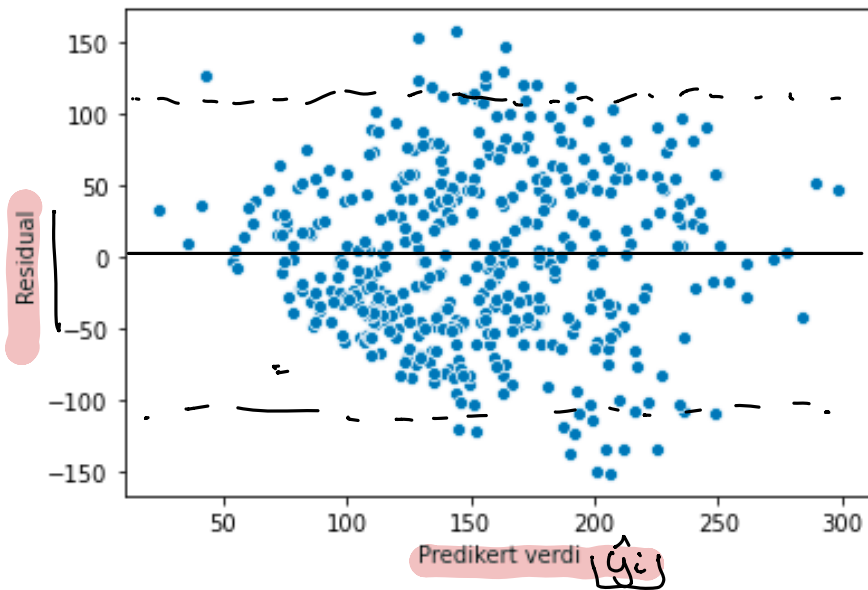
OLS Regression Results

```
=====
Dep. Variable:          prog      R-squared:          0.518
Model:                  OLS      Adj. R-squared:       0.506
Method:                 Least Squares      F-statistic:       46.25
Date:                   Tue, 27 Oct 2020    Prob (F-statistic): 4.01e-62
Time:                   12:34:44    Log-Likelihood:    -2386.0
No. Observations:      442      AIC:              4794.
Df Residuals:          431      BIC:              4839.
Df Model:               10
Covariance Type:       nonrobust
=====
```

bedre modell
enn den
ned
5 kovariater

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    -356.6440      67.020     -5.321      0.000     -488.370     -224.918
age           -0.0353       0.217     -0.163      0.871     -0.462       0.391
sex           -22.7923       5.837     -3.905      0.000     -34.264     -11.321
bmi           5.5955       0.717       7.799      0.000       4.185       7.006
map           1.1159       0.225       4.954      0.000       0.673       1.559
tc            -1.0829       0.573     -1.890      0.059     -2.209       0.043
ldl           0.7391       0.530       1.394      0.164     -0.303       1.781
hdl           0.3678       0.783       0.470      0.639     -1.171       1.906
tch           6.5405       5.960       1.097      0.273     -5.173      18.254
ltg          157.1761      36.048       4.360      0.000      86.324     228.028
glu           0.2815       0.273       1.030      0.304     -0.256       0.819
=====
```

```
=====
Omnibus:          1.482      Durbin-Watson:      2.029
Prob(Omnibus):    0.477      Jarque-Bera (JB):    1.386
Skew:             0.015      Prob(JB):            0.500
Kurtosis:         2.727      Cond. No.            7.76e+03
=====
```



diabetes & 5 korærshr

$$\hat{y}_i$$

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

← passer godt

Ingen grunn til stor bekymring

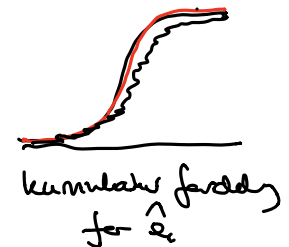
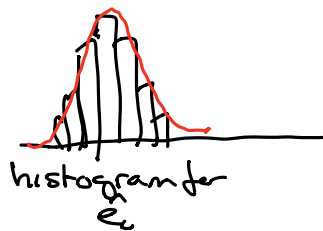
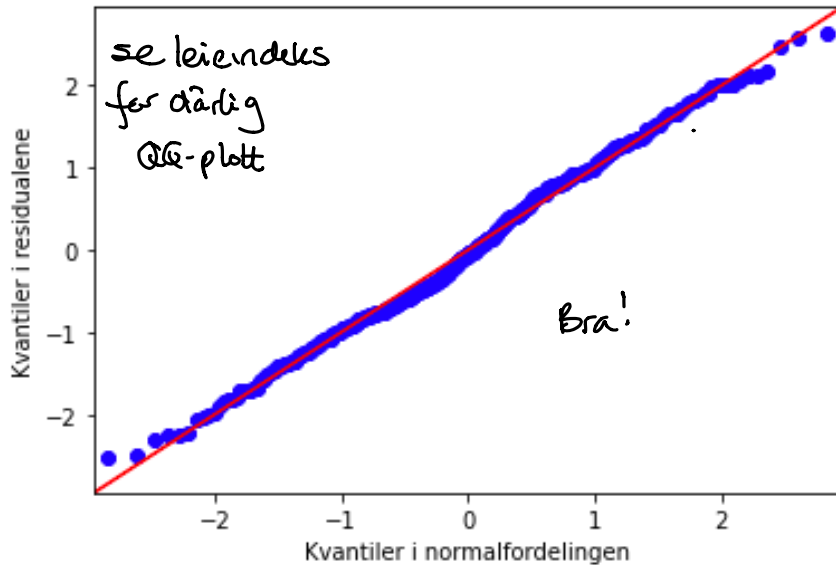
Antagelser

— linear sammenheng x og y

$$e_i \sim N(0, \sigma)$$

uavhengige

for at $\hat{e}_i$ god predikasjon for e_i



Regresjonsanalyse i fem steg

Steg 1: Bli kjent med dataene ved å se på oppsummeringsmål og ulike typer plott

Steg 2: Spesifiser en matematisk modell

Steg 3: Initialiser og tilpass modellen

Steg 4: Presenter resultater fra den tilpassede modellen

Steg 5: Evaluer om modellen passer til dataene

Vi har ikke snakket så mye om hvordan finne en god modell, men har snakket om hvordan sammenligne to modeller med justert R^2 .

tolke ulike plott av x og y

regresjonskoeffisienter

tolke estimerte koeffisienter

forskjell enkel og multippel

konfidensintervall

hypotesetest og p-verdi

kontinuerlig kovariat

kategorisk kovariat:
dummyvariabelkoding

estimert regresjon (linje/hyperplan)

velg mellom to modeller

R^2 justert R^2

$\hat{y}_i$
predikert verdi

$\hat{e}_i = y_i - \hat{y}_i$
residual

modellantagelser

hvilke

hvordan
evaluere

residualplott

QQ-plott

passer modellen?

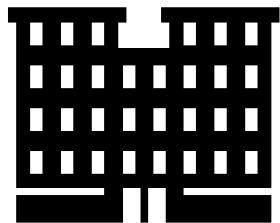
Pensum: hva er spurt om i Oppgave 1

Plan: Multippel lineær regresjon

1. Læringsmål, pensum og læringsressurser
2. Datasett: Diabetesprogresjon
3. Fra enkel til multippel regresjon (oppsett og resultater, tolke estimerte koeffisienter, prediksjon, KI og hypotesetest)
4. Velge modell med justert R^2
5. Datasett: Leieindeks i München
6. Dummyvariabelkoding av kategoriske kovariater
7. Passer modellen - og hva hvis ikke?
8. Har vi dekket pensum?
9. Etter timen

(Inkludert 15 minutters pause)

Hvor mye bør det koste å leie en leilighet?



månedlig
leie (i Euro)

respons

area (i m²)

kvalitet bad

kvalitet kjøkken

sentralvarme

beliggenhet

forklaringsvariabler

kovariater

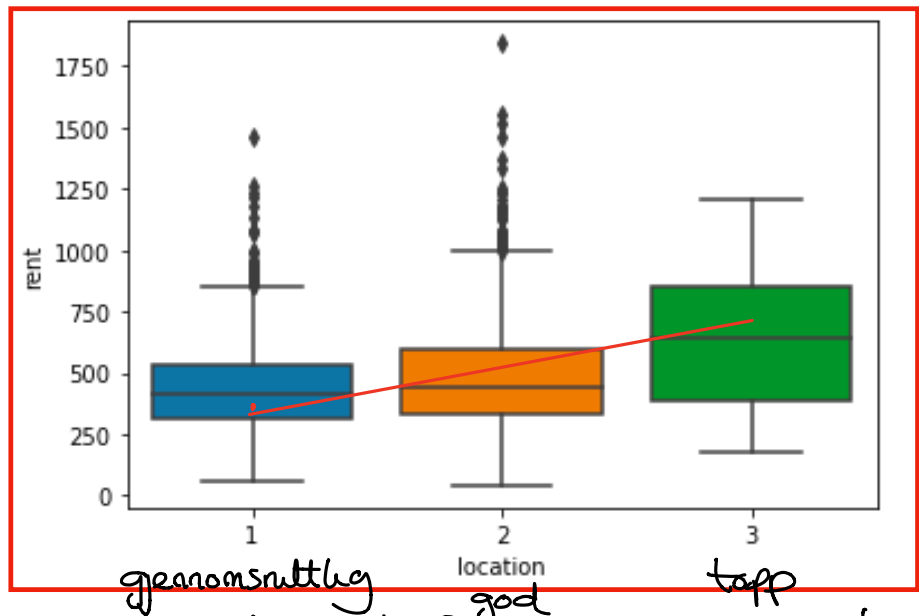
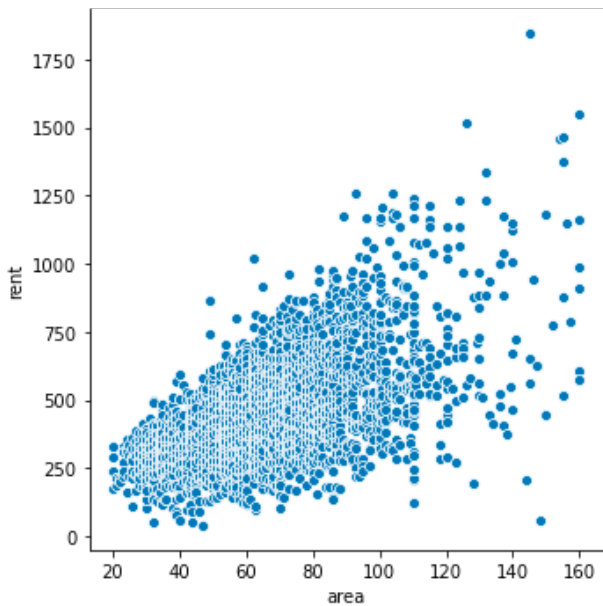
prediktorer

kontinuerlig variabel

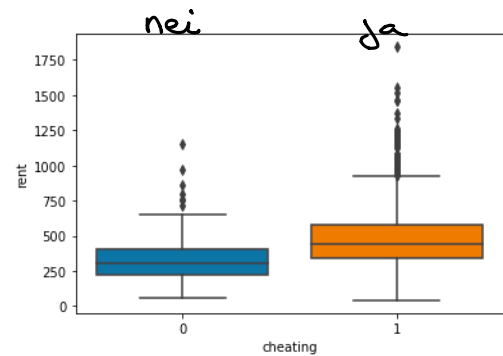
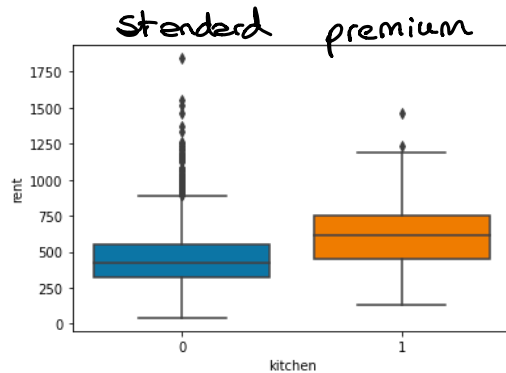
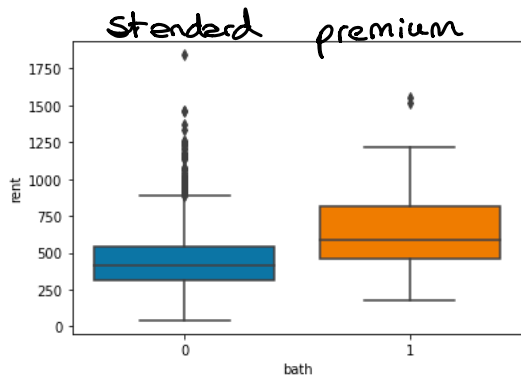
respons

$(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$

forklaringsvariabler - kontinuerlige eller kategoriske



Vi vil ikke tringe modellen til at det nåværende sammenheng fra
gjenn til god som bra god til tapp



Forklaringsvariabler

Kategorisk: kodes ved dummy-variabel koding - kalles one-hot koding i maskinlæring

Kategorisk med tre kategorier: gjennomsnittlig, god og topp beliggenhet.

- vi velger “gjennomsnittlig som referansekategori og lager **to forklaringsvariabler**
- x_{1i} : kodes som 1 hvis beliggenhet er “god” og 0 ellers.
- x_{2i} : kodes som 1 hvis beliggenhet er “topp” og 0 ellers.

Gjennomsnittlig beliggenhet: $x_{1i} = 0$ og $x_{2i} = 0$

God beliggenhet: $x_{1i} = 1$ og $x_{2i} = 0$

Topp beliggenhet: $x_{1i} = 0$ og $x_{2i} = 1$

OLS Regression Results

```
=====
Dep. Variable:          rent    R-squared:                0.450
Model:                  OLS     Adj. R-squared:           0.449
Method:                 Least Squares    F-statistic:              420.0
Date:                  Tue, 27 Oct 2020   Prob (F-statistic):       0.00
Time:                  12:59:56    Log-Likelihood:          -19712.
No. Observations:      3082    AIC:                     3.944e+04
Df Residuals:          3075    BIC:                     3.948e+04
Df Model:              6
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-21.9733	11.655	-1.885	0.059	-44.826	0.879
bath[T.1]	74.0538	11.209	6.607	0.000	52.076	96.031
kitchen[T.1]	120.4349	13.019	9.251	0.000	94.908	145.962
cheating[T.1]	161.4138	8.663	18.632	0.000	144.428	178.400
location[T.2] <i>god</i>	39.2602	5.447	7.208	0.000	28.580	49.941
location[T.3] <i>topp</i>	126.0575	16.875	7.470	0.000	92.971	159.144
area	4.5788	0.114	40.055	0.000	4.355	4.803

$$\hat{y} = -21.97 + 0 + 1 \cdot 126.43 + 0 + 126.06 + 4.57 \cdot 60$$

≈ 500 Euro

Hva er predikert leie for leilighet med

- standard bad
- premium kjøkken
- uten sentralvarme
- med topp beliggenhet
- areal 60m²

tolke ulike plott av x og y

estimert regresjon (linje/hyperplan)

regresjonskoeffisienter

velg mellom to modeller

tolke estimerte koeffisienter

R^2 justert R^2

forskjell enkel og multippel

predikert verdi

konfidensintervall

residual

hypotesetest og p-verdi

modellantagelser

hvilke

hvordan
evaluere

kontinuerlig kovariat

residualplott

kategorisk kovariat:
dummyvariabelkoding

QQ-plott

passer modellen?

Pensum: hva er spurt om i Oppgave 1

Videre denne og neste uken

- Få på plass gruppe for å jobbe med!
- Jobb med Oppgave 1 på prosjektet!
- Still korte spørsmål i Forumet
- og lange spørsmål på digital veiledning tirsdag 12.15-14.

Forberedelser til neste ukes zoom-time

- **Lenker og info er under uke 45: Klassifikasjon**
- Se på videoene (viktigst: den som heter Introduksjon og k-nærmeste nabo) og bla i kompendiet, og
- titt på Prosjektoppgaven Oppgave 2.

Prosjektoppgaven

Vi ser hvor informasjonen ligger på Blackboard og hvordan melde seg på gruppe.

Snarvei til Bb 1003: <https://s.ntnu.no/ist1003>

Vi ser på prosjektoppgaven på <https://s.ntnu.no/isthub>

- Oppgave 1: Regresjon. 8 spørsmåls punkt.
- Oppgave 2: Klassifikasjon. 6 spørsmåls punkt.
- Oppgave 3: Klyngeanalyse. 6 spørsmåls punkt

Jobbe med prosjektoppgaven

Uke- tema- oppgave på prosjektet

Uke 43: Enkel lineær regresjon:

Oppgave 1: Q1.1-Q1.4

Uke 44: Multippel lineær regresjon:

Oppgave 1: Q1.5-Q1.8

Uke 45: Klassifikasjon

Oppgave 2

Uke 46: Klyngeanalyse

Oppgave 3