Lineær regresjon (enkel og multippel)

ISTx1003 Statistisk læring og Data Science

Stefanie Muff, Institutt for matematiske fag

Oktober 20 og 21, 2025

Anerkjennelse

Disse slides bygger på slides fra Mette Langaas, 2020.

Takk til Mette for at jeg fikk bruke noen av materialene.

Plan for i dag og morgen 14:15-15:00 (tema "Regresjon")

- Læringsmål og plan for prosjektmodulen ISTx1003.
- De tre temaene i modulen:
  - regresjon
  - klassifikasjon og
  - klyngeananlyse
- Pensum og læringsressurser
- Prosjektoppgaven og Blackboard-informasjon
- Tema: Enkel og multippel lineær regresjon

Hvem er vi?

- **Studentene**: BIDATA og BDIGSEC, omtrent 250 studenter totalt.
- **Faglig ansvarlig** for innholdet i modulen er Stefanie Muff (stefanie.muff@ntnu.no).
- I **veilederteamet** (for prosjektet) inngår i tillegg
  - Trondheim: studentassistentene Scott du Plessis og Oscar Kehinde Asplin Martins
  - Gjøvik: Bare digital veiledning
  - Ålesund: Digital veiledning og support av Siebe B. van Albada

Hva er statistisk læring og data science?

- *Statistisk læring* inneholder stort sett alle metoder som hjelper oss å lære av data.
- *Data science* er et konsept for å forene statistikk, dataanalyse, informatikk og tilhørende metoder for å "forstå og analysere relle fenomener med data".

Læringsmål (av modulen)

Etter du har gjennomført denne modulen skal du kunne:

- forstå når du kan bruke regresjon, klassifikasjon og klyngeananlyse til å løse et ingeniørproblem
- kunne gjennomføre multippel lineær regresjon på et datasett
- bruke logistisk regresjon og nærmeste nabo for å utføre en klassifikasjonsoppgave
- bruke hierarkisk og k-means klyngeanalyse på et datasett, forstå begrepet avstandsmål
- og kunne kommunisere resultatene fra regresjon/ klassifikasjon/klyngeanalyse til medstudenter og ingeniører
- bli en kritisk leser av resultater fra statistikk/maskinlæring/ statistisk læring/data science/kunstig intelligens når disse rapporteres i media, og forstå om resultatene er realistiske ut fra informasjonen som gis
- kunne besvare prosjektoppgaven på en god måte!

Pensum og læringsressurser

Pensum er definert som "svarene på det du blir spurt om på prosjektoppgaven'' og de kan du finne ved å bruke læringsressursene.

Alle ressurser er tilgjengelig her:

https://wiki.math.ntnu.no/istx100y/2025h/1003

Tema Regresjon:

- **Kompendium**: Regresjon (pdf og html, by Mette Langaas)
- **Korte videoer**: (by Mette Langaas)
  - Multippel lineær regresjon: introduksjon (14:07 min)
  - Multippel lineær regresjon: analyse av et datasett (15:20 min)
- Denne forelesningen
- **Disse slides** med notater

Prosjektoppgaven

- Vi ser hvor informasjonen ligger på Blackboard.
- Vi ser på prosjektoppgaven på https://s.ntnu.no/isthub. Velg mappe **ProjectISTx1003**.
- Karakteren teller 30% til den endelige karakteren.
- Vi bruker prosentvurderingsmetoden: Konverterer poengene i en % (heltall, avrundet) og så bruker vi følgende skala:

Karakterskala for prosentvurderingsmetoden

A: 89–100 poeng

B: 77–88 poeng

C: 65–76 poeng

D: 53–64 poeng

E: 41–52 poeng

F: 0–40 poeng

Veiledning til prosjektoppgaven

Fysisk veiledning for Trondheim:

Alle fredager: 24.10, 31.10., 7.11, og 14.11., 10:15-12:00, Sentralbygg S5.

Digital veiledning for Ålesund og Gjøvik:

Veiledning via *Whereby*

- Man 3.11. og 10.11., 14:15-15:00
- Tir 4.11. og 11.11., 14:15-16:00

Det er et kø-system så dere må evt vente litt:

https://whereby.com/stefanies-whereby

Mattelab forum for ALLE

- https://mattelab2025h.math.ntnu.no/c/istx1003-project-module/57
- Åpent 24/7, men stenger fredag, 14.11., 17:00.

Læringsmål for regresjon

- Du kan lage en modell for å forstå sammenhengen mellom en respons og én eller flere forklaringsvariabler.
- Du kan lage en modell for å predikere en respons fra en eller flere forklaringsvariabler.

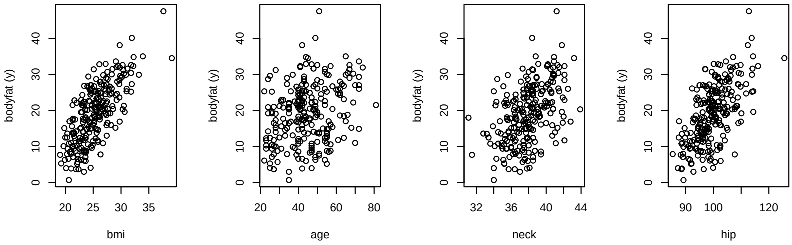
Regresjon – motiverende eksempel

(Veiledet læring - vi kjenner responsen)

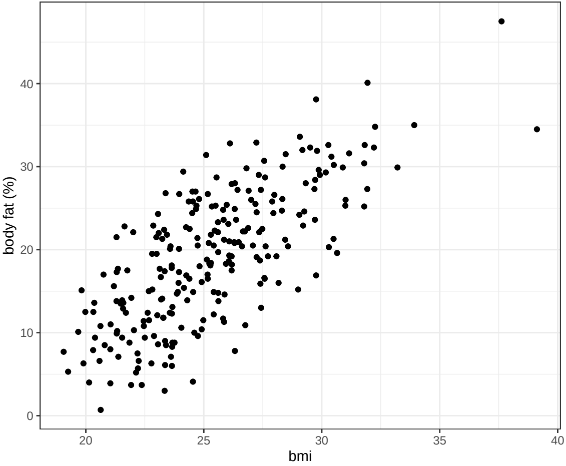
- Kropssfett er en viktig indikator for overvekt, men vanskelig å måle.

Spørsmål: Hvilke faktorer tillater præsis estimering av kroppsfettet?

Vi undersøker 243 mannlige deltakere. Kroppsfett (%), BMI og andre forklaringsvariabler ble målet. Kryssplott:



For en model for funker god for prediksjon trenger vi *multippel lineær regresjon*. Men vi begynner med *enkel lineær regresjon* (bare en forklaringsvariabel):



Enkel lineær regresjon

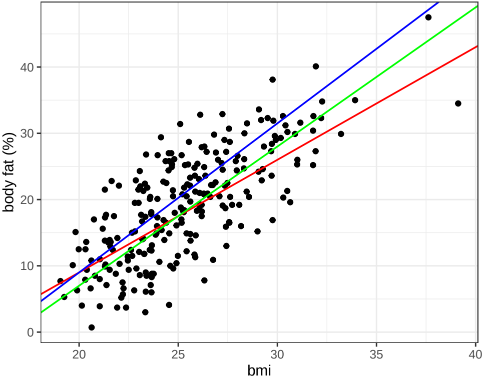
- En kontinuerlig respons variabel Y
- Bare *en forklaringsvariabel* x_1
- Relasjon mellom Y og x_1 er antatt å være *lineær*.

Hvis den lineære relasjonen mellom Y og x er perfekt, så gjelder

$$y_i = \beta_0 + \beta_1 x_{1i}$$

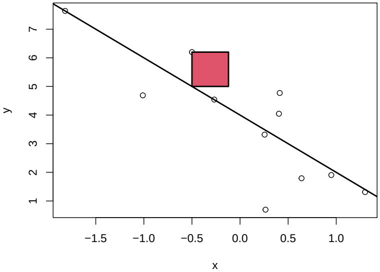
for alle i. Men..

Hvilken linje er best?



Enkel lineær regresjon

a) Kan vi tilpasse den "rette" linjen til dataene?



- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$.
- $\hat{e}_i = \hat{y}_i - y$
- $\hat{\beta}_0$ og $\hat{\beta}_1$ velges slik at

$$SSE = \sum_i \hat{e}_i^2$$

b) Kan vi tolke linja? Hvor sikkert er jeg på $\hat{\beta}_1$ og linja? Vi trenger antakelser, KI og hypothesetest.

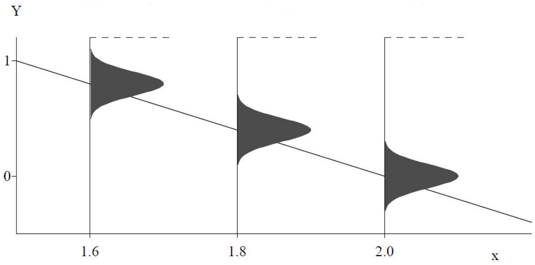
c) Fremtidige presisjoner av predikert y (kroppsfett)?

Lineær regresjon – antakelser

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{i1}}_{\hat{y}_i} + e_i$$

med

$$e_i \sim \mathsf{N}(0, \sigma^2)\ .$$



Do-it-yourself “by hand”

Her kan du finne de beste parametrene selv:

Bruk denne lenken:

https://gallery.shinyapps.io/simple_regression/

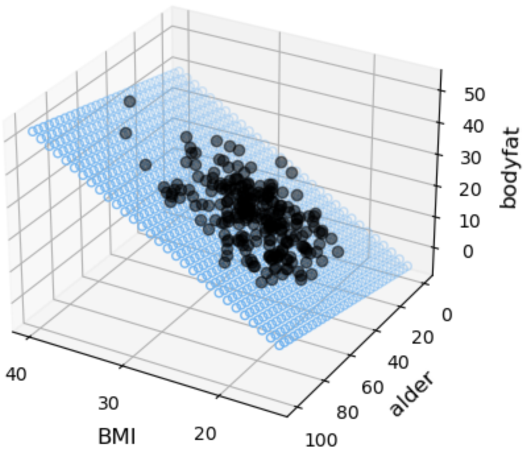
Multippel lineær regresjon

Nesten det samme some enkel lineær regresjon, vi bare summerer flere forklaringsvariabler:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + e_i \ , \quad e_i \sim \mathsf{N}(0, \sigma^2) \ .$$

For eksempel:

$$\text{bodyfat}_i = \beta_0 + \beta_1 \text{bmi}_i + \beta_2 \text{age}_i + e_i \ .$$



0
10
20
30
40
50
bodyfat
0
20
40
60
80
100
alder
40
30
20
BMI

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: Tilpass modellen

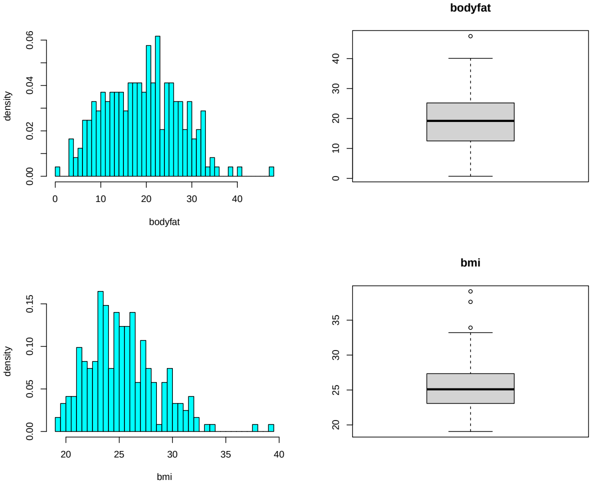
Steg 4: Presenter resultatene fra den tilpassede modellen

Steg 5: Evaluer om modellen passer til dataene

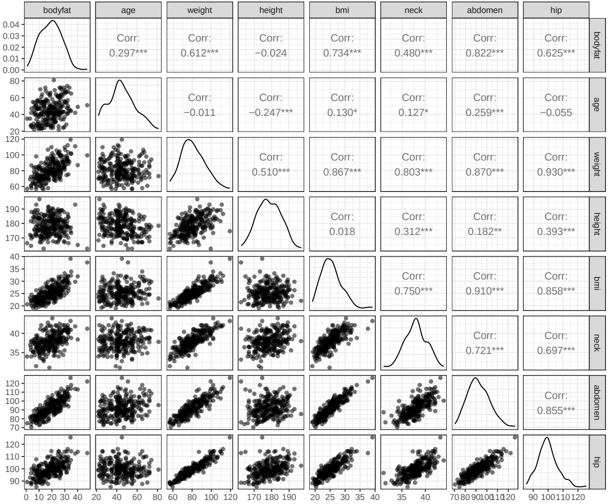
Vi skal ikke snakke så mye om hvordan man finner en god modell, men om hvordan man sammenligner to modeller (med justert R^2) .

Steg 1: Bli kjent med dataene

Vi kan for eksempel se på histogram og boxplot:



Ellers en *parplot* med kryssplotter for alle forklaringsvariable(r) $(x_1, \ldots, x_p)$ og respons y:



```
##     bodyfat           age              weight           height
##  Min.   : 0.70   Min.   :22.00   Min.   : 56.75   Min.   :162.6
##  1st Qu.:12.50   1st Qu.:35.50   1st Qu.: 72.30   1st Qu.:173.7
##  Median :19.20   Median :43.00   Median : 80.02   Median :177.8
##  Mean   :19.11   Mean   :44.83   Mean   : 80.91   Mean   :178.5
##  3rd Qu.:25.20   3rd Qu.:54.00   3rd Qu.: 89.32   3rd Qu.:183.5
##  Max.   :47.50   Max.   :81.00   Max.   :119.29   Max.   :196.8
##       bmi             neck           abdomen            hip
##  Min.   :19.06   Min.   :31.10   Min.   : 70.40   Min.   : 85.30
##  1st Qu.:23.07   1st Qu.:36.40   1st Qu.: 84.90   1st Qu.: 95.55
##  Median :25.10   Median :38.00   Median : 91.00   Median : 99.30
##  Mean   :25.34   Mean   :37.96   Mean   : 92.38   Mean   : 99.69
##  3rd Qu.:27.34   3rd Qu.:39.40   3rd Qu.: 99.15   3rd Qu.:103.15
##  Max.   :39.12   Max.   :43.90   Max.   :126.20   Max.   :125.60
```

I Python får du en oppsummering av datasettet (`df`) med `df.describe()`.

Steg 2: Spesifiser modellen

Nå må vi spesifisere en modell med å velge hvilke forklaringsvariabler vi vil bruke

$$y \sim x_1 + x_2 + x_3 \ .$$

I Python er det

```
formel='y ~ x1 + x2 + x3'.
```

Eksempel 1:

$$\text{bodyfat} \sim \text{bmi}$$

hvis den matematiske modellen er

$$\text{bodyfat}_i = \beta_0 + \beta_1 \text{BMI}_i + e_i \ ,$$

Python: `formel='bodyfat ~ bmi'`.

Eksempel 2:

$$\text{bodyfat} \sim \text{bmi} + \text{age}$$

hvis den matematiske modellen er

$$\text{bodyfat}_i = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{age}_i + e_i \ .$$

Python: `formel='bodyfat ~ bmi + age'`.

Steg 3: Tilpass modellen

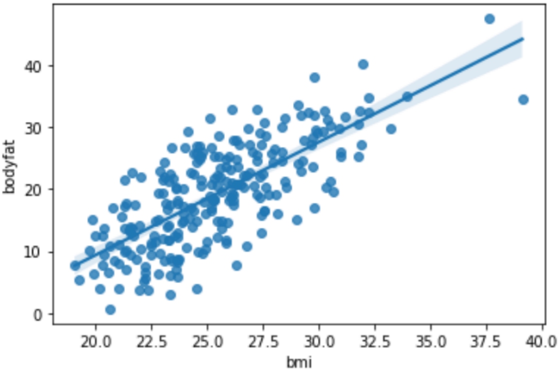
"Tilpasse" betyr:

- Vi estimerer β_0, β_1, ... , og vi får estimater $\hat{\beta}_0$, $\hat{\beta}_1$,...
- I tillegg estimerer vi også σ^2.

Steg 4: Resultat og tolkning av estimatene

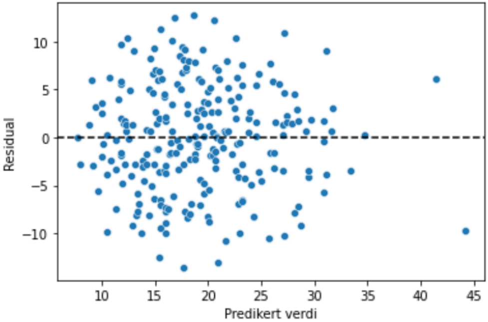
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                bodyfat   R-squared:                       0.539
Model:                            OLS   Adj. R-squared:                  0.537
Method:                 Least Squares   F-statistic:                     281.8
Date:                Wed, 08 Sep 2021   Prob (F-statistic):           2.06e-42
Time:                        18:58:47   Log-Likelihood:                -761.28
No. Observations:                 243   AIC:                             1527.
Df Residuals:                     241   BIC:                             1534.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -26.9844      2.769     -9.746      0.000     -32.439     -21.530
bmi            1.8188      0.108     16.788      0.000       1.605       2.032
==============================================================================
Omnibus:                        5.031   Durbin-Watson:                   2.311
Prob(Omnibus):                  0.081   Jarque-Bera (JB):                3.079
Skew:                          -0.033   Prob(JB):                        0.215
Kurtosis:                       2.452   Cond. No.                         198.
==============================================================================
```

- Tilpasset regresjonslinje og 95% konfidensintervall for regressjonslinja (forventningsverdien $\mathrm{E}(Y)$).
- 95% prediksjonsintervall for nye observasjoner (kroppsfett for nye personer; håndtegnet).



Steg 5: Passer modellen?

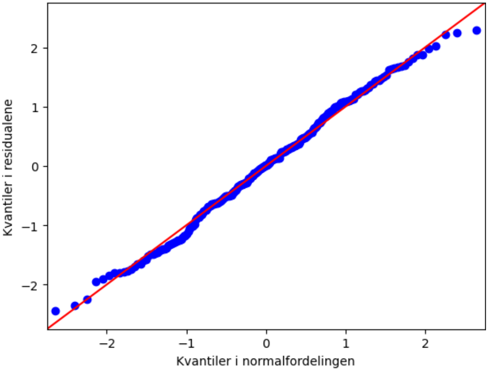
Tukey-Anscome (TA) diagram:



Her vil man

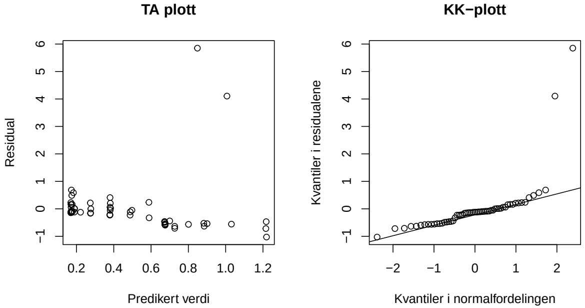
- Ikke noe struktur
- Sentrering rundt 0 verdien

Kvantil-kvantil plot:



Her ser man at observasjonene ligger mer og mindre på linja.

Hvordan ser det ut når en modell *ikke* passer?



Multippel lineær regresjon

Gjenta samme analyse med to kovariabler (`formel='bodyfat ~ bmi + age'`):

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                bodyfat   R-squared:                       0.580
Model:                            OLS   Adj. R-squared:                  0.577
Method:                 Least Squares   F-statistic:                     165.9
Date:                Wed, 08 Sep 2021   Prob (F-statistic):           5.67e-46
Time:                        19:55:28   Log-Likelihood:                -749.88
No. Observations:                 243   AIC:                             1506.
Df Residuals:                     240   BIC:                             1516.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -31.2545      2.790    -11.203      0.000     -36.750     -25.759
bmi            1.7526      0.104     16.773      0.000       1.547       1.958
age            0.1327      0.027      4.857      0.000       0.079       0.186
==============================================================================
```

Med fem kovariabler:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                bodyfat   R-squared:                       0.726
Model:                            OLS   Adj. R-squared:                  0.720
Method:                 Least Squares   F-statistic:                     125.3
Date:                Thu, 09 Sep 2021   Prob (F-statistic):           1.73e-64
Time:                        09:16:49   Log-Likelihood:                -698.26
No. Observations:                 243   AIC:                             1409.
Df Residuals:                     237   BIC:                             1429.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -35.2802      6.073     -5.809      0.000     -47.245     -23.316
bmi            0.3881      0.224      1.730      0.085      -0.054       0.830
age            0.0038      0.027      0.141      0.888      -0.050       0.058
weight        -0.1141      0.029     -3.883      0.000      -0.172      -0.056
neck          -0.4581      0.216     -2.123      0.035      -0.883      -0.033
abdomen        0.8888      0.085     10.486      0.000       0.722       1.056
==============================================================================
Omnibus:                        5.492   Durbin-Watson:                   2.345
Prob(Omnibus):                  0.064   Jarque-Bera (JB):                3.310
Skew:                           0.056   Prob(JB):                        0.191
Kurtosis:                       2.439   Cond. No.                     4.64e+03
==============================================================================
```

Hva betyr alt dette?

- `coef`: $\hat{\beta}_j$
- `std err`: $\widehat{\mathrm{SE}}(\hat{\beta}_j)$
- `t`: $\frac{\hat{\beta}_j - 0}{\mathrm{SE}(\hat{\beta}_j)}$
- `P>|t|`: p-verdi (obs! $p = 0.000$ er *ikke* mulig, det betyr egentlig $p < 0.0005$)

Hva betyr alt dette?

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-35.2802	6.073	-5.809	0.000	-47.245	-23.316
bmi	0.3881	0.224	1.730	0.085	-0.054	0.830
age	0.0038	0.027	0.141	0.888	-0.050	0.058
weight	-0.1141	0.029	-3.883	0.000	-0.172	-0.056
neck	-0.4581	0.216	-2.123	0.035	-0.883	-0.033
abdomen	0.8888	0.085	10.486	0.000	0.722	1.056

Prediksjon:

$\hat{y} =$

Prediker bodyfat for en ny person med
`bmi=25`, `age=50`, `weight=75`, `neck=40`, `abdomen=95`:

$\hat{y} =$

$= 21.88$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-35.2802	6.073	-5.809	0.000	-47.245	-23.316
bmi	0.3881	0.224	1.730	0.085	-0.054	0.830
age	0.0038	0.027	0.141	0.888	-0.050	0.058
weight	-0.1141	0.029	-3.883	0.000	-0.172	-0.056
neck	-0.4581	0.216	-2.123	0.035	-0.883	-0.033
abdomen	0.8888	0.085	10.486	0.000	0.722	1.056

- Hva betyr $\hat{\beta}_0$?
- Hva betyr $\hat{\beta}_{abdomen} = 0.89$?

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-35.2802	6.073	-5.809	0.000	-47.245	-23.316
bmi	0.3881	0.224	1.730	0.085	-0.054	0.830
age	0.0038	0.027	0.141	0.888	-0.050	0.058
weight	-0.1141	0.029	-3.883	0.000	-0.172	-0.056
neck	-0.4581	0.216	-2.123	0.035	-0.883	-0.033
abdomen	0.8888	0.085	10.486	0.000	0.722	1.056

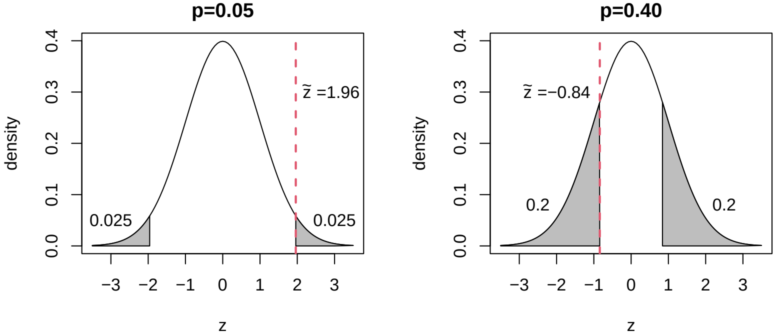
- 95% konfidensintervall: Intervall vi har stor tro at den inneholder den sanne stigningen β_j.
- $[\hat{\beta}_j \pm \underbrace{t_{\alpha/2,df}}_{\approx 1.96} \cdot \text{SE}(\hat{\beta}_j)]$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-35.2802	6.073	-5.809	0.000	-47.245	-23.316
bmi	0.3881	0.224	1.730	0.085	-0.054	0.830
age	0.0038	0.027	0.141	0.888	-0.050	0.058
weight	-0.1141	0.029	-3.883	0.000	-0.172	-0.056
neck	-0.4581	0.216	-2.123	0.035	-0.883	-0.033
abdomen	0.8888	0.085	10.486	0.000	0.722	1.056

- p-verdier og hypotesetester

Recap: Formell definisjon av p-verdien

p**-verdien** er sannsynligheten for det vi *har* observert eller noe mer ekstremt, dersom H_0 er sant.



R^2 og justert R^2

Dep. Variable:	bodyfat	R-squared:	0.726
Model:	OLS	Adj. R-squared:	0.720
Method:	Least Squares	F-statistic:	125.3
Date:	Thu, 09 Sep 2021	Prob (F-statistic):	1.73e-64
Time:	09:16:49	Log-Likelihood:	-698.26
No. Observations:	243	AIC:	1409.
Df Residuals:	237	BIC:	1429.
Df Model:	5		
Covariance Type:	nonrobust		

$$R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2},$$

med

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

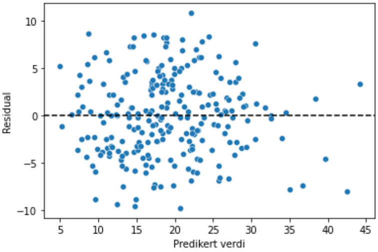
som måler den totale variabiliteten i $(y_1, \dots, y_n)$.

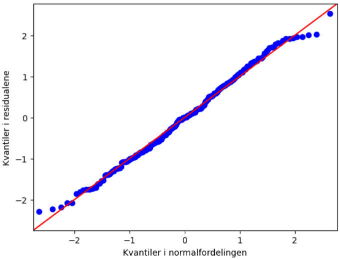
Problemet med R^2: Verdien blir alltid større når flere variabler er lagt til.

For modellvalg bruker vi derfor en justert versjon:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

TA og kvantil-kvantil plot





Binære forklaringsvariabler

Den enkleste modellen er

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

Hva betyr det når x_{1i} er enten 0 eller 1 (binær)?

$$\begin{array}{ll} \beta_0 + e_i & \text{hvis } x_{1i} = 0 \ , \\ \beta_0 + \beta_1 + e_i & \text{hvis } x_{1i} = 1 \ . \end{array}$$

Eksempel: Studie om kvikksølv (Hg)

Modell:

$$\log(Hg_{urin})_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + e_i \ ,$$

Med

- $\log(Hg_{urin})$: log konsentrasjon av Hg i urin.
- x_1 binær variabel som er 1 hvis person røyker, ellers 0
- x_2 antall amalgam fillinger i tennene.
- x_3 antall fiskemåltider per måned.

Interpretasjon av regresjon med binær variabel

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.1136	0.100	-21.101	0.000	-2.311	-1.916
smoking	0.3317	0.257	1.292	0.198	-0.175	0.839
amalgam_quant	0.1799	0.039	4.566	0.000	0.102	0.258
fisk_quant	0.0678	0.017	4.088	0.000	0.035	0.101

Modell for røyker:

Modell for ikke-røyker:

Kategoriske forklaringsvariabler

- Vi gjør ting enda mer fleksible (eller kompliserte!) når vi også tillater kategoriske forklaringsvariabler.
- Eksempel med 3 kategorier: Bildatasett med y=bensinforbruk og forklaringsvariabler `vekt` og `origin` som er en av de tre kategoriene {American,European,Japanese}.

```
formel='mpg ~  vekt + origin'
```

- Idé: dummy-variabel koding – kalles *one-hot koding* i maskinlæring.
  - $x_{2i} = 0$ og $x_{3i} = 0$ hvis `origin` er "American"
  - $x_{2i} = 1$ og $x_{3i} = 0$ hvis `origin` er "European"
  - $x_{2i} = 0$ og $x_{3i} = 1$ hvis `origin` er "Japanese"

Modellen: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$

```
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.700
Method:                 Least Squares   F-statistic:                     304.7
Date:                Mon, 13 Sep 2021   Prob (F-statistic):          1.28e-101
Time:                        14:42:02   Log-Likelihood:                -1123.9
No. Observations:                 392   AIC:                             2256.
Df Residuals:                     388   BIC:                             2272.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept                43.7322      1.113     39.277      0.000      41.543      45.921
origin[T.European]        0.9709      0.659      1.474      0.141      -0.324       2.266
origin[T.Japanese]        2.3271      0.665      3.501      0.001       1.020       3.634
weight                   -0.0070      0.000    -21.956      0.000      -0.008      -0.006
==============================================================================
```

Så hva er modellene for de tre opprinnelsene (`origin`) av bilene?

Videre denne uken

- Se på videoene om multippel lineær regresjon (hvis du ikke har allerede gjort det).
- Se på videoene om klassifikasjon.
- Les i kompendiet.
- Begyn å jobbe med prosjektoppgaven – problem 1.
- Se her for mer informasjon: https://wiki.math.ntnu.no/istx100y/2025h/1003