Fra statistisk modell for enkel lineær regresjon til statistisk modell for multippel lineær regresjon

- Tolkning: En under-populasjon for hver verdi av *x*
- Forventninga er en rettlinja funksjon av *x*
- Data = Forventning angitt av linja + Residual
- Residualene er uavhengige N(0,σ) for alle verdier av *x*

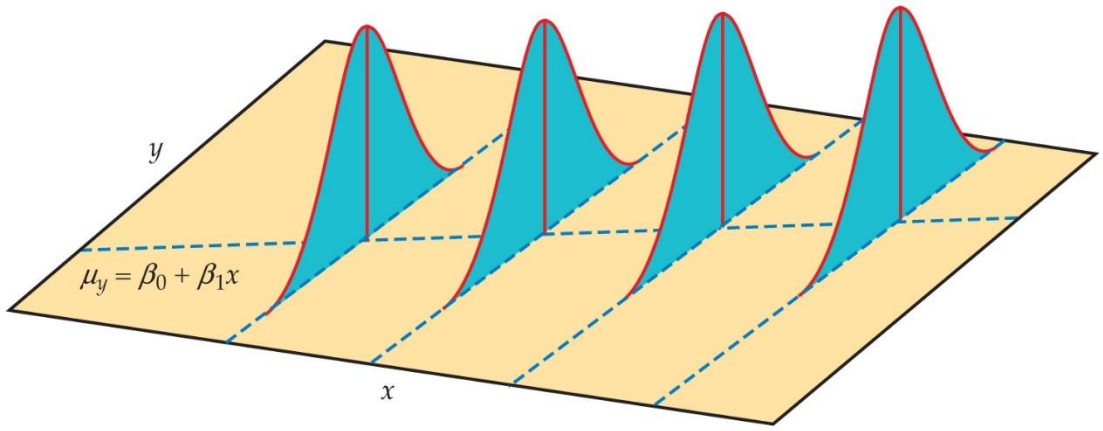


Figure 10.2
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Kap 11: Multippel lineær regresjon

- Data
- Modeller
- Estimere og tolke parametere
- Konfidensintervall og statistisk hypotesetest
- Kvadrert multippel korrelasjon: R^2
- Prediksjon av nye observasjoner

Data for multippel regresjon

Dataene i en enkel lineær regresjon består av *n* observasjoner (x_i, y_i) av to variabler.

Dataene for multippel lineær regresjon består av verdien til en responsvariabel *y*, og *p* forklaringsvariabler $(x_1, x_2, \ldots, x_p)$ for hver av de *n* observasjonene.

Vi skriver dataene slik:

Kalles design-matrise

	Variabler				
Observasjon	x_1	x_2	…	x_p	y
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}	y_n

I multippel lineær regresjon utvider vi perspektivet for til å omfatte flere forklaringsvariabler $x_1, x_2, \ldots, x_p$

Vi utvider modellen på en naturlig, intuitiv måte:
Prediktoren får ett lineært ledd for hver forklaringsvariabel.

Det oppstår en del nye problemstillinger med flere forklaringsvariabler, men vi vil konsentrere oss om det som er felles med enkel lineær regresjon.

Det er fortsatt en responsvariabel *y*.

Multippel lineær regresjonsmodell

- Mens vi i enkel lineær regresjon har en respons y og en forklaringsvariabel x, og forventa respons gitt x som $\mu_y = \beta_0 + \beta_1 x$

- Har vi nå: En respons y, p forklaringsvariable $x_1, x_2, \ldots, x_p$ og forventa respons gitt $x_1, x_2, \ldots, x_p$

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Den statistiske modellen for individuell respons y_i gitt $x_{i1}, x_{i2}, \ldots, x_{ip}$, er

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad \text{for } i=1,\ldots,n$$

- Vi fortsetter med antagelsen om at individuell variasjon ε_i er uavhengig og normalfordelt $N(0, \sigma)$

Antagelsene i multippel lineær regresjon er

- 1. Linearitet: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$
- 2. Uavhengighet: Gitt x-ene, er $y_1, y_2, \ldots, y_n$ uavhengige.
  (tilsvarende for $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$)
- 3. Konstant varians: Gitt x-ene, har $y_1, y_2, \ldots, y_n$ like standardavvik σ.
  (tilsvarende for $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$)
- 4. Normalitet: Gitt x-ene er $y_1, y_2, \ldots, y_n$ normalfordelte
  (tilsvarende for $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$)

Hvorfor multippel regresjon?

- Reelle fenomener er typisk multifaktorielle: responsen y har sammenheng med flere forklaringsvariable $x_1, x_2, \ldots, x_p$.

- Vi kan bruker multippel regresjon til å
  - identifisere viktige forklaringsvariable for en respons.
  - studere effekten av en av en eller flere forklaringsvariabler, justert for de andre forklaringsvariablene
  - predikere responsen y i tilfeller der verdier for forklaringsvariablene $x_1, x_2, \ldots, x_p$ er kjent

Systolisk blodtrykk: data fra 32 voksne menn

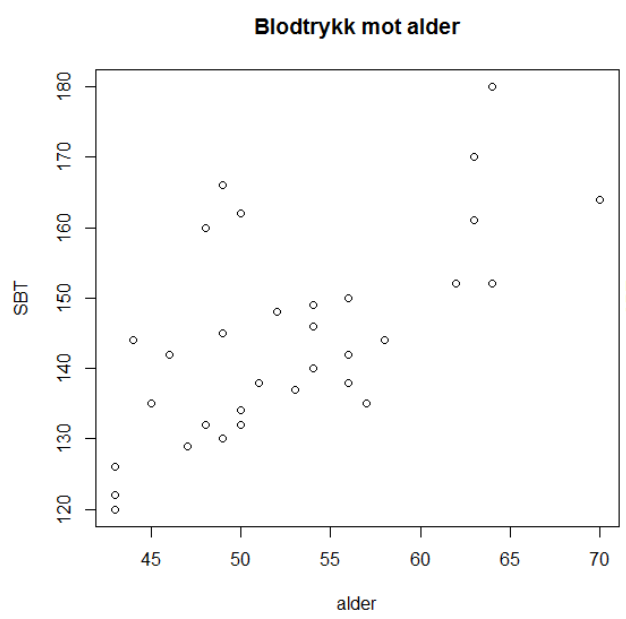
- Variabler:
  - Systolisk blodtrykk (SBT): Målt i mmHg
  - Alder (alder): Målt i år
  - Body Mass Index (BMI): Målt i kg/m^2
  - Røykestatus (røyk): 0=ikke-røyker, 1=røyker
  - (ID (1-32))
- Dvs, for hvert individ i=1,...n:
  - Responsvariabel yi: SBT
  - 3 forklaringsvariable : x_{i1}=alder, x_{i2}=BMI, x_{i3}=røyk
- Vi vil bruke multippel lineær regresjonsmodell for å modellere hvordan blodtrykket varierer med de 3 forklaringsvariablene

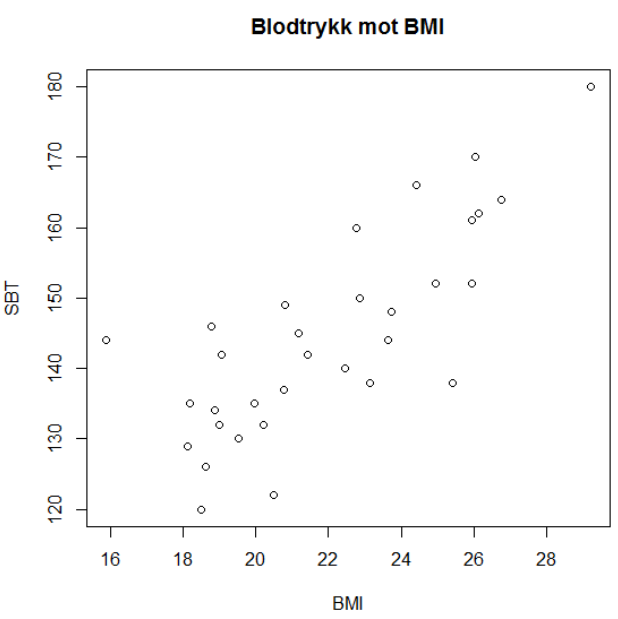
ID	SBT	r?yk	BMI	alder
1	135	0	18.18	45
2	122	0	20.48	43
3	130	0	19.53	49
4	148	0	23.74	52
5	146	1	18.77	54
6	129	1	18.12	47
7	162	1	26.12	50
8	160	1	22.76	48
9	144	1	15.90	44
10	180	1	29.21	64
11	166	1	24.43	49
12	138	1	25.40	51
13	152	0	25.93	64
14	138	0	23.14	56
15	140	1	22.44	54
16	134	1	18.89	50
17	145	1	21.17	49
18	142	1	19.05	46
19	135	0	19.98	57
20	142	0	21.43	56
21	150	1	22.86	56
22	144	0	23.63	58
23	137	0	20.76	53
24	132	0	20.22	50
25	149	1	20.80	54
26	132	1	19.01	48
27	120	0	18.50	43
28	126	1	18.62	43
29	161	0	25.93	63
30	170	1	26.03	63
31	152	0	24.96	62
32	164	0	26.76	70

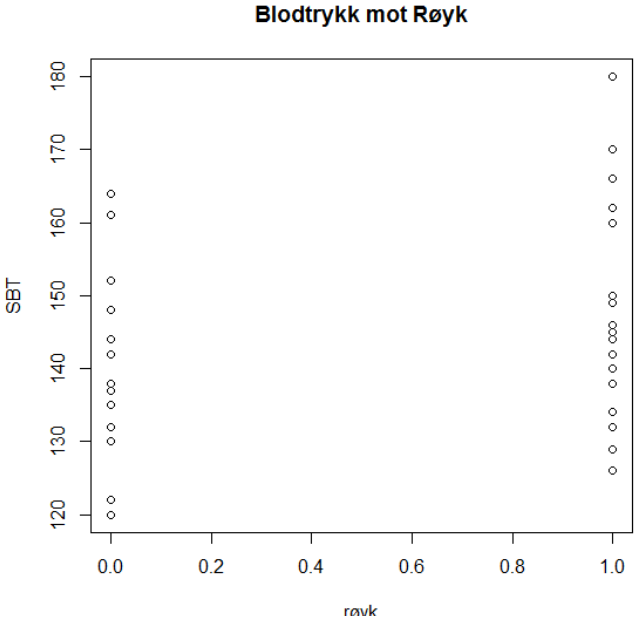
Design-matrisa: alle forklaringsvariablene

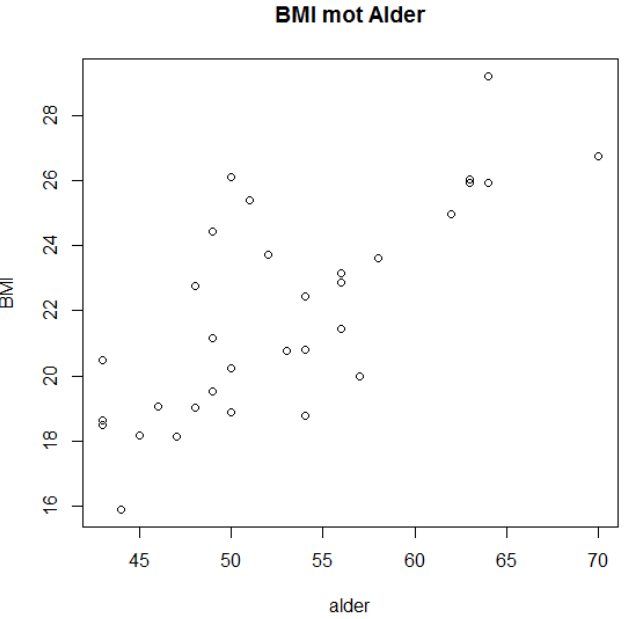
Respons-variabelen

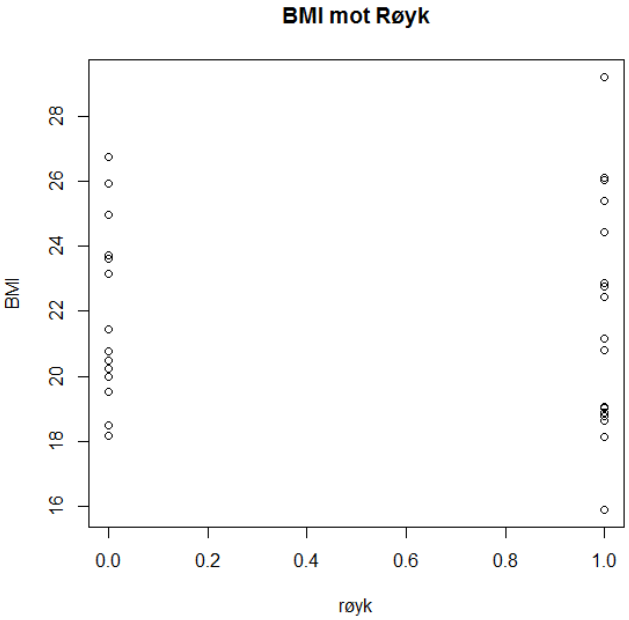
Se på data: Spredningsplott av alle parvise kombinasjoner av variablene

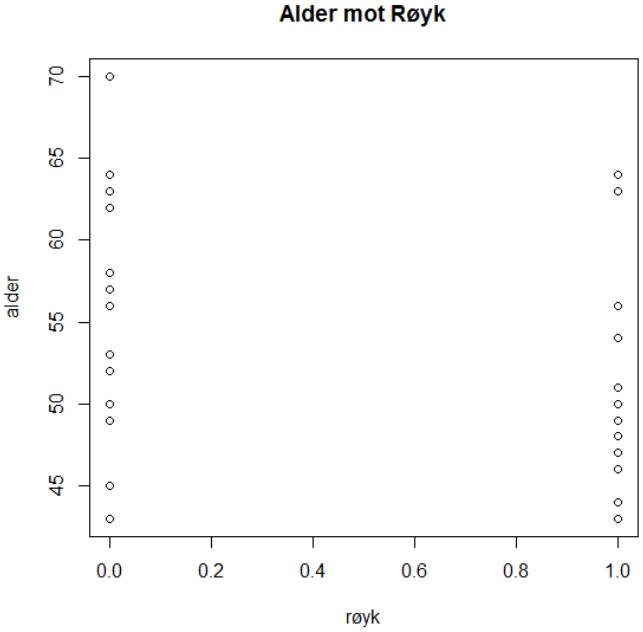












Estimering av regresjonsparametre

- Populasjonsparametre: β_0 , β_1 , …. , β_p og σ
- Data: y_i, x_{i1}, x_{i2}, …. , x_{ip} for i = 1, … , n

- Estimater b_0 , b_1 , …. , b_p for populasjons-regresjonsparameterne β_j finner vi med minste kvadraters metode; dvs vi velger b_0 , b_1 , …. , b_p som minimerer

$$\sum(error)^2 = \sum(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2$$

- I multippel lineær regresjon er formlene for b_0 , b_1 , …. , b_p, litt mer kompliserte enn i enkel lineær regresjon, men det er uansett enkelt å be R eller annen programvare om å beregne dem for oss.

R-utskrift av regresjonsparametre

```
> lmfit <- lm(SBT~BMI+alder+røyk,data=blodtrykkdata)
> summary(lmfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.3831    11.3488   4.263 0.000207 ***
BMI           2.4774     0.6640   3.731 0.000860 ***
alder         0.6831     0.3127   2.185 0.037421 *
røyk         10.6210     2.8588   3.715 0.000897 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.663 on 28 degrees of freedom
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7167
F-statistic: 27.14 on 3 and 28 DF,  p-value: 1.952e-08
```

Å tolke modell-parameterne i multippel lineær regresjon

- Modell: $\mu_y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- Regresjonslinje: $\hat{\mu}_y = b_0 + b_1 x_1 + \cdots + b_p x_p$

- b_0 - estimert forventa respons y når alle forklaringsvariablene $x_1, x_2, \ldots, x_p$ har verdi lik 0
- b_1 - estimert forventa økning i respons *y* når x_1 øker med en enhet samtidig som alle andre forklaringsvariablene har uendra verdi.
  - Tilsvarende for $b_2, \ldots, b_p$

Å estimere forventning og predikert verdi

Når vi har estimert regresjonskoeffisientene, er det naturlig å angi punktestimat for forventa respons gitt forklaringsvariablene $x^*_1 \ldots x^*_p$ som

$$\hat{\mu}_y = b_0 + b_1 x_1^* + \cdots + b_p x_p^*$$

Vi bruker samme punkt-estimat

$$\hat{y} = b_0 + b_1 x_1^* + \cdots + b_p x_p^*$$

når vi skal predikere en ny observasjon $y = \mu_y + \varepsilon = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^* + \varepsilon$
der vi kjenner verdien av forklaringsvariablene $x_1^*, x_2^*, \ldots, x_p^*$

Residualene, og å estimere σ

For observasjon y_i defineres residualen ved

$$e_i = \text{observert verdi - predikert verdi}$$
$$= y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}$$

Ved hjelp av residualene estimerer vi variansen σ^2 ved

$$s^2 = \frac{1}{n-(p+1)} \sum e_i^2 = \frac{1}{n-(p+1)} \sum (y_i - \hat{y}_i)^2$$

Vi har estimert p+1 parametere

Standardavviket σ estimeres da ved

$$s = \sqrt{s^2}$$

Konfidensintervall for β_j

husk at b_j er normalfordelt

- Når vi skal gjøre statistisk inferens for $\beta_0, \ldots, \beta_j, \ldots, \beta_p$ bruker vi estimatene $b_0, \ldots, b_j, \ldots, b_p$, og estimatet s:
- $\frac{b_j - \beta_j}{SE_{bj}}$ vil være t-fordelt med n-p-1 frihetsgrader.
- Standardfeilene SE_{bj} avhenger av s, og beregnes av programvare.

Et **konfidensintervall for β_j på nivå C** blir da:

$$b_j \pm t^* SE_{b_j}$$

hvor SE_{b_j} er standardfeilen til b_j og t^* er den kritiske verdien for $t(n-p-1)$-fordelingen, som gir areal C mellom $-t^*$ og t^*.

Hvilke av forklaringsvariablene er viktige?

- Hypotesetesting for regresjons-koeffisientene er essensielt
- Må teste null-hypotesen H_0: $\beta_j = 0$,

  som sier at den j-te forklaringsvariabelen ikke er viktig for å forklare responsen i modellen slik den er angitt
- Baserer testen på t-observator

$$t_j = \frac{b_j}{SE_{b_j}}$$

Finnes av programvare, basert på s

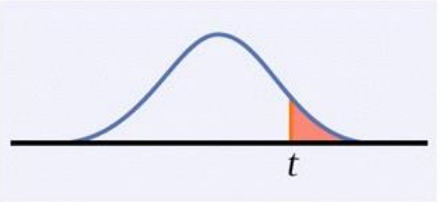
- Disse er alle t-fordelte med n-p-1 frihetsgrader når respektive nullhypoteser gjelder.

Hypotesetesting forts.

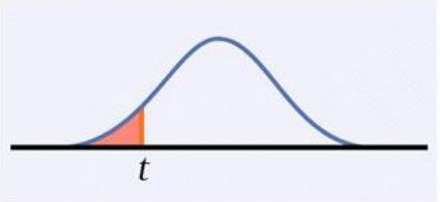
Når vi skal teste hypotesen H_0: $\beta_j = 0$ mot en tosidig (evt. ensidig) alternativhypotese, beregner vi t-observatoren $t_j = \frac{b_j}{SE_{b_j}}$

Denne har t-fordeling med n-p-1 frihetsgrader når H_0 er sann.

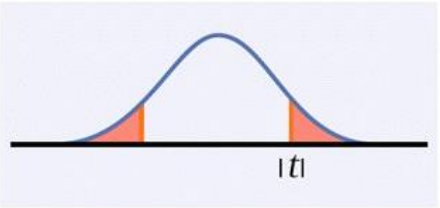
H_a: $\beta_j > 0$ is $P(T \geq t)$



H_a: $\beta_j < 0$ is $P(T \leq t)$



H_a: $\beta_j \neq 0$ is $2P(T \geq |t|)$



R og annen programvare bruker alltid tosidig alternativ

Signifikanstesten for β_j er for β_j alene

Anta at vi tester H_0: $\beta_j = 0$ for hver j, og finner at ingen av de p testene er signifikante.

Betyr dette at vi skal konkludere med at ingen av forklaringsvariablene har noen sammenheng med responsen?

På ingen måte!

Når vi ikke forkaster H_0: $\beta_j = 0$, betyr dette at vi antageligvis ikke trenger x_j i modellen når alle de andre variablene er til stede.

Manglende forkastning av hver av hypotesene for β_j betyr bare at det er trygt å ta bort minst én av variablene, men sier ikke nødvendigvis noe om hvilken eller hvilke, eller hvor mange.

Man kan gjøre videre analyser, for å undersøke hvilken delmengde av variablene som gir den beste modellen.

Hypotesetesting av koeffisientene ved bruk av R-utskriften for multippel lineær regresjon

```
> summary(lmfit)$coef

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.3831    11.3488   4.263 0.000207 ***
BMI           2.4774     0.6640   3.731 0.000860 ***
alder         0.6831     0.3127   2.185 0.037421 *
røyk         10.6210     2.8588   3.715 0.000897 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Oppsummerer blodtrykk-eksempelet

- Alle regresjons-koeffisientene β_j er signifikant ulik 0 på et 5% signifikansnivå (tosida tester)
- Det er en signifikant lineær sammenheng mellom systolisk blodtrykk og hver av de tre forklarings-variablene alder, BMI og røykestatus - når alle tre forklaringsvariablene er med i modellen
- Alle tre variablene gir forventa økt systolisk blodtrykk (siden de estimerte stigningstallene er positive)

- Men før vi gir en endelig konklusjon må vi gjøre residualanalyse for å sjekke at modellen passer til dataene

Vi må gjøre modellsjekk som en del av analysen for multippel lineær regresjon

Vi har antatt

1. Linearitet
2. Uavhengighet gitt x-ene
3. Konstant varians gitt x-ene
4. Normalitet gitt x-ene

Vi skal først se på hvordan antagelsene sjekkes visuelt.

Deretter skal vi diskutere konsekvenser og mulige forbedringer.

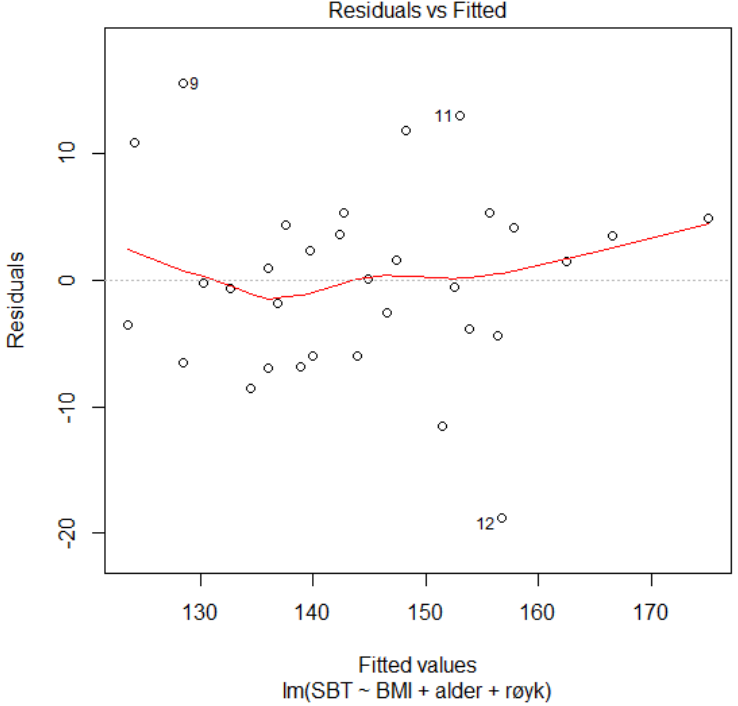
Å sjekke Linearitetsantagelsen $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

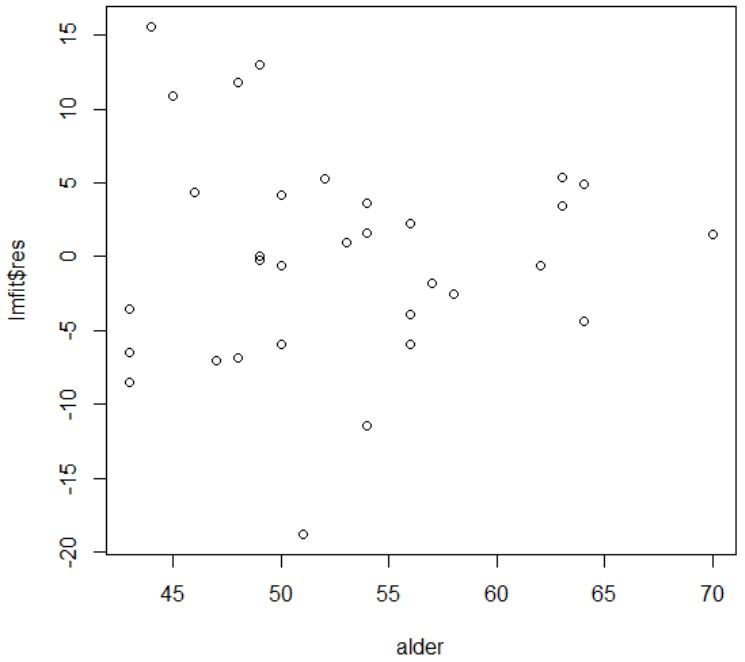
De to vanligste plottene for å sjekke om lineariteten holder er

- Residualer mot predikerte verdier
- Residualene mot hver av forklaringsvariablene

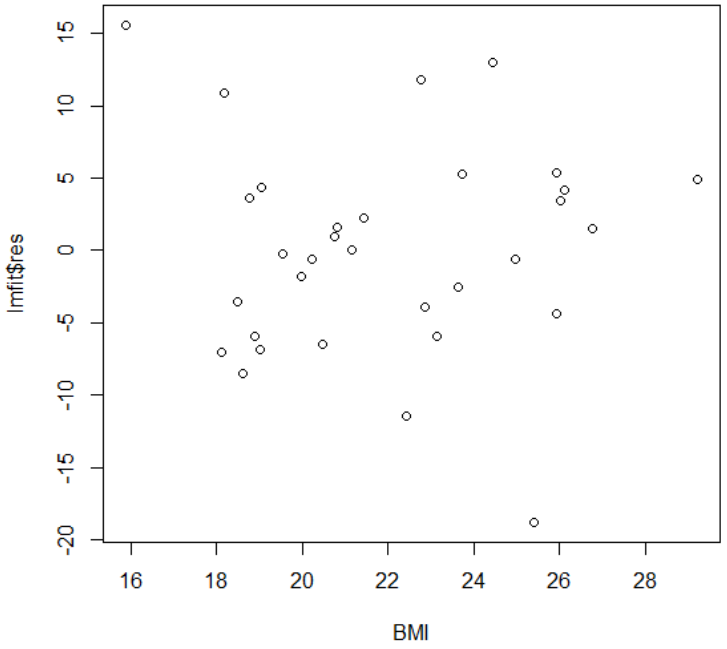
Merk: For enkel lineær regresjon er det samme *form* på disse to fordi predikert verdi er en lineærtransformasjon av forklaringsvariabelen.

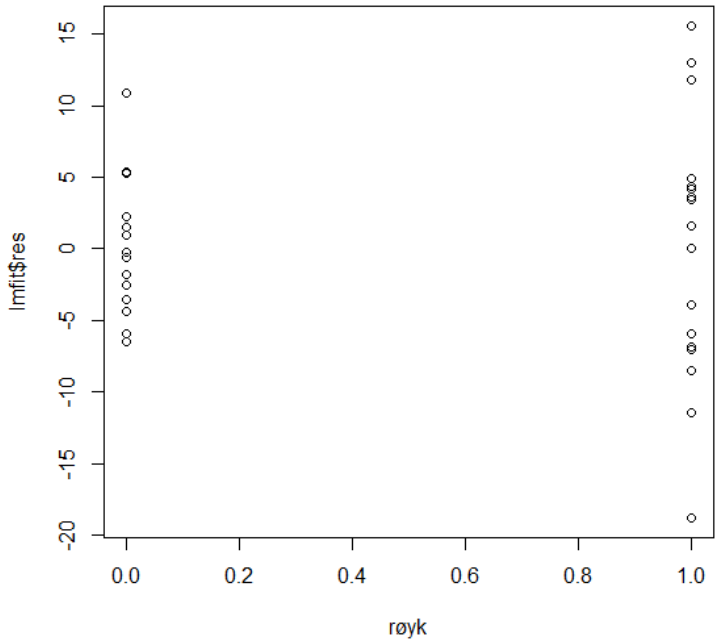
Et mønster i residualene antyder avvik fra linearitetsantagelsen.
Et mulig mønster kan være først positive residualer, deretter negative og så til slutt positive residualer igjen.





Her ser vi ikke tydelige avvik fra linearitetsantagelsen





Å sjekke uavhengighets-antagelsen:

Dersom dataene er hentet inn i en rekkefølge, er det naturlig å plotte residualer mot observasjonsnummer for å se etter mønster.

Mønstre indikerer avhengighet.

Også andre typer avhengigheter som innen familier, skoleklasser og andre grupper, kan oppdages ved å inspisere residualene for de ulike gruppene.
Det er for eksempel mistenkelig dersom (nesten) alle residualene innenfor en klasse systematisk er positive.

Men: i mange situasjoner er det ikke mulig å sjekke dette, f.eks. ved blodtrykksdataene.

Sjekk av antagelsen om konstant varians:

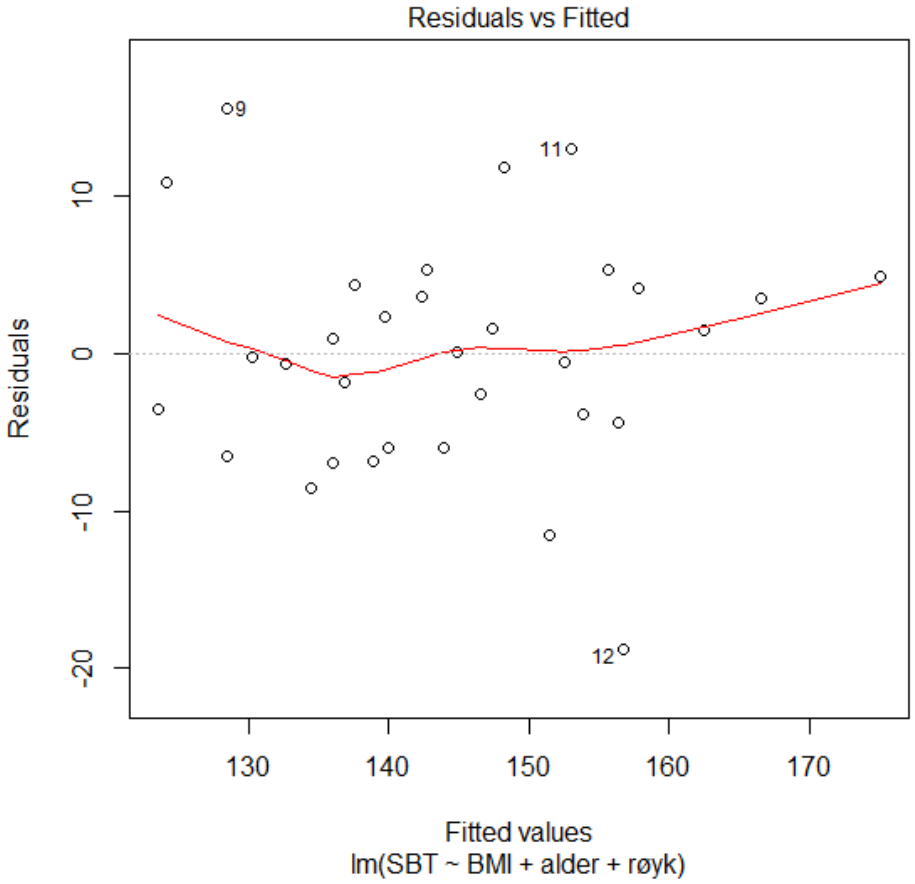
Se etter mønster i plottene av

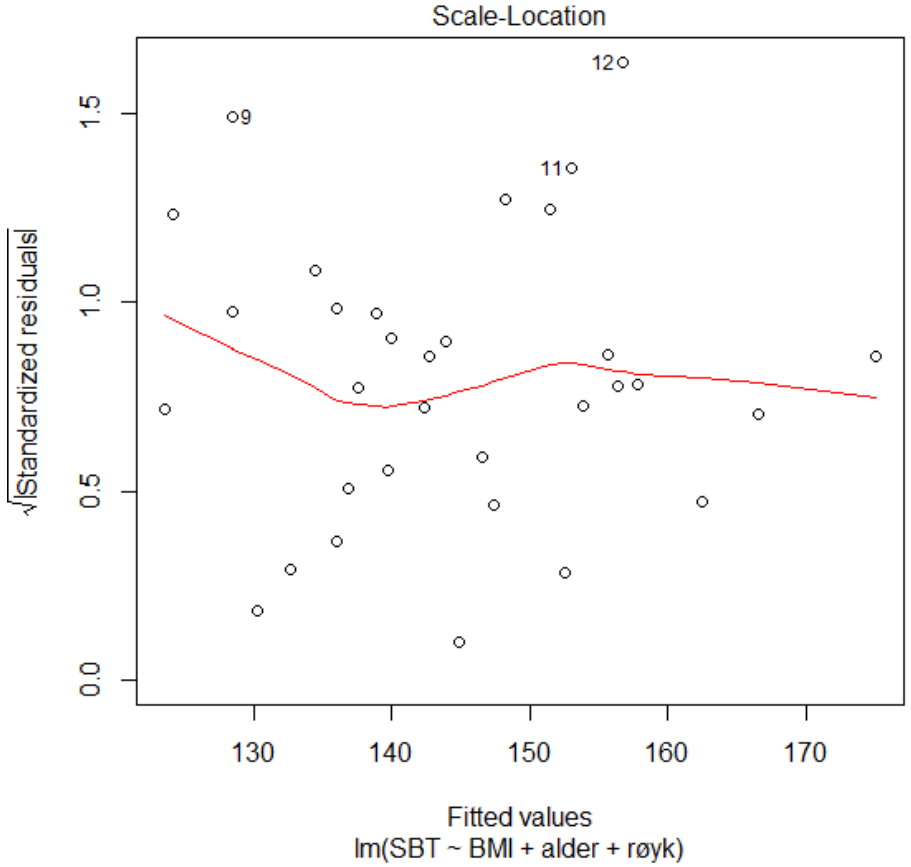
- residualer mot predikerte verdier

- absoluttverdi av residualene mot predikerte verdier

Et mulig mønster er en ”vifteform” i residualene med stadig større spredning.
Oppdager du et slikt mønster, har du grunn til å tro at antagelsen ikke holder.

Antagelsen om konstant varians ser OK ut.



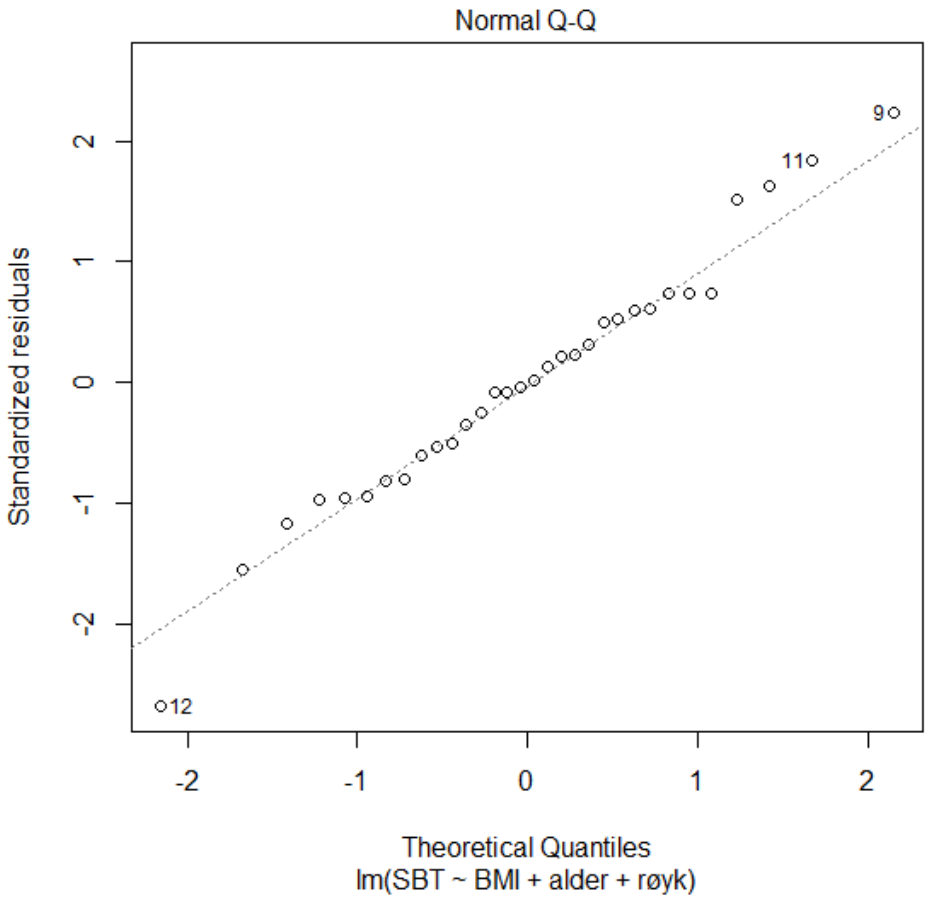


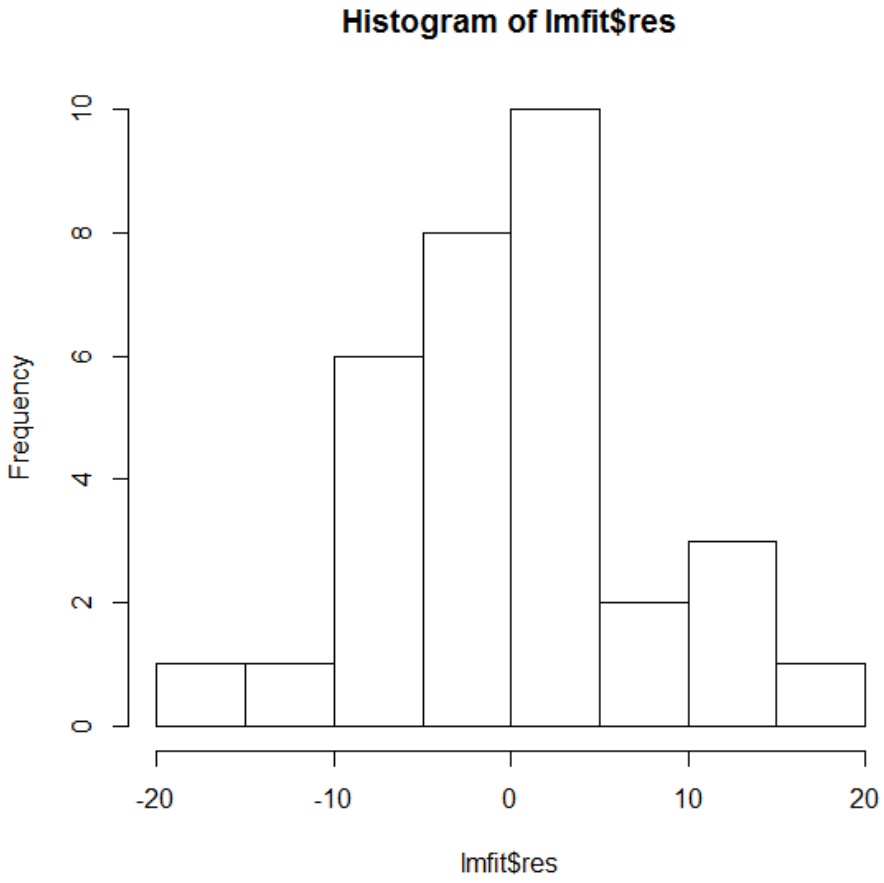
Sjekk av normalitetsantagelsen:

- Normal kvantilplott (qq-plott) av residualene
- Histogram over residualene

Disse plottene skal helst indikere normalfordeling.

Vi ser hverken klare avvik fra normalitetsantagelsen eller opplagte uteliggere.





Konsekvenser av avvik fra antagelsene i multippel lineær regresjon

- Avvik fra linearitet: Modellen er gal, og hvis avvikene er markante *må* en mer passende modell estimeres.

- Avhengighet mellom observasjonene kan gi dårlig estimat for standardfeil.
  - Dette kan medføre betydelige feil i statistisk inferens (KI og tester). Men punkt-estimatene er likevel OK.

- Avvik fra konstant varians: Standardfeil kan bli gale.
  - Dette kan medføre betydelige feil i statistisk inferens (KI og tester). Men punkt-estimatene er likevel OK.

- Avvik fra normalitet: Vi har ikke lenger at testobservatorer etc. er t-fordelt. Ikke kritisk hvis utvalgstørrelsen (n) er stor og dataene ikke har betydelige uteliggere

Mulige forbedringer av lineær regresjonsmodell

Dersom linearitetsantagelsen $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ svikter kan man prøve bl.a.

- Transformer x-er (f.eks. log-trans.).
- Innfør et nytt ledd $\beta_{p+1} x_j^2$ i tillegg til $\beta_j x_j$ (polynomiell regresjon).
- Avansert: glattingsteknikker
- Transformer y-ene

Andre forbedringsmuligheter for å oppfylle antakelser:

- Uavhengighet: Hvis grupper av avhengige enheter er store, kan man innføre en gruppe-variabel
- Konstant varians: Transformér responsvariabelen y
- Normalitet: Sjekk uteliggere

Forklart andel av varians: R^2

Ved enkel lineær regresjon hadde vi at forklart andel av variasjon

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

faktisk var lik kvadratet av korrelasjonskoeffisienten ($R^2 = r^2$).

For multippel lineær regresjon vil den generelle definisjonen

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

fortsatt fortelle hvor mye av variasjonen i de opprinnelige dataene som kan forklares ved regresjonen.

Kvadratrota $R = \sqrt{R^2}$ kalles den multiple korrelasjonskoeffisient, men kan ikke relateres til en forklaringsvariabel alene.

R^2 er kvadratet av korrelasjonskoeffisienten mellom observasjonene y_i og prediksjonene $\hat{y}_i$.

- $0 \leq R^2 \leq 1$
- R^2 kan ikke avta når vi inkluderer en ny forklaringsvariabel i modellen, men vil som oftest øke

Siden R^2 vil øke med antall forklaringsvariable også når disse har helt marginal betydning, vil den overestimere betydningen av alle forklaringsvariablene. Derfor oppgis også andre varianter av dette målet i statistikkpakker, bl.a.: justert (adjusted) R^2.

Brukes til å sammenligne modeller, velge kovariater

```
> summary(lmfit)

Call:
lm(formula = SBT ~ BMI + alder + røyk, data = blodtrykkdata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7704  -4.7471  -0.0826   4.1910  15.5471

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.3831    11.3488   4.263 0.000207 ***
BMI           2.4774     0.6640   3.731 0.000860 ***
alder         0.6831     0.3127   2.185 0.037421 *
røyk         10.6210     2.8588   3.715 0.000897 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.663 on 28 degrees of freedom
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7167
F-statistic: 27.14 on 3 and 28 DF,  p-value: 1.952e-08
```

Vi husker: Konfidensintervall for forventa respons i enkel lineær regresjon

Med en spesifisert verdi av forklaringsvariablen x=x* blir forventa respons:

$$\mu_y = \beta_0 + \beta_1 x^*$$

Den naturlige estimatoren for μ_y er $\hat{\mu}_y = b_0 + b_1 x^*$

Denne $\hat{\mu}_y$ er en tilfeldig variabel med forventning μ_y, og en litt komplisert formel for standardavviket. Spesielt er $\hat{\mu}_y$ normalfordelt dersom observasjonene y_i er det.

Fordi σ estimeres med s, blir et nivå C konfidensintervall for μ_y:

$$\hat{\mu}_y \pm t^* \, SE_{\hat{\mu}_y}$$

når t* velges slik at P(-t* < t(n-2) < t*) = C og $SE_{\hat{\mu}_y}$ er standardfeilen til $\hat{\mu}_y$.

Vi husker: Prediksjonsintervall for ny observasjon i enkel lineær regresjon

Verdien av den nye observasjonen $y = \mu_y + \varepsilon = \beta_0 + \beta_1 x^* + \varepsilon$ estimeres med $\hat{y} = b_0 + b_1 x^*$, altså samme verdi som estimert forventning.

Men for å ta høyde for dens usikkerhet må vi ta hensyn til variasjonen i feilleddet ε som har standardavvik σ. Variansen til prediksjonen $\hat{y}$ blir dermed $SE^2_{\hat{\mu}_y} + \sigma^2$

Et prediksjonsintervall med nivå C for en ny verdi av y når x=x* gis dermed ved

$$\hat{y} \pm t^* \sqrt{SE^2_{\hat{\mu}_y} + s^2}$$

når P(-t* < t(n-2) < t*) = C .
Når vi ser på den nye observasjonen som en tilfeldig variabel, er P(ny verdi faller innenfor prediksjonsintervallet)=C.

Estimering av forventa respons og prediksjon av ny verdi *y*:

Akkurat som ved enkel lineær regresjon vil vi være interessert i å estimere, gitt forklaringsvariable $x^*_1, x^*_2, \ldots, x^*_p$,

- forventninga til y gitt verdi for hver av forklaringsvariablene

$$\mu_y = \beta_0 + \beta_1 x^*_1 + \beta_2 x^*_2 + \cdots + \beta_p x^*_p$$

- verdien av ny *y* gitt verdi for hver av forklaringsvariablene

$$y = \mu_y + \varepsilon = \beta_0 + \beta_1 x^*_1 + \beta_2 x^*_2 + \cdots + \beta_p x^*_p + \varepsilon$$

Her er ε individuell variasjon, som antas å ha forventning 0 og standardavvik σ, og antas uavhengig og normalfordelt.

For begge størrelsene benytter vi samme punktestimat

$$\hat{\mu}_y = b_0 + b_1 x^*_1 + b_2 x^*_2 + \cdots + b_p x^*_p$$

men feilmarginen blir forskjellig!

Usikkerhet i forventa respons μ_y og i predikert ny verdi *y*

Standardfeilen $SE_{\hat{\mu}_y}$ til $\hat{\mu}_y = b_0 + b_1 x_1^* + b_2 x_2^* + \cdots + b_p x_p^*$ avhenger av usikkerheten (varianser og kovarianser) til minste kvadraters estimatorene $b_0, b_1, b_2, \cdots, b_p$ samt av verdiene av forklarings-variablene $x^*_1, x^*_2, \ldots, x^*_p$.

Et 95% konfidensintervall for $\hat{\mu}_y$ blir nå gitt ved $\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$

der t^* er 97.5 persentilen i t-fordelinga med n-p-1 frihetsgrader.

Tilsvarende blir til et 95% prediksjonsintervall for ny *y* gitt ved

$$\hat{\mu}_y \pm t^* SE_{\hat{y}}$$

der $SE^2_{\hat{y}} = s^2 + SE^2_{\hat{\mu}_y}$ er estimert varians for en ny observasjon *y*.

Konfidens- og prediksjonsintervall i R

```
> nydata=data.frame(BMI=28,røyk=1,alder=50)

> predict(lmfit,nydata,interval="confidence")
       fit       lwr       upr
1 162.5285 152.6616 172.3954

> predict(lmfit,nydata,interval="predict")
       fit       lwr       upr
1 162.5285 143.9873 181.0698
```

Oppsummering

- Modell
- Estimering
- Fortolkning av estimater
- Konfidensintervall & testing av koeffisienter
- Residualer og modellsjekk
- Forklart andel av variasjon R^2
- Prediksjon for ny y gitt x-verdier

Neste uke: (Multippel) logistisk regresjon (kap. 14) OG Sannsynlighetsmaksimeringsprinsippet ♫♫♩

Kvadratsumsoppspaltning

Ikke pensum, men inkludert for de interesserte

Som indikert i forbindelse med R^2 med har vi følgende sammenheng

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Dvs. Total kvadratsum SST = $\sum (y_i - \bar{y})^2$ kan uttrykkes som

SST = SSE + SSM

der SSE = $\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ er kvadratsum for avviket fra modellen og SSM = $\sum (\hat{y}_i - \bar{y})^2$ er kvadratsum forklart av modellen.

For SSE svarer det df_E = n-p-1 frihetsgrader og tilsvarende har modellen p parametre og vi bruker df_M = p frihetsgrader .

F-observator

Ikke pensum, men inkludert for de interesserte

Som dere husker estimeres variansen σ^2 med

$$s^2 = \frac{1}{n-p-1}\sum (y_i - \hat{y}_i)^2 = \frac{SSE}{dfE} = MSE$$

Dersom det ikke er noen sammenheng mellom forklaringsvariable og responser, dvs. når alle regresjonsparametrene $\beta_1 = \beta_2 = \ldots. = \beta_p = 0$ så er faktisk også

$$MSM = \frac{SSM}{dfM} = \frac{1}{p}\sum (\hat{y}_i - \bar{y})^2$$

et estimat for σ^2 . Hvis minst en $\beta_j \neq 0$ så vil MSM tendere til å være større enn σ^2

Dette kan vi bruke til å test om minst én $\beta_j \neq 0$ ved testobservator F = MSM / MSE som er ”F-fordelt med p og n-p-1 frihetsgrader”.

F-observator i R

Ikke pensum, men inkludert for de interesserte

```
> summary(lmfit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.3831    11.3488   4.263 0.000207 ***
BMI           2.4774     0.6640   3.731 0.000860 ***
alder         0.6831     0.3127   2.185 0.037421 *
røyk         10.6210     2.8588   3.715 0.000897 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.663 on 28 degrees of freedom
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7167
F-statistic: 27.14 on 3 and 28 DF,  p-value: 1.952e-08
```

Store verdier av F taler altså mot H_0: $\beta_1 = \cdots = \beta_p = 0$.
Her finner vi en stor F = 27.14 - Og vi så jo tidligere at alle 3 koeffisienter var signifikant forskjellig fra 0

Ikke pensum, men inkludert for de interesserte

F-observator i enkel lineær regresjon

I enkel lineær regresjon $\mu_y = \beta_0 + \beta_1 x$ er det bare en regresjons-parameter β_1 som angir sammenheng mellom respons og forklaringsvariabel.

Den vanlige testen for H_0: $\beta_1 = 0$ basert på $t = \frac{b_1}{SE_{b_1}}$

Denne samsvare perfekt med F-testen.
Med bare en regresjonsparameter har vi faktisk at $F = t^2$
og begge testene får samme p-verdi (tosida test for β_1).

Ikke pensum, men inkludert for de interesserte

Vi ser $F=t^2$ i R-utskrift for eksempeldata og enkel lineær regresjon

```
> summary(lm(BMI~PA)) #Skritt-teller-eksempelet!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
PA           -0.6547     0.1583  -4.135  7.5e-05 ***
---

Residual standard error: 3.655 on 98 degrees of freedom
Multiple R-squared: 0.1485,  Adjusted R-squared:  0.1399
F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

Så vi får $t^2 = (-4.135)^2 = 17.1 = F$

Også p-verdiene for t og F er like.