Kort overblikk over kurset sålangt

- Kapittel 1: Deskriptiv statististikk for en variabel
- Kapittel 2: Deskriptiv statistikk for samvariasjon mellom to variable (regresjon)
- Kapittel 3: Metoder for å innhente data
- Kapittel 4: Sannsynlighet og sannsynlighetsmodeller for en variabel
- Kapittel 5: Sannsynlighetsfordelingen til et gjennomsnitt
- Kapittel 6: Statistisk inferens for en variabel
- Kapittel 7: Statistisk inferens for sammenligning av to grupper

Kapittel 10: Inferens i regresjon

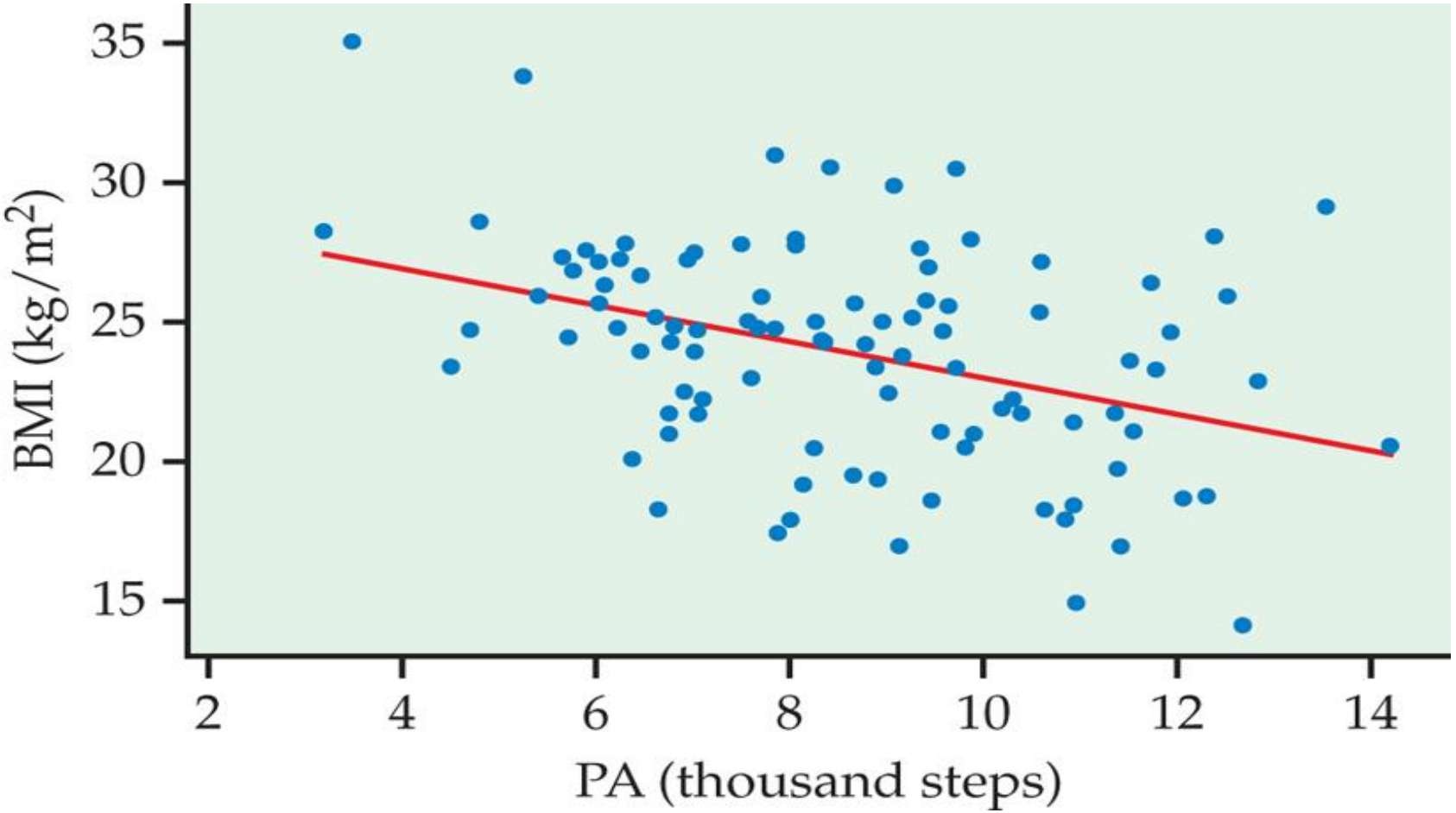
I Kapittel 1, 4, 5 og 6 så vi på en variabel i en populasjon.
Først med deskriptiv statistikk, deretter med sannsynlighetsmodeller og tilslutt inferensmetoder for forventning som konfidensintervall og hypotesetesting.

Her skal vi først repetere den deskriptive statistikken for samvariasjon mellom to variable, spesielt med lineær regresjon.
Deretter setter vi opp samvariasjonen som en sannsynlighetsmodell og dermed utvikles metoder for statistisk inferens når det er en responsvariabel og en forklaringsvariabel.

Eksempel BMI (body mass index) og fysisk aktivitet, n=100

Respons: BMI

Forklaringsvariabel: Fysisk aktivitet målt ved skritteller i en uke.



De deskriptive teknikkene og
numeriske mål vi har sett vil vise seg nyttige:

- Spredningsplott
- Tilpasning av regresjonslinje med minste kvadraters metode
- Korrelasjon.

Husk hvordan gjennomsnitt / forventning er blitt behandlet:

- Data $x_1, x_2, \ldots, x_n$. Gjennomsnittet $\bar{x} = \frac{1}{n}\sum x_i$ er senter i fordeling
- Data er realisasjoner av tilfeldig variable X_i med forventning μ
- Gjennomsnittet $\bar{x}$ har (tilnærmet) fordeling $N(\mu\ ,\ \sigma/\sqrt{n})$
- Et 95% konfidensintervall for μ blir gitt ved $\bar{x} \pm t^* s/\sqrt{n}$

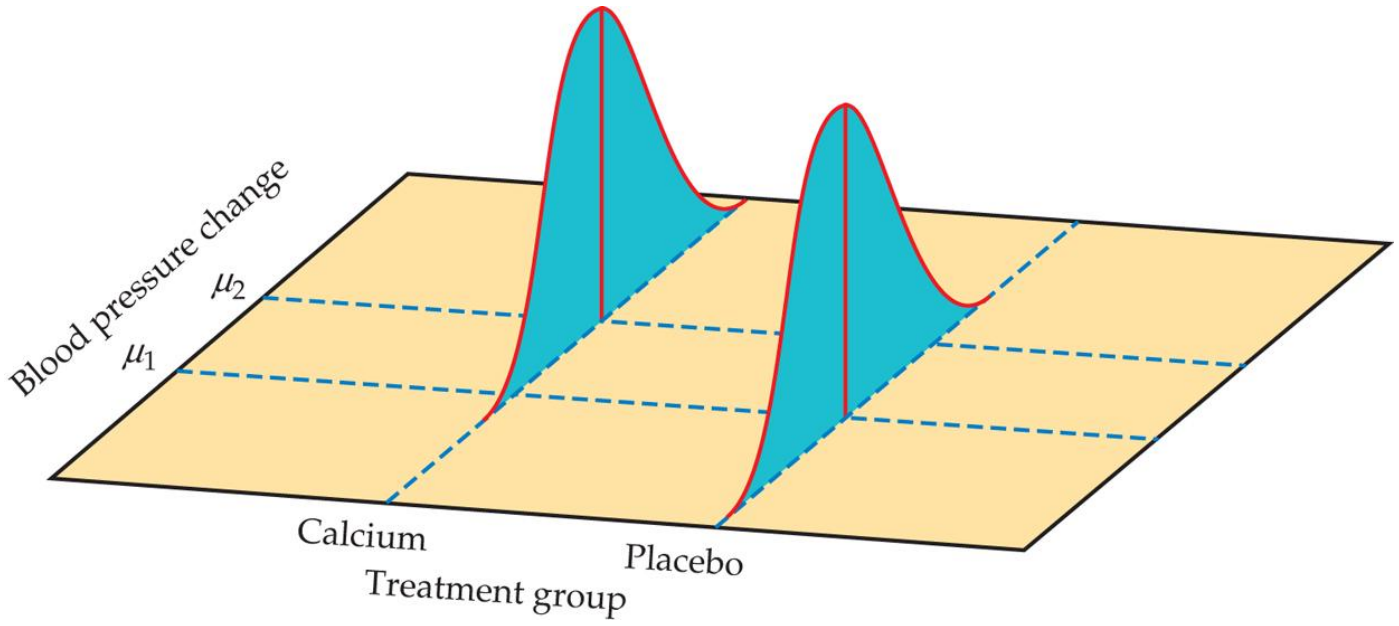
  og man kan tilsvarende teste hypoteser om μ.

Vi skal følge et tilsvarende oppsett for lineær regresjon

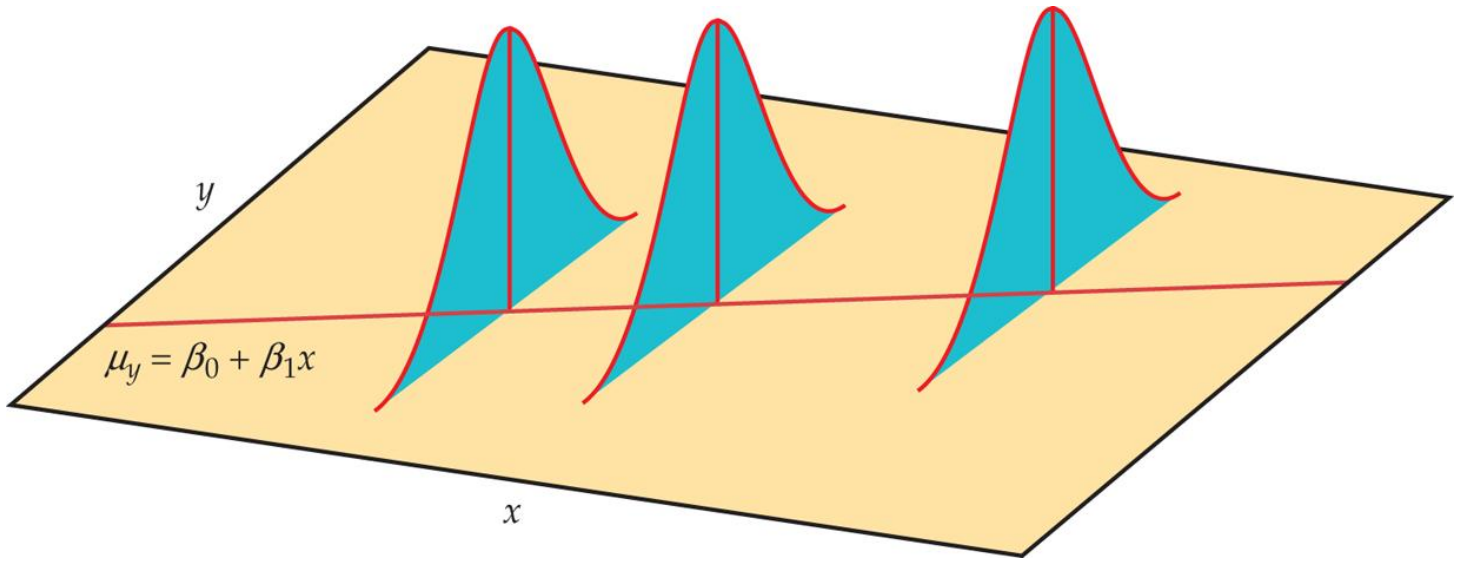
- Vi har data $(x_1, y_1), \ldots, (x_n, y_n)$
- Vi postulere en lineær sammenheng mellom x-er og y-er og estimerer en regresjonslinje $\hat{y} = b_0 + b_1 x$ med minste kvadraters metode for å beskrive denne sammenhengen.
- Vi antar at responsene y_i er realisasjoner av tilfeldige variable Y_i med forventning $\mu_y = \beta_0 + \beta_1 \mathrm{x}$ gitt $x_i = x$
- De estimerte b_0 og b_1 blir tilfeldige variable med forventninger β_0 og β_1 samt med standardfeil som kan beregnes.
- Basert på dette kan vi teste hypoteser om og lage konfidensintervall
. for β_0 og β_1 og dessuten for verdien av regresjonslinja $\mu_y = \beta_0 + \beta_1 \mathrm{x}$.

Enkel lineær regresjon dreier seg om forholdet mellom en responsvariabel y og en forklaringsvariabel x . Her er det viktig å legge merke til at **fordelingen til responsvariabelen avhenger av verdien til forklaringsvariabelen**.

Dette generaliserer situasjonen vi har sett tidligere med to utvalg. Der er det to populasjoner.


Blood pressure change
μ_2
μ_1
Calcium
Placebo
Treatment group

I lineær regresjon kan forklaringsvariabelen ha mange verdier. For hver fast, fiksert verdi defineres en **subpopulasjon**, som består av alle enhetene som har denne spesielle verdien.



Legg merke til at fordelingen for responsene y_i avhenger av forklaringsvariablen x_i gjennom spesifikasjonen av forventningen

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

Dette er populasjonsmodellen for enkel lineær regresjon - som behandles i Kapittel 10.

I Kapittel 11 skal vi utvide dette til å se flere forklaringsvariable samtidig og se på en multippel lineær regresjonsmodell

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

Responsverdiene y som observeres for en bestemt verdi av forklaringsverdien x vil variere rundt verdien av linja for denne verdien, dvs rundt $\beta_0 + \beta_1 x$

Spredningen rundt denne verdien antas å være den samme, og gitt ved standardavviket σ, for alle x .

Modellen har altså tre parametre, β_0, β_1 og σ.

I tillegg vil vi anta at variasjonen rundt $\beta_0 + \beta_1 x$ er normalfordelt, dvs. y er $N(\beta_0 + \beta_1 x, \sigma)$.

Dette er det samme som at vi kan skrive $y = \beta_0 + \beta_1 x + \varepsilon$
der feilleddene (støyen/varisjonen) ε er $N(0, \sigma)$

Vi vil nå bruke det apparatet vi har sett i aksjon i andre sammenhenger til inferens, spesielt konfidensintervaller og hypotesetester, om

- stigningsforholdet β_1 og skjæringspunktet med y-aksen β_0 for regresjonslinja

- forventet respons $\mu_y = \beta_0 + \beta_1 x$ for en gitt verdi av forklaringsvaribelen x

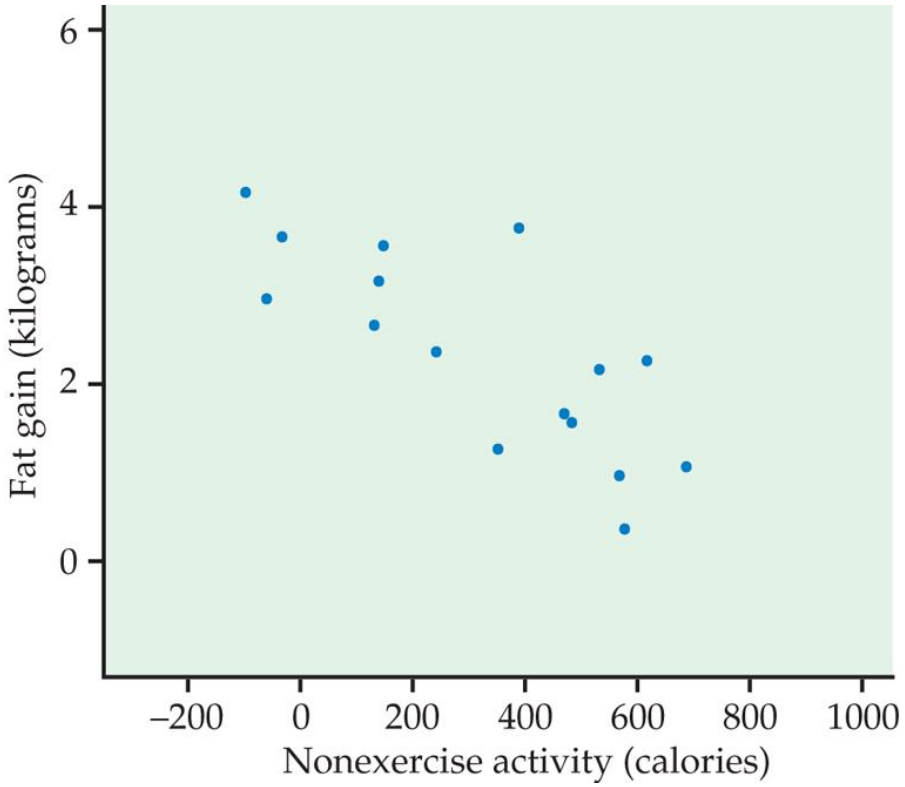
- for en ny responsverdi $\hat{y}$ for en gitt verdi av forklaringsvaribelen x

Men først må parametrene β_0 , β_1 og σ estimeres.

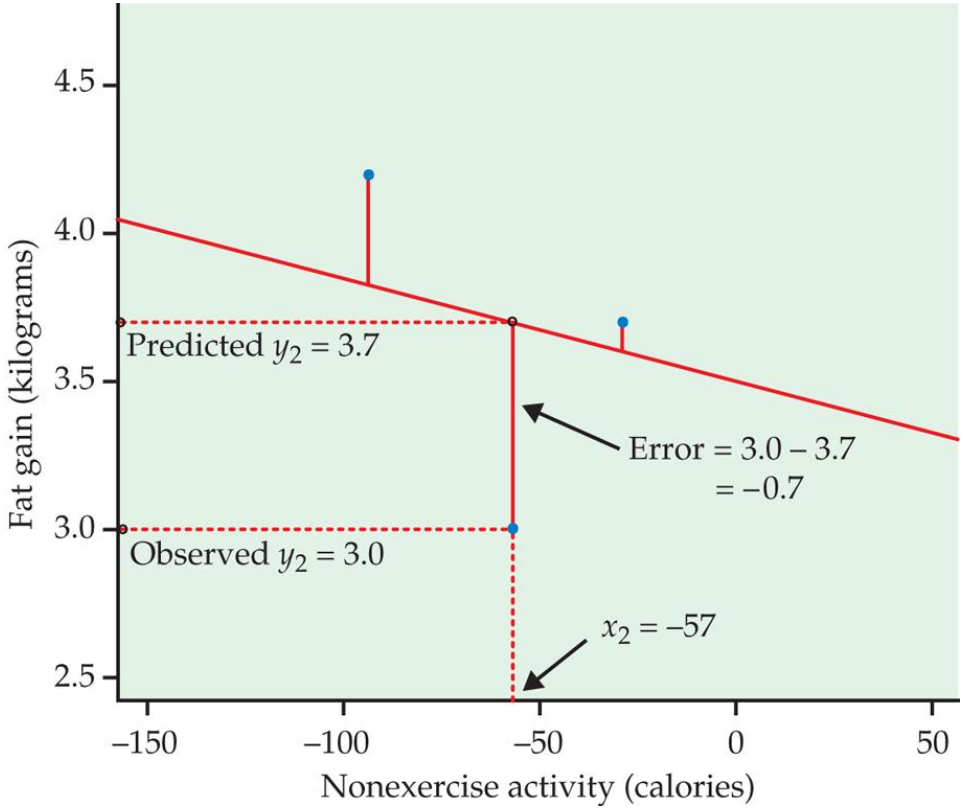
For å estimere β_0 og β_1 brukes koeffisientene b_0 og b_1 fra minste kvadraters metode.

Hovedtrekkene i utledningen av minste kvadraters metode repeteres på følgende slider.

Eksempel Spredningsplottet nedenfor viser sammenheng mellom *endring* i vekt i kg og *endring* i aktivitet som ikke er knyttet til trening, AIT målt i kalorier, i en studie av 16 unge voksne. Rastsløse personer har typisk høy AIT.



Kriterium: Minste kvadraters regresjon av y på x er den linja som minimerer summen av kvadratetene av den vertikale avstanden fra datapunktene $(x_1, y_1),\ldots,(x_n, y_n)$ til linja.



Anta at $\mu = \beta_0 + \beta_1 x$ er regresjonslinja. Med denne får vi prediksjoner $\hat{y}_i = b_0 + b_1 x_i$ når forklaringsvariablen er lik x_i.

Da er $(x_i, \hat{y}_i)$ det punktet på regresjonslinja som ligger rett over eller under (x_i, y_i).

Avstanden mellom de to punkten er

$$| y_i - \hat{y}_i | = | y_i - b_0 - b_1 x_i |$$

og den kvadrerte avstanden blir

$$(y_i - \hat{y}_i)^2 = (y_i - b_0 - b_1 x_i)^2$$

slik at summen av de kvadrerte avstandene blir

$$\sum (y_i - b_0 - b_1 x_i)^2$$

Skjæringspunktet b_0 og stigningsforholdet b_1 til regresjonslinja er de verdiene som minimerer denne kvadratsummen. Alle statistikkpakker har rutiner for å beregne disse koeffesientene numerisk.

Men de kan også uttrykkes eksplisitt og den hjelper på forståelsen.

Formlene for minste kvadraters estimatorene for b_0 og b_1 er gitt ved

$$b_1 = r\,\frac{s_y}{s_x} \quad \text{og} \quad b_0 = \bar{y} - b_1\,\bar{x}$$

Her er som før

$\bar{x}$ gjennomsnittet av $x_1,\ldots,x_n$

$\bar{y}$ gjennomsnittet av $y_1,\ldots,y_n$

s_x standardavviket for $x_1,\ldots,x_n$

s_y standardavviket for $y_1,\ldots,y_n$

r korrelasjonen for $(x_1,y_1),\ldots,(x_n,y_n)$

Merk at b_0 og b_1 avhenger av observasjonene $y_1,\cdots,y_n$ som er tilfeldige variable (samt av $x_1,\cdots,x_n$).

Da er også b_0 og b_1 tilfeldige variable!

Det kan vises at b_0 og b_1 er forventningsrette estimatorer for hhv. β_0 og β_1 :

$$\mu_{b_0} = \beta_0$$

og

$$\mu_{b_1} = \beta_1$$

Dessuten er både b_0 og b_1 lineærkombinasjoner av $y_1, y_2, \cdots, y_n$ som vi har antatt er normalfordelte.

Det følger at da er også b_0 og b_1 er normalfordelte.

Med en utvidet versjon av sentralgrenseteoremet kan vi også vise at b_0 og b_1 er tilnærmet normalfordelt selv om $y_1, y_2, \cdots, y_n$ og $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ ikke selv er normalfordelte.

Et annet resultat som kan vises er at variansen til estimatoren b_1 for stigningsforholdet β_1 er lik

$$\sigma_{b_1}^2 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)\, s_x^2}$$

der s_x^2 er den empiriske variansen til x-ene.

Dette betyr at variansen til estimatoren b_1 typisk blir mindre når n vokser og at b_1 vil ligge nær β_1 når n er stor.

Tilsvarende resultater holder for estimatoren b_0 for konstantleddet β_0.

Størrelsen $\hat{y}_i = b_0 + b_1 x_i$ er den predikerte verdi av responsen y_i med forklaringsvariabel x_i.

Da er $\hat{y}_i$ en forventningsrett estimator for $\mu_y = \beta_0 + \beta_1 x_i$

Residualene er definert ved

$$e_i = \text{observert verdi - predikert verdi}$$
$$= y_i - \hat{y}_i$$

Residualene e_i er de empiriske motstykkene til feilleddene ε_i.

Blant annet har vi $\sum e_i = 0$ svarende til at forventning til ε_i er lik $\mu_\varepsilon = 0$.

Men feilleddene ε_i kan ikke observeres, så f.eks. modellsjekk må gjøres med residualene e_i.

Det gjenstår å estimere standardavviket $\sigma_\epsilon = \sigma$.

Estimatet for variansen σ^2i enkel lineær regresjon er «gjennomsnittet» av de kvadrerte residualene

$$s^2 = \frac{\sum \epsilon_i^2}{n-2} = \frac{\sum (y_i - \widehat{y_i})^2}{n-2}.$$

Legg merke til at man deler med n-2. Det gjør at s^2 blir en forveningsrett estimator for σ^2. Tidligere brukte vi av samme grunn n-1 for å estimere variansen i et utvalg.

Estimatet for standardavviket $\sigma_\epsilon = \sigma$ er $s=\sqrt{s^2}$.

Eksempel: BMI og fysisk aktivitet.

Det skal ikke stor verdi av n til før det lønner seg å beregne verdien av estimatorene, estimatene, i en enkel lineær regresjon ved hjelp av en statistikkpakke.

På de neste slidene følger utskrift fra noen vanlige programpakker:

Regresjonslinja er BMI = 29.5578 – 0.655 PA

Men først strukturen for data som skal benyttes i regresjon:

Det er hensiktsmessig å organisere observasjonenen i en matrise, en såkalt **datamatrise**, som kan beskrives ved

Enhet	Respons	Forklaringsvariabel
1	y_1	x_1
2	y_2	x_2
	…	
n	y_n	x_n

Eksempel: BMI og fysisk aktivitet
y: BMI
x: gjennomsnittlig antall skritt pr dag

SPSS

Model Summary

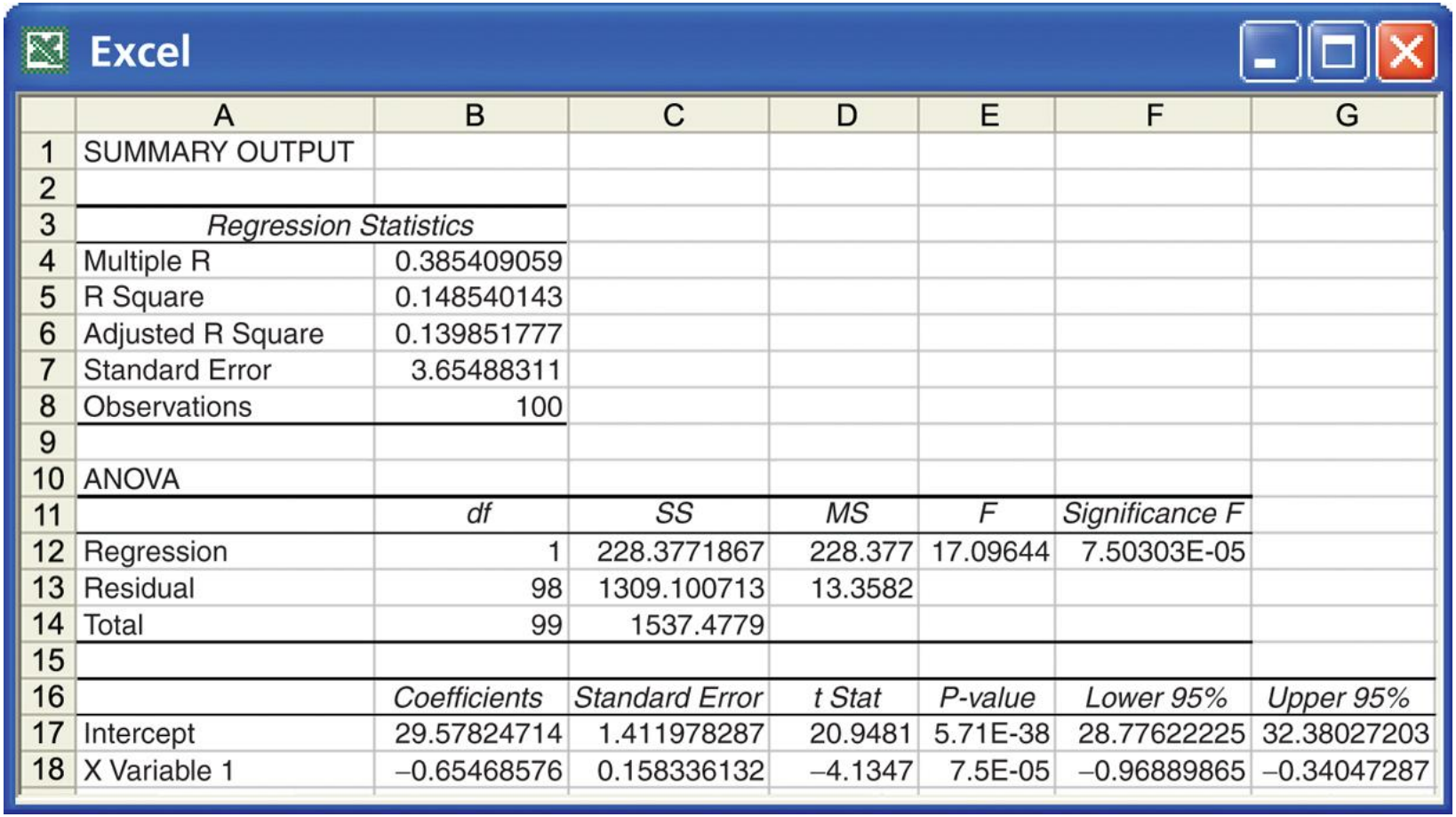
Model	R	R Square	Std. Error of the Estimate
1	.385[a]	.149	3.6549

Model		Unstandardized Coefficients		t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	29.578	1.412	20.948	.000	26.776	32.380
	PAT	−.655	.158	−4.135	.000	−.969	−.340

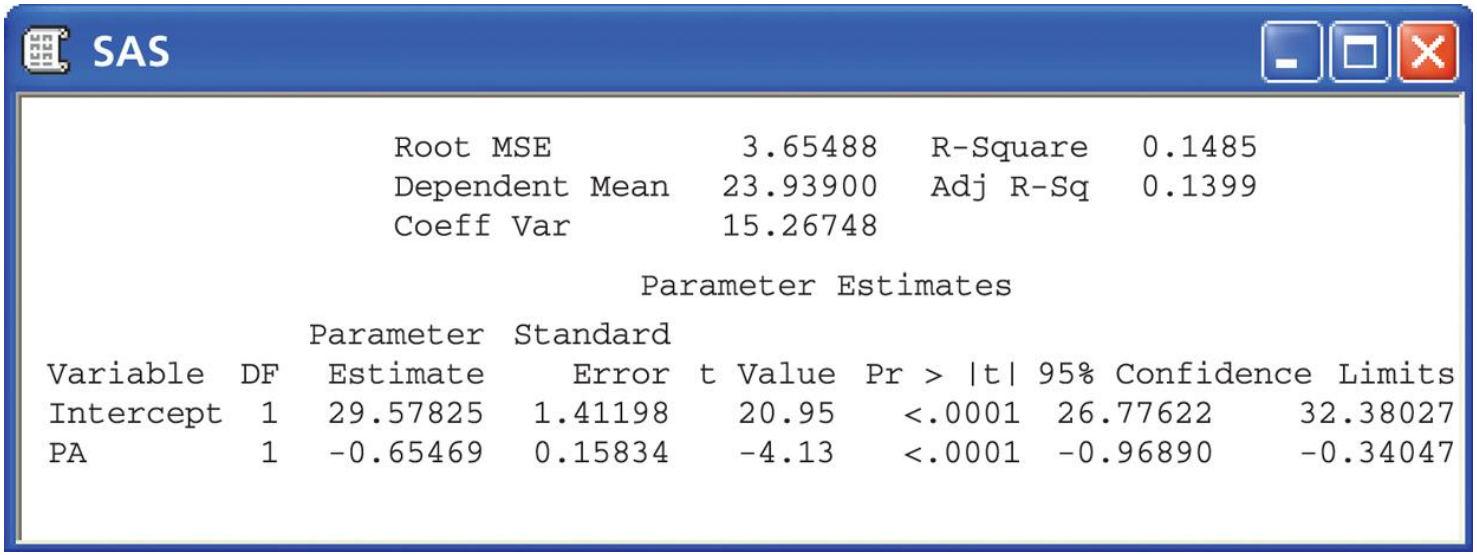
Minitab

```
The regression equation is
BMI = 29.6 - 0.655 PA

Predictor         Coef      SE Coef           T         P
Constant        29.578        1.412       20.95     0.000
PA             -0.6547       0.1583       -4.13     0.000
```



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.385409059					
5	R Square	0.148540143					
6	Adjusted R Square	0.139851777					
7	Standard Error	3.65488311					
8	Observations	100					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	228.3771867	228.377	17.09644	7.50303E-05	
13	Residual	98	1309.100713	13.3582			
14	Total	99	1537.4779				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	29.57824714	1.411978287	20.9481	5.71E-38	28.77622225	32.38027203
18	X Variable 1	–0.65468576	0.158336132	–4.1347	7.5E-05	–0.96889865	–0.34047287


SAS
Root MSE 3.65488 R-Square 0.1485
Dependent Mean 23.93900 Adj R-Sq 0.1399
Coeff Var 15.26748
Parameter Estimates
Variable DF Parameter Estimate Standard Error t Value Pr > |t| 95% Confidence Limits
Intercept 1 29.57825 1.41198 20.95 <.0001 26.77622 32.38027
PA 1 -0.65469 0.15834 -4.13 <.0001 -0.96890 -0.34047

Etter å ha tilpasset modellen og estimert regresjonslinja kan man beregne predikerte verdier og residualer.

Hvis PA er 8000 skritt per dag er predikert BMI

$$29.5578 - 0.655 \times 8 = 24.338.$$

Hvis observert BMI er 25.655, er residualet

$$y - \hat{y} = 25.655 - 24.338 = 1.317.$$

En viktig del av modeltilpasning er å undersøke hvorvidt modelforutsetningene holder.

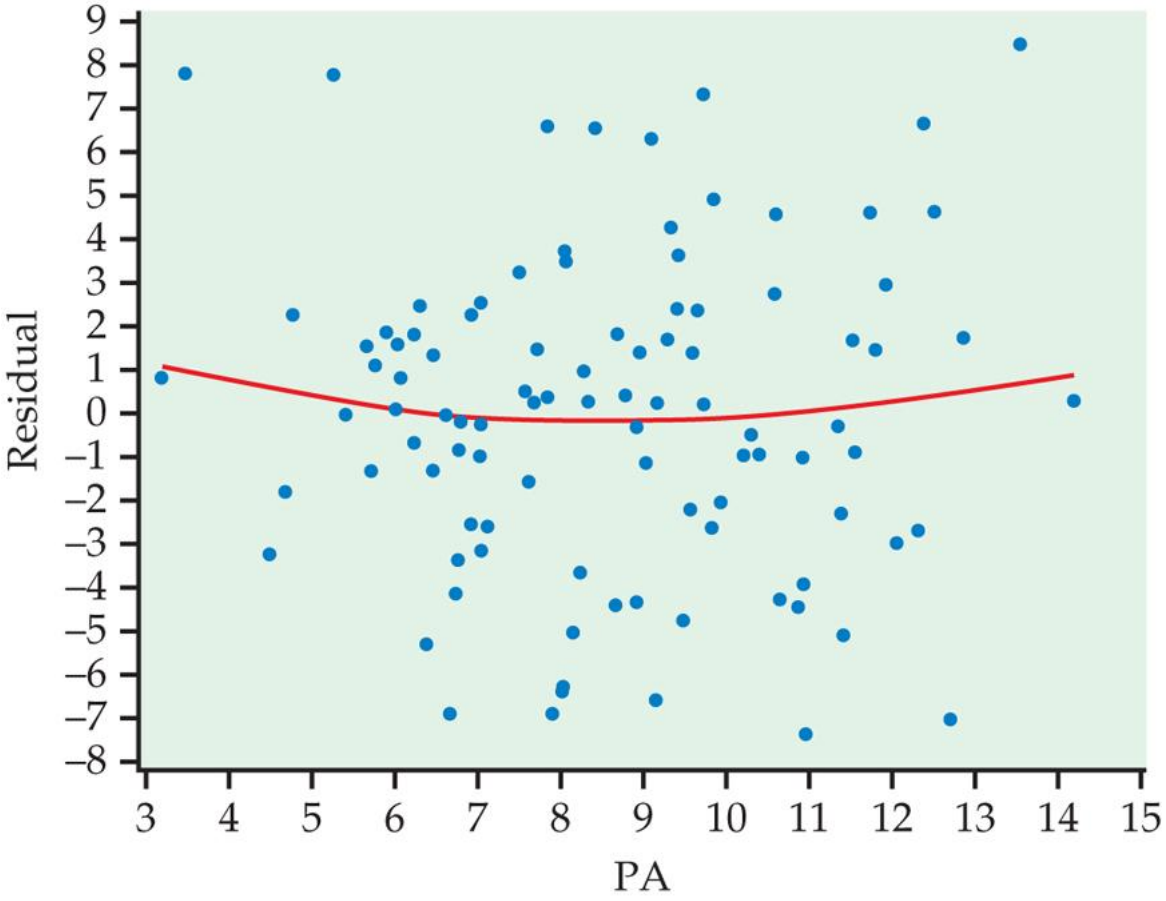
Det kan omfatte residualsjekk for å avdekke

- normalitet
- konstant varians
- andre mønstre i data, som ikke-linearitet og outliere.
- avhengige data, f.eks. ved at dataene er samlet inn i en «rekkefølge»

Første steg er å plotte residualene mot forklaringsvariablene og eventuelt «rekkefølgen» på observasjonene.

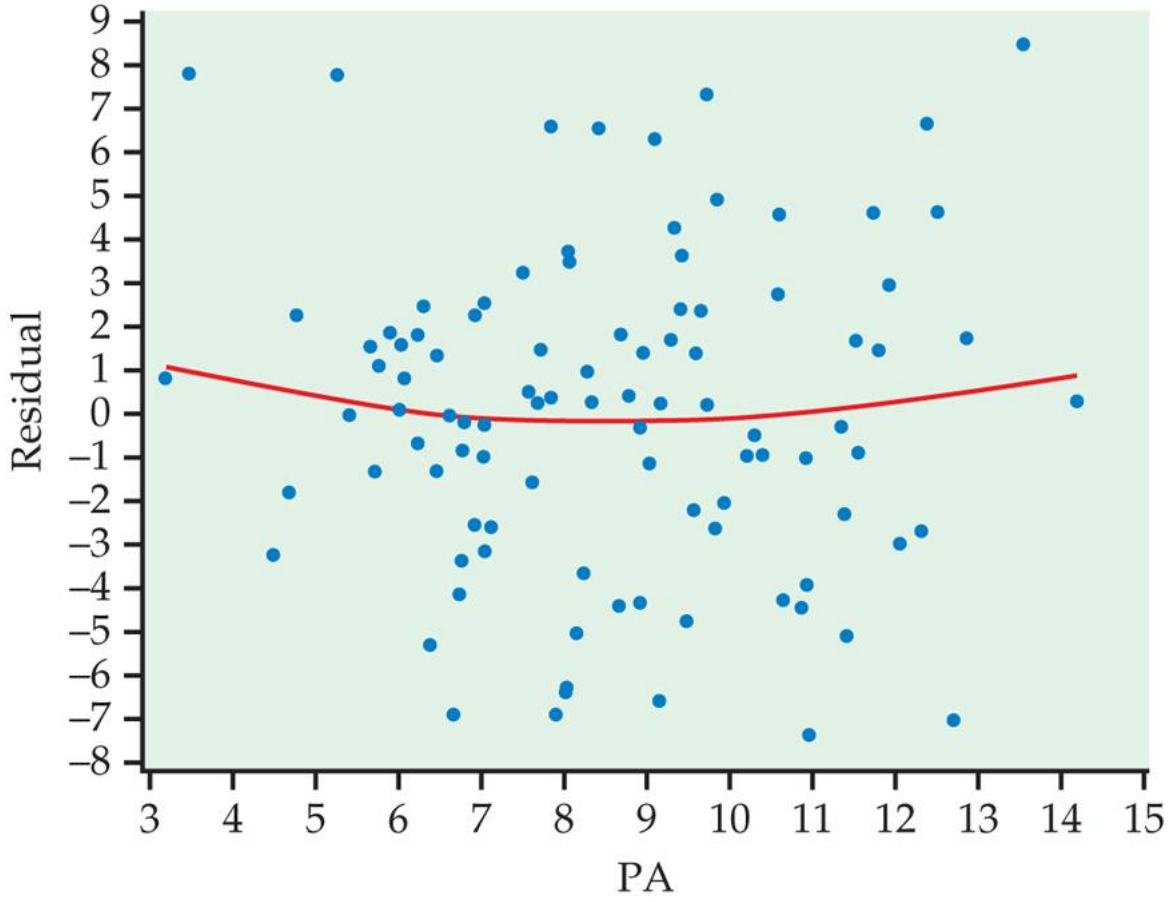
Eksempel: BMI og fysisk aktivitet.

Plott av residual mot forklaringsvariabelen PA



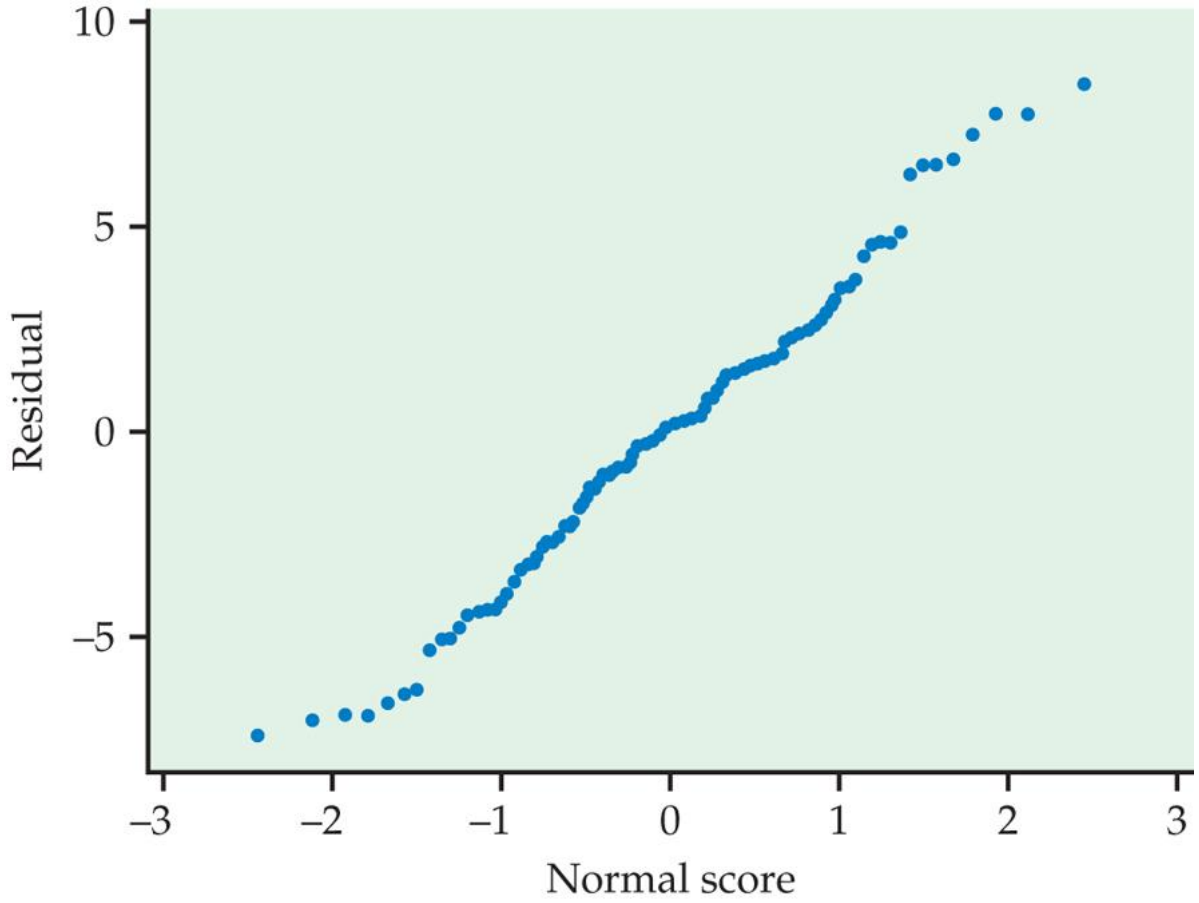
Hvis man ser klare mønstre i dette residualplottet så tyder dette på ikke-linearitet. I dette eksempelet er avviket ikke markant.

Eksempel: BMI og fysisk aktivitet.



Det samme plottet kan også avsløre om variansen er konstant, noe som svikter hvis det er markant større spredning ved store eller små verdier av x-variablen. Her: Ingen grunn til å tro at variansen ikke er konstant.

For å sjekke om feilleddene ε_i er normalfordelte kan vi lage et normalfordelingsplott over residualene e_i.



Alternativt kan vi se på et histogram over e_i-ene.

Men: (Små) avvik fra normalfordeling er ikke kritisk!

Konfidensintervaller for β_0 og β_1 utledes i prinsippet på samme måte som konfidensintervaller for forventning i et utvalg og konfidensintervaller for differansen mellom forventningene i to utvalg. Formlene blir litt mer kompliserte i dette tilfellet.

Men alle konfidensintervaller har formen

$$\text{estimat} \pm \text{t}^* \, SE_{estimat}$$

der t* er kritisk verdi i en t-fordeling.

Et konfidensintervall med konfidenskoeffisient lik C for β_0 har grenser

$$b_0 \pm \text{t}^* \, SE_{b_0}$$

Et konfidensintervall med konfidenskoeffisient lik C for β_0 har grenser

$$b_1 \pm \text{t}^* \, SE_{b_1}$$

Her er t* den verdien som gjør at arealet under tetthetskurven til en t(n-2) fordeling mellom -t* og t* er C .

For å teste nullhypotesen $H_0: \beta_1 = 0$ benyttes t-statistikken

$$t = \frac{b_1}{SE_{b_1}}$$

Når $H_0: \beta_1 = 0$ holder vil denne t-statistikken være trukket fra en T-fordeling med n-2 frihetsgrader.

La i det følgende T være en tilfeldig variabel fra denne fordelingen.

P-verdien for testen vil avhenge av hvilket alternativ vi bruker. Standard er å rapportere P-verdier fra tosidige tester.

P-verdien for $H_0: \beta_1 = 0$ blir

med ensidig alternativ hypotese $H_a: \beta_1 < 0$:

$$P(T < t)$$

med ensidig alternativ hypotese $H_a: \beta_1 > 0$:

$$P(T > t)$$

med tosidig alternativ hypotese $H_a: \beta_1 \neq 0$:

$$2\, P(T > |t|)$$

Tilsvarende finner vi P-verdier for testing av H_0: $\beta_0 = 0$.

Utskrifter fra programpakker inneholder vanligvis slike P-verdier,

men legg merke til at de egentlig ikke er spesielt interessante.
Nullhypotesen H_0: $\beta_0 = 0$ forteller oss bare hva μ_y er når x=0.

Nullhypotesen H_0: $\beta_1 = 0$ er vesentlig mer interessant!

For hvis $\beta_1 \neq 0$ så vil forventningen $\mu_y = \beta_0 + \beta_1 x$ avhenge av x.

Tilsvarende hvis $\beta_1 = 0$ så er vi tilbake til en
ett-utvalgssituasjon $\mu_y = \beta_0$ for alle x.

Eksempel: BMI og fysisk aktivitet.

Utskriften fra statistikkpakkene inneholder typisk noe av typen:

Forklaringsvaribel	Koeffisient	St. feil	T	P
konstant	29.578	1.412	20.95	0.000
PA	-0.6547	0.1583	-4.13	0.000

Sjekker at det stemmer:

t= $\frac{b_1}{SE_{b_1}} = \frac{-0.6547}{0.1583}$ = -4.1358

For H_0:β_1 =0 og tosidig alternativ H_a:β_1 ≠0 er
P-verdien 2P(T≥ |t|)= 2P(T ≥ 4.14) =0.000 for en t-fordeling med n-2= 100-2= 98 frihetsgrader.

Vi vil kanskje i utgangspunktet tro at $\beta_1 < 0$ og det kan derfor være naturlig med ensidig alternativ hypotese $H_a : \beta_1 < 0$.

Det gir en p-verdi som er halvparten så stor, noe som i dette tilfellet kan regnes ut til å være 0.0000368.

Dette tallet kan vi ikke finne eksakt i Tabell D.
Det nærmeste vi finner er at

med 80 frihetsgrader er $P(T \leq -3.300) = 0.0005$
og med 100 frihetsgrader blir $P(T \leq -3.390) = 0.0005$.

Så vi innser at P-verdien må være veldig liten.

95% Konfidensintervall for β_1:

Trenger kritisk verdi for T-fordelingen med n-2=98 frihetsgrader, men i Tabell D finner vi bare 80 og 100 frihetsgrader.

Vi bruker da 80 frihetsgrader siden dette er konservativt, dvs. dette gir et litt for bredt intervall.

97.5 persentilen med 80 frihetsgrader er $t^* = 1.99$. Dermed blir intervallet

$$b_1 \pm 1.99\, SE_{b_1} = \text{-}0.655 \pm 1.99 \text{ x } 0.1583 = -0.655 \pm 0.315$$

$$= (-0.970\ ,\ -0.340)$$

97.5 persentilen med 98 frihetsgrader er $t^* = 1.984$. Det endrer intervallet til $(-0.969\ ,\ -0.340)$

som er eksakt det som SAS og SPSS oppga.

Vi skal nå se på to utvidelser:

- Konfidensintervall for forventet respons μ_y når forklaringsvariabelen har en bestemt verdi x*

- «Konfidensintervall» for responsen til en ny observasjon når forklaringsvariabelen har en bestemt verdi x*

  Denne typen intervall kalles "prediksjonsintervall".

Forventet respons μ_y når forklaringsvariabelen har en bestemt verdi x* er $\mu_y = \beta_0 + \beta_1$ x*. Den estimeres med

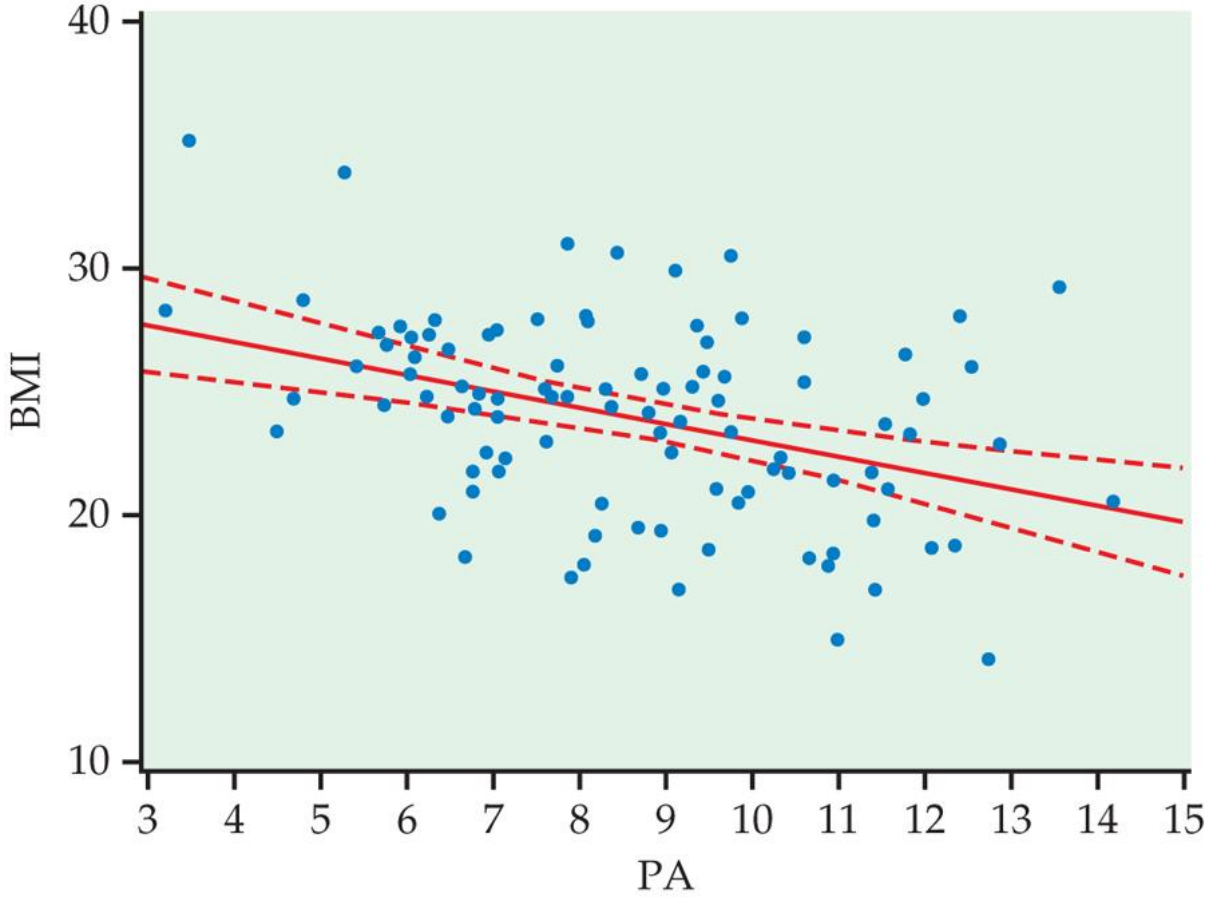
$$\widehat{\mu_y} = b_0 + b_1 \text{ x}^*.$$

Et konfidensintervall med konfidenskoeffisient lik C for μ_y har grenser

$$\widehat{\mu_y} \pm \text{t}^* \; SE_{\widehat{\mu}_y}.$$

Her er t* den verdien som gjør at arealet under tetthetskurven til en t(n-2) fordeling mellom -t* og t* er C .

Eksempel: BMI og fysisk aktivitet.
Konfidensintervaller for forventet respons.



Eksempel: BMI og fysisk aktivitet.

Anta x*= 9000 skritt per dag.

Da er konfidensintervallet for forventet verdi av BMI

$$\widehat{\mu_y} \pm \text{t}^* \ SE_{\widehat{\mu}_y} = b_0 + b_1 \ \text{x}^* \ \pm \text{t} \ * \ SE_{\widehat{\mu}_y}$$

Estimatet for forventningen er 29.578 – 0.655 × 9.000= 23.7 og ved å bruke en statistikkpakke får man at et 95% konfidensintervall er (23.0, 24.4) kr/m^2.

Hvis vi ønsker å predikere en ny observasjon når forklaringsvariabelen har verdi x*, er esimatoren den samme som for den forventede verdien, $\widehat{\mu_y}$, så

$$\hat{y} = b_0 + b_1 \, x^* \, .$$

Men husk at fra formuleringen av modellen

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \, ,$$

Når vi estimerer forventningen tar vi bare hensyn til den systematiske delen $\beta_0 + \beta_1 x_i$.

Når vi skal finne et konfidensintervall for en ny observasjon, må vi også ta hensyn til den tilfeldige delen.

Feilmarginen for å predikere en ny observasjon blir derfor større enn for å estimere forventningen.

Siden $\hat{y}$ er prediksjonen av **ny** observasjon, mens $\widehat{\mu_y}$ er en estimator for en forventning, som kan oppfattes som en grense for gjennomsnittet av **veldig mange** observasjoner, er standardfeilen til $\hat{y}$ større enn standardfeilen til $\widehat{\mu_y}$.

Vi har siden den nye observasjonen er uavhengig at

$$\sigma_{\hat{y}}^2 = \sigma_\epsilon^2 + \sigma_{\widehat{\mu_y}}^2 = \sigma^2 + \left(SE_{\widehat{\mu}_y}\right)^2 .$$

Intervaller som brukes for å predikere en framtidig observasjon kalles **prediksjonsintervaller.**

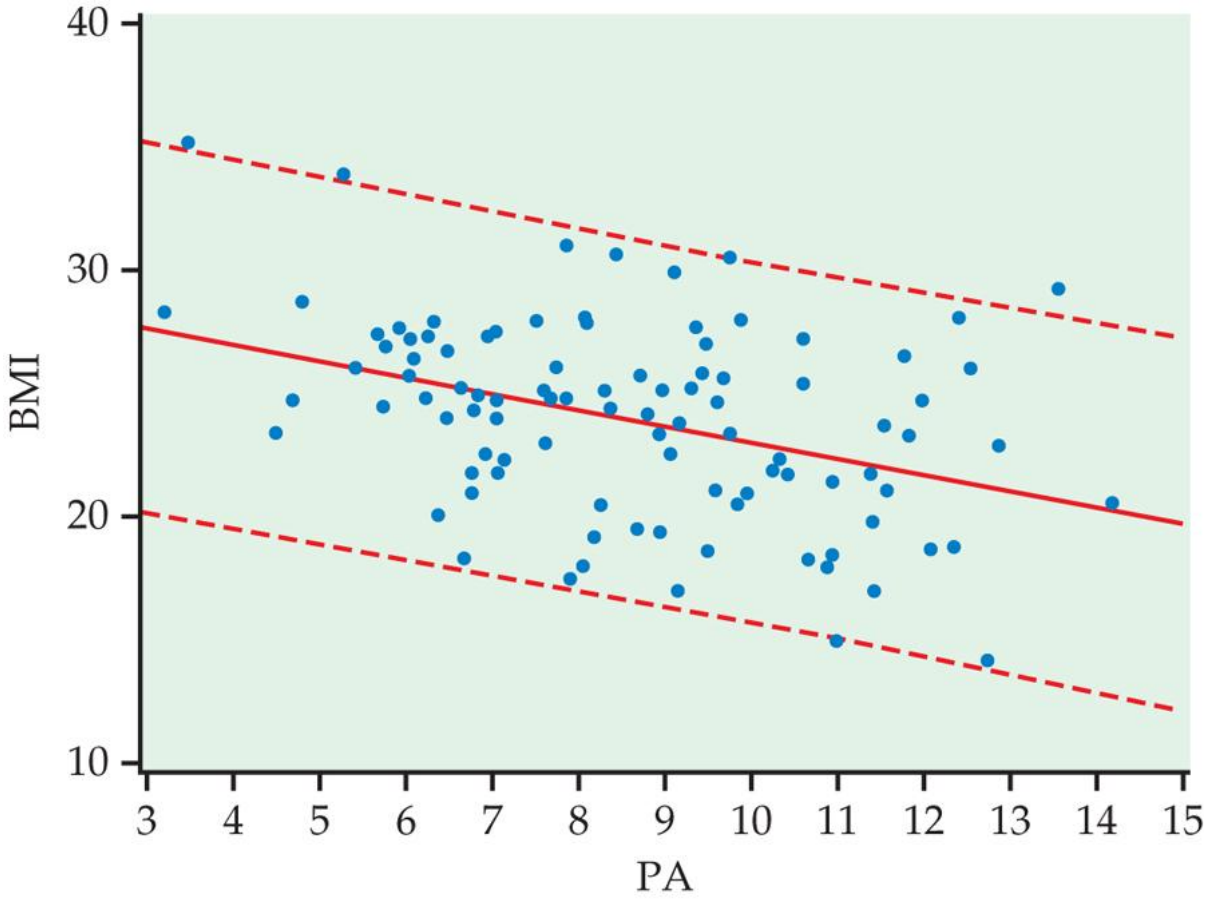
Legg merke til at selv om $\hat{y}$ er en tilfeldig variabel, er tolkningen av prediksjonsintervaller den samme som for konfidensintervaller.

Prediksjonsintervall: Et prediksjonsintervall med konfidenskoeffisient lik C for en ny observasjon når forklaringsvariabelen antar verdien x=x* har grenser

$$\hat{y} \pm \text{t}^* \, SE_{\hat{y}} \, .$$

Her er t* den verdien som gjør at arealet under tetthetskurven til en t(n-2) fordeling mellom -t* og t* er C .

Eksempel: BMI og fysisk aktivitet.
Prediksjonsintervaller for BMI.



Eksempel: BMI og fysisk aktivitet.

Anta x*= 9000 skritt per dag.

Da er prediksjonsintervallet for BMI i en ny observasjon der forklaringsvariabelen har verdi x*

$$\hat{y} \pm \text{t}^* \, SE_{\hat{y}} = b_0 + b_1 \, \text{x}^* \pm \text{t}^* \, SE_{\hat{y}}.$$

Estimatet for $\hat{y}$ er som før 23.7 kg/m^2.
Men nå er feilmarginen større slik at et 95% prediksjonsintervall blir (16.4, 31.0) kg/m^2. Dette fås fra en statistikkpakke.

95% konfidensintervall for μ_y er (23.0, 24.4) kr/m^2, altså betraktelig kortere.

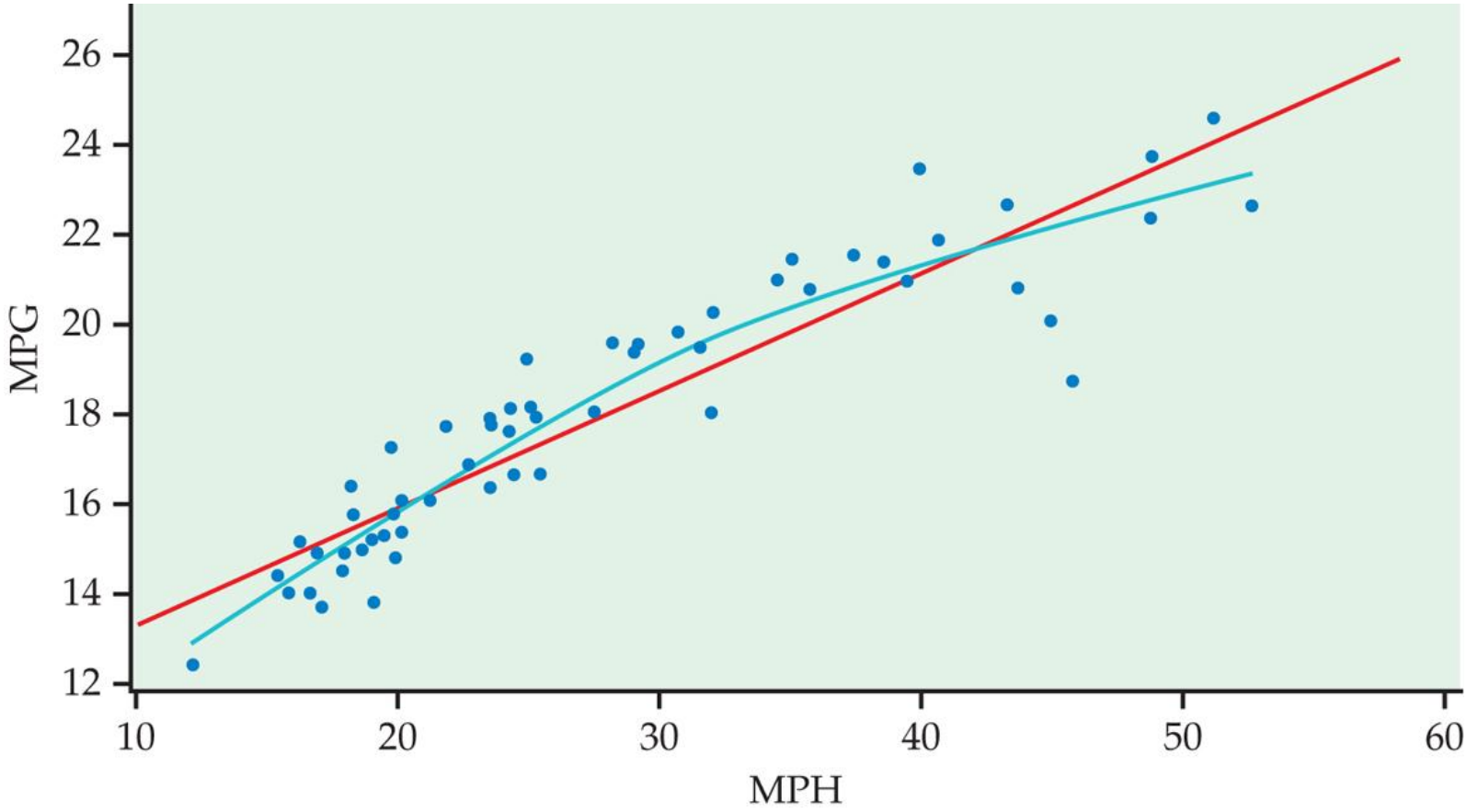
En viktig modelforutsetning er at sammenhengen mellom respons- og forklaringsvariabelen kan beskrives ved en lineær sammenheng, dvs en regresjonslinje.

Når denne forutsetningen ikke er til stede, kan transformasjon av dataene hjelpe.

Eksempel: Bilers ytelse

MPG: antall miles pr gallon, dvs bensinforbruk
MPH: gjennomsnittshastighet.

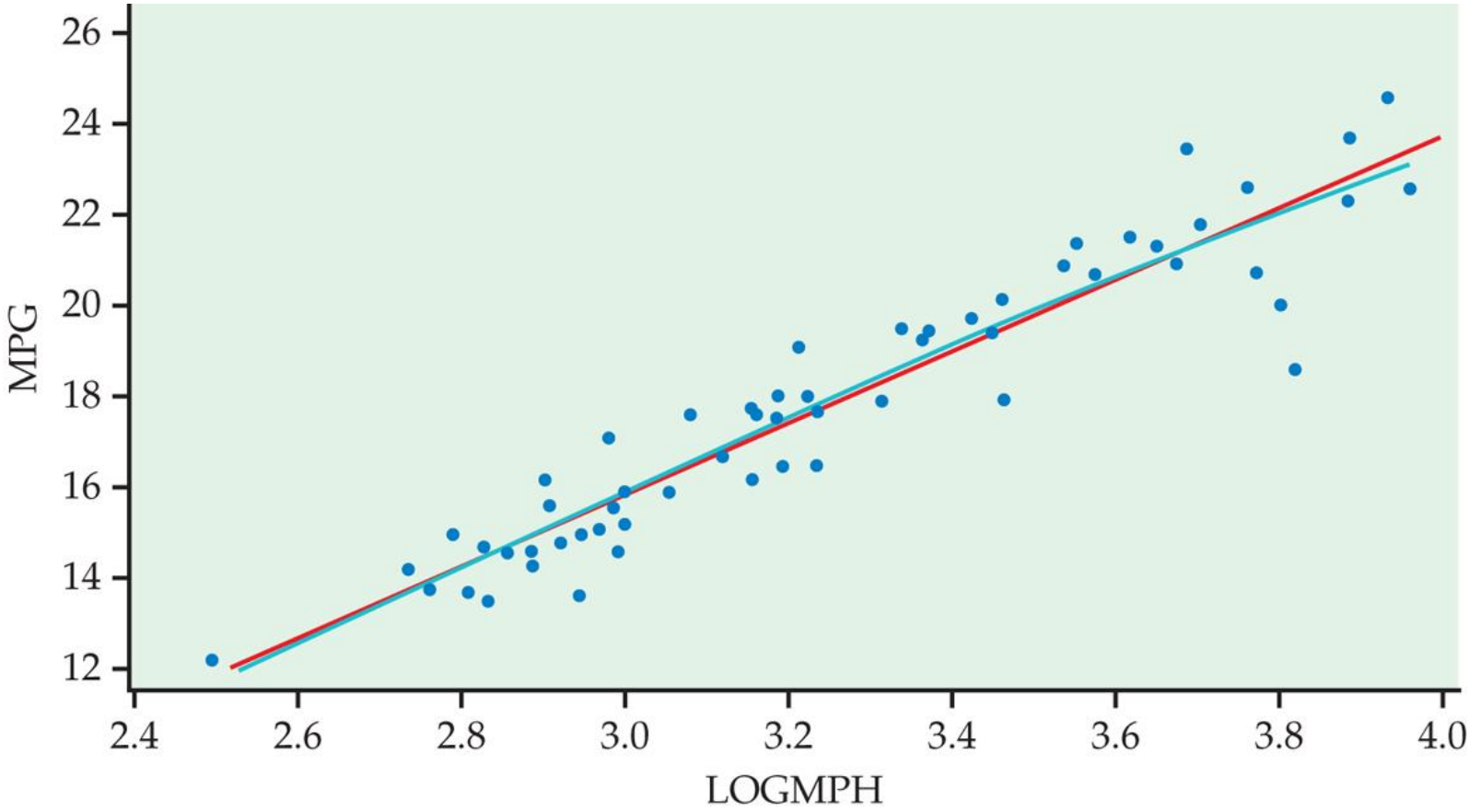
Fra et plott med 60 observasjoner ser man at selv om det er positiv sammenheng, er den ikke helt lineær.



Rødt: Minste kvadrater

Blått: Glatting

Men MPG mot log MPH ser mye mer lineært ut.



Rødt: Minste kvadrater

Blått: Glatting