I dag skal vi lære om statistisk inferens for enkel lineær regresjon

- Vi bygger på minste kvadraters regresjon (kapittel 2)

Statistisk modell for enkel lineær regresjon

- Vi skal estimere modell-parametere
- Vi lærer mer om residualer og modellsjekk
- Statistisk inferens:
  - Hypotestetesting
  - Konfidensintervall
  - Prediksjonsintervall

Hva betyr statistisk inferens i regresjon?

Vi har sett på én variabel i én populasjon i kapitlene 1, 4, 5, 6 og 7 : Først med deskriptiv statistikk, deretter sannsynlighetsmodeller, og til slutt inferensmetoder for forventningsverdi (KI og hyp-test).

Vi vil i dag repetere deskriptiv statistikk for samvariasjon mellom to variable, spesielt minste kvadraters lineær regresjon.

Deretter setter vi opp samvariasjonen som en sannsynlighetsmodell og derfra kan vi legge frem metoder for statistisk inferens for en responsvariabel og en forklaringsvariabel.

Statistisk inferens om forventningsverdi slik vi kjenner det, standard oppsett

- Data $x_1, x_2, \ldots, x_n$. Gjennomsnittet $\bar{x} = \frac{1}{n}\sum x_i$ estimerer 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 forventningsverdien μ.

Vi har undersøkt om forventa respons varierer med type behandling ved å sammenligne to utvalg

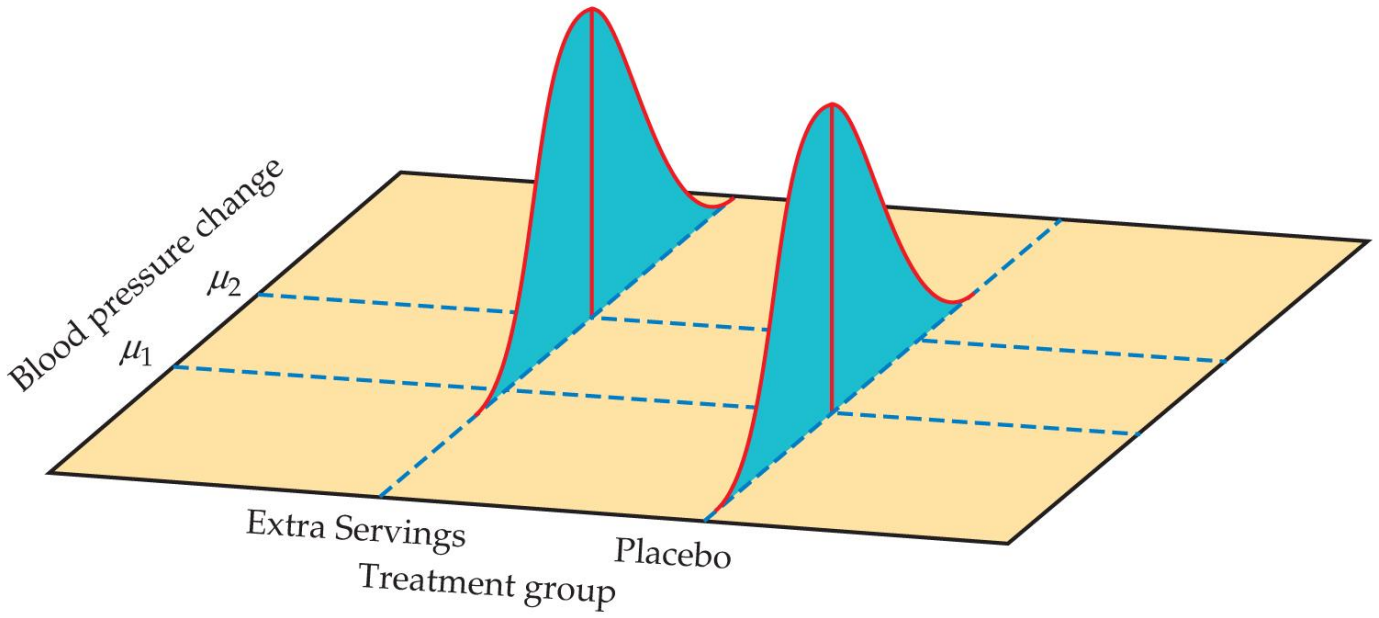


Figure 10.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

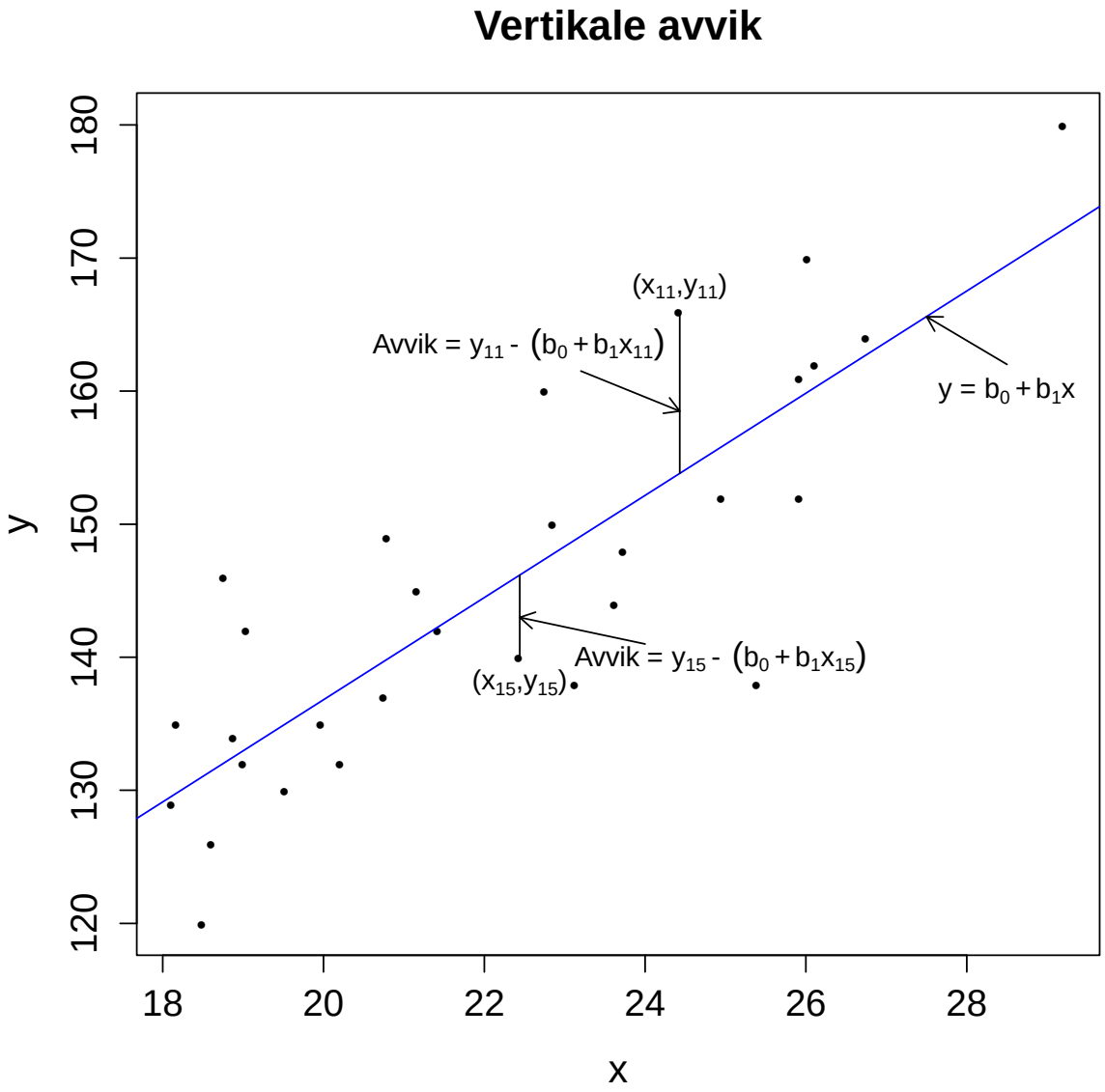
Inferens om modellparametre i regresjon

Vi skal følge et tilsvarende oppsett for lineær regresjon

- Vi har data $(x_1, y_1), \ldots, (x_n, y_n)$
- Vi postulerer 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 $\mu_y = \beta_0 + \beta_1 x$ gitt verdien av $x = x_i$
- De estimerte b_0 og b_1 blir tilfeldige variable med forventninger β_0 og β_1 , og de har 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 $\mu_y = \beta_0 + \beta_1 x$

Estimere linje: Minste kvadraters regresjon

- Vi husker at som regel vil ingen linje vil gi perfekt tilpasning
- Velger å søke minst mulig *vertikal* avstand mellom linje og observert *y*-verdi med minste kvadraters regresjonslinje
- Observasjoner (x_1,y_1) , (x_2,y_2) , ... , (x_n,y_n)
- Minimerer $\sum(error)^2 = \sum(y_i - b_0 - b_1 x_i)^2$


Vertikale avvik
y
x
120
130
140
150
160
170
180
18
20
22
24
26
28
(x11,y11)
Avvik = y11 - (b0 + b1x11)
y = b0 + b1x
(x15,y15)
Avvik = y15 - (b0 + b1x15)

Regresjonslinja er basert på y i rollen som respons og x i rollen som forklaringsvariabel

En responsvariabel måler utfall. En forklaringsvariabel brukes til å forklare endring i responsvariabelen, og dermed til å forklare (deler av) variasjonen i responsvariabelen

Den matematiske relasjonen mellom x og y i lineær regresjon blir som regel ikke det samme («speilet») om man bytter om rollene og velger *x* som responsvariabel og *y* som forklaringsvariabel!

På den annen side, er korrelasjonen *r* symmetrisk i *x* og *y*.

Vi har tidligere sett formlene for minste kvadraters regresjonslinje

Minste kvadraters regresjonslinja blir $\hat{y} = b_0 + b_1 x$

med stigningskoeffisient lik

$$b_1 = r \frac{s_y}{s_x}$$

og konstantledd lik

$$b_0 = \bar{y} - b_1 \bar{x}$$

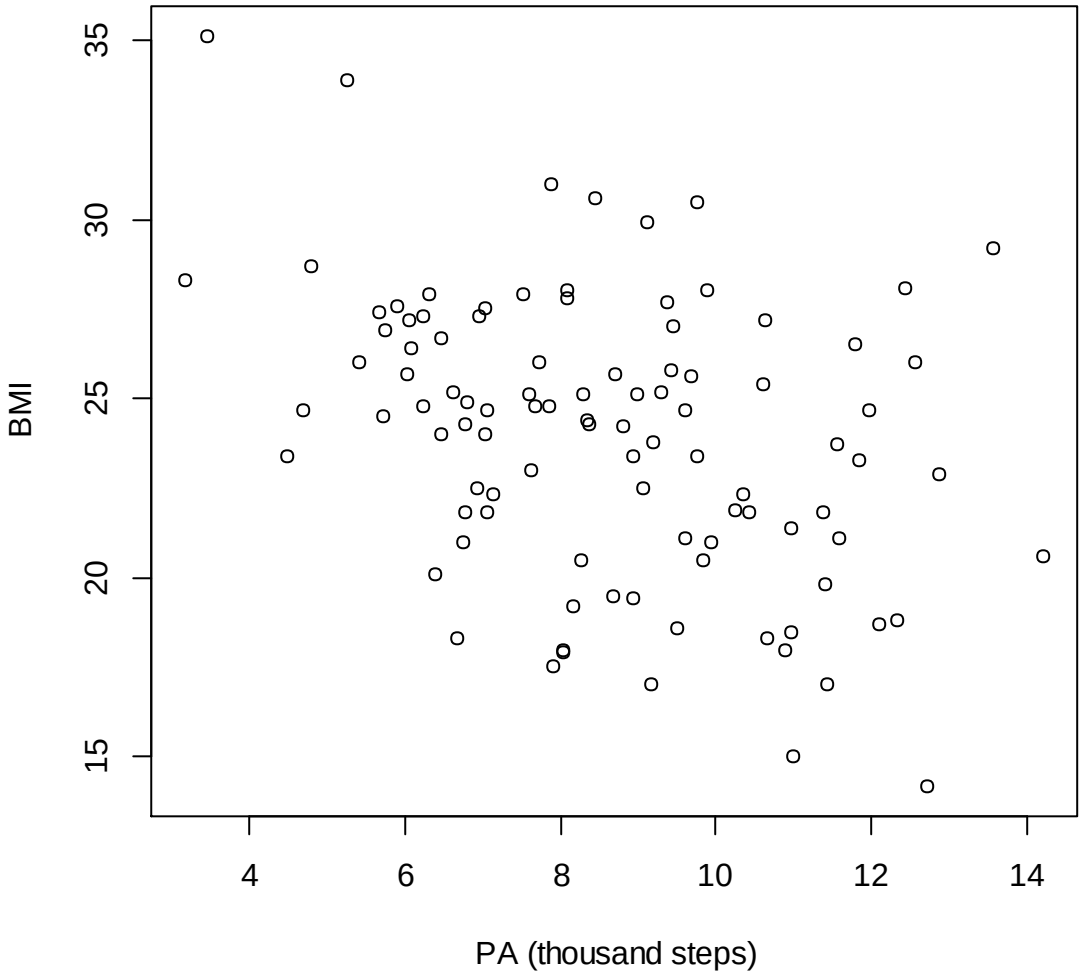
Her er

- r den empiriske korrelasjonen mellom *x* og *y*
- s_x og s_y de empiriske standardavvikene til *x* og *y*
- $\bar{x}$ og $\bar{y}$ gjennomsnittene til *x* og *y*

Eksempel BMI og ukentlig fysisk aktivitet, n=100

Respons: BMI (body mass index)

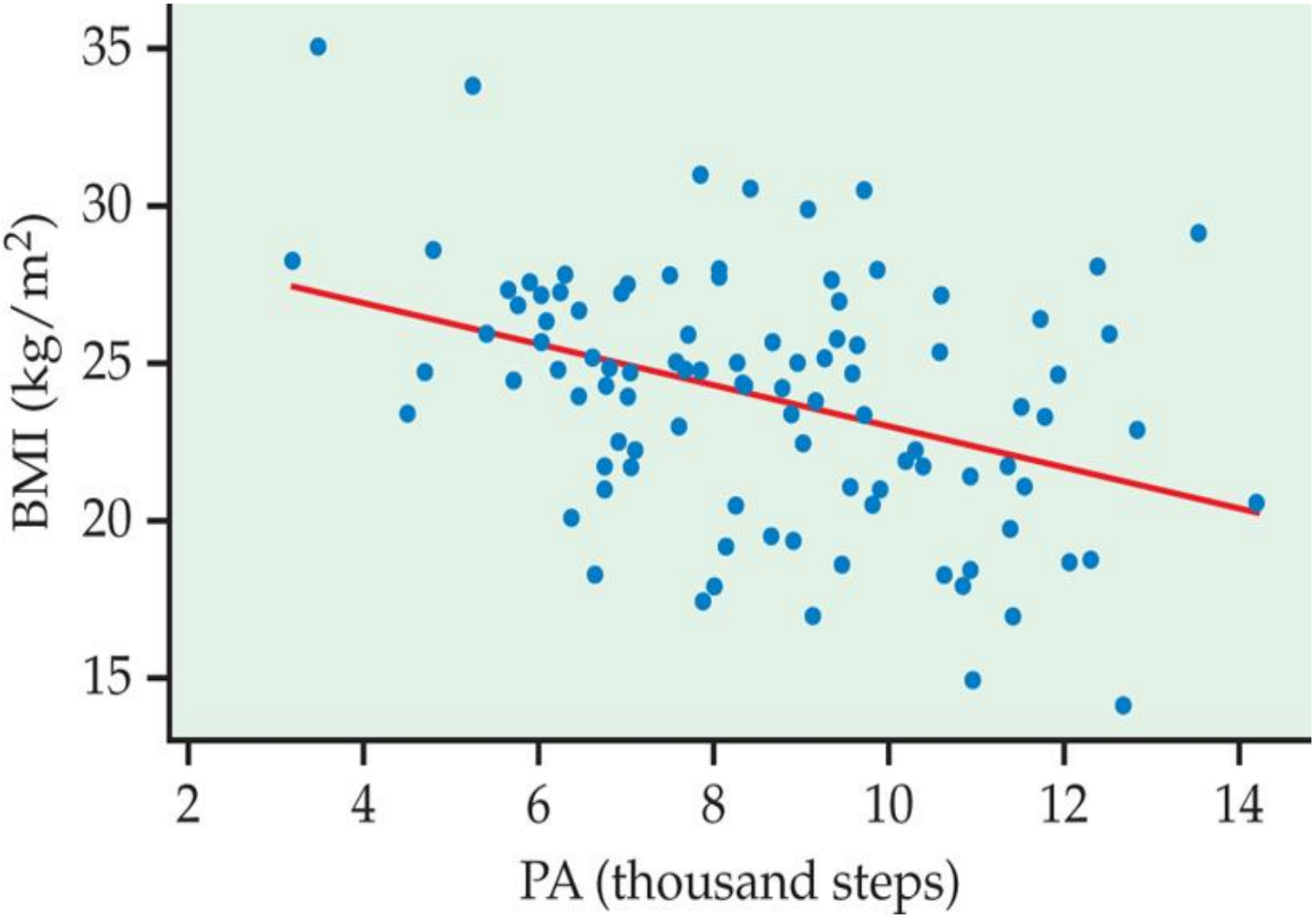
Forklaringsvariabel: Fysisk aktivitet målt ved skritteller i en uke.
Antall tusen skritt per dag.



Bloomberg/Contributor/Getty Images

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

Regresjonslinja blir $BMI = 29.58 - 0.655\, PA$



Enkel lineær regresjon er en statistisk modell for relasjonen mellom en responsvariabel y og en forklaringsvariabel x

Fordelinga til responsvariabelen avhenger av verdien til forklaringsvariabelen.

Dette generaliserer situasjonen der vi sammenligner to utvalg.
Vi går fra to populasjoner til mange ulike verdier for x.

Vi har undersøkt om forventa respons varierer med type behandling ved å sammenligne to utvalg

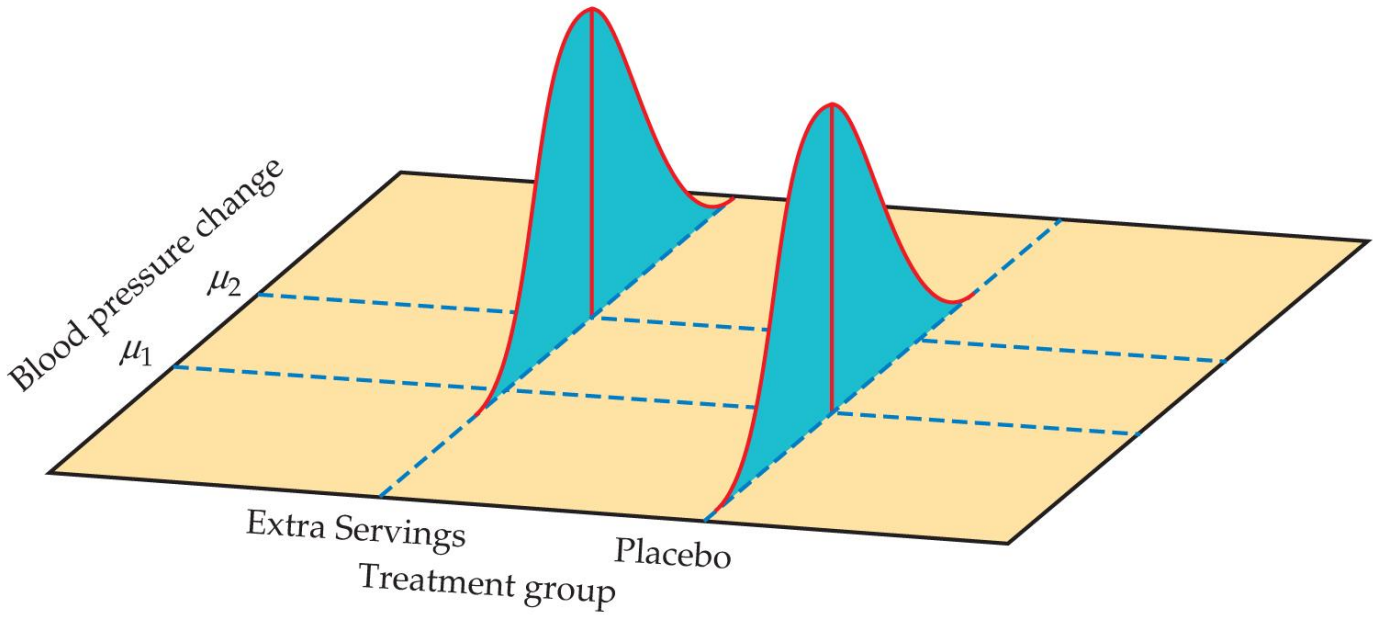
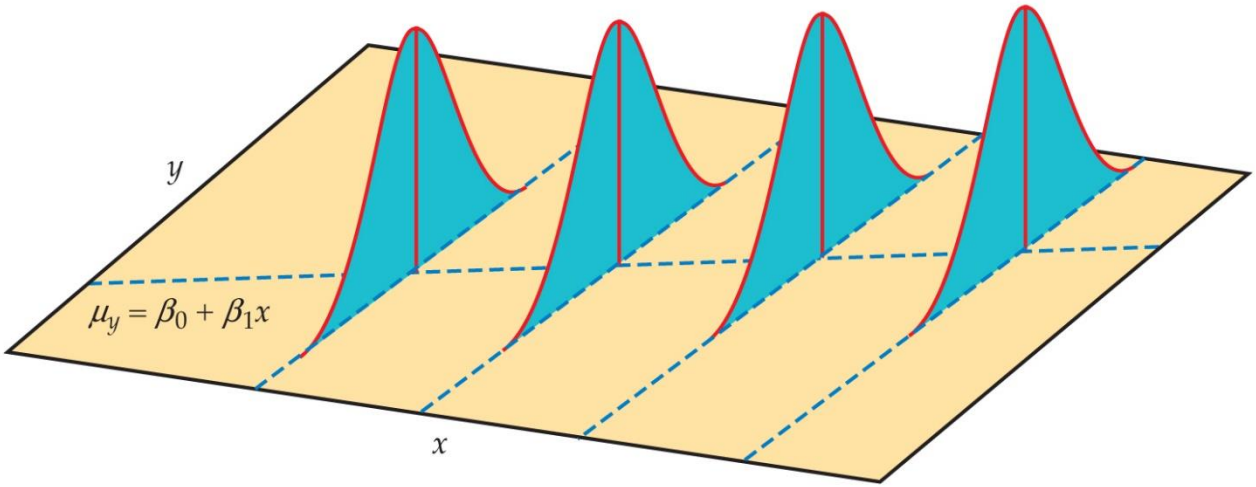


Figure 10.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

I regresjon har vi mange sub-populasjoner, en for hver verdi av forklaringsvariabelen x

Data y = Forventning μ_y + Residual

- Forventninga μ_y til y er en lineær funksjon av *x*
- Observerte *y*-er for gitt verdi av *x* vil variere rundt forventninga μ_y
- Modellen antar at spredning av datapunktene rundt linja, målt av standardavviket σ, er den samme for alle verdier av *x*



Statistisk modell for enkel lineær regresjon

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

Spredninga rundt denne verdien antas å være den samme, angitt ved standardavviket σ, for alle x .

Vi har dermed tre populasjons-parametre, β_0, β_1 og σ.

Antar: variasjonen rundt sub-populasjonsforventninga $\beta_0 + \beta_1 x$ er normalfordelt, dvs. hver y er $N(b_0 + b_1 x, s)$

Dette er det samme som $y = \beta_0 + \beta_1 x + \varepsilon$
der individuell variasjon (feilledd/støy) ϵ er $N(0, s)$

Inferens i regresjon betyr i praksis inferens om

- Stigningskoeffisienten β_1
- Konstantleddet/Skjæringspunktet β_0
- Forventa respons μ_y for gitt verdi *x*
- Individuell fremtidig respons *y* for gitt *x*

Og inferens betyr

- Konfidensintervall
- Statistisk hypotesetest

Statistisk inferens i enkel lineær regresjon utvikles på bakgrunn av egenskapene til minste kvadraters estimatorene, som igjen bygger på modellformuleringa.

Egenskaper minste kvadraters estimatorer

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 $r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

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

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 :

$$\mu_{b_0} = \beta_0 \quad \text{og} \quad \mu_{b_1} = \beta_1$$

Forventninga til en estimator er lik det den skal estimere

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

Det følger at da er også b_0 og b_1 normalfordelte.

Med en utvida versjon av sentralgrenseteoremet kan vi også vise at b_0 og b_1 er tilnærma normalfordelt selv om $y_1, y_2, \ldots, y_n$ og $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 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.

Predikerte verdier og residualer

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.

Estimere σ

Feilleddene ε har standardavvik σ som er en ukjent parameter og som må estimeres. Estimatet er gitt som

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

der $s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$ er estimatet for variansen σ^2 .

Merk at vi deler på *n-2*. Dette skyldes at vi har estimert to parametre β_0 og β_1. Vi sier at vi har brukt 2 frihetsgrader.

Tilsvarende for standardavvik for ett utvalg delte vi på *n-1*. Da hadde vi bare estimert en parameter: forventninga μ.

Vi hadde altså at b_1 er normalfordelt med

$$\mu_{b_1} = \beta_1 \quad \text{og} \quad \sigma^2_{b_1} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)\, s_x^2}$$

Når vi estimerer σ med s, finner vi den estimerte standardfeilen til b_1 , som vi kaller SE_{b1}

I R eller annen programvare finner vi alltid

b_1 og SE_{b1}
b_0 og SE_{b0}

og alt annet vi trenger for å gjøre inferens!

Skritt-teller-eksempelet analysert i R

```
> summary(lm(BMI~PA))

Call: lm(formula = BMI ~ PA)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3819 -2.5636  0.2062  1.9820  8.5078

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Signifikanstester

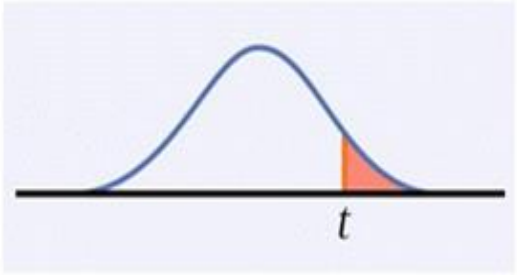
- Ønsker (som oftest) å teste H_0: $\beta_1 = 0$
  - Svarer til at y ikke har noen lineær sammenheng med x
  - Dvs. $\mu_y = \beta_0$ for alle x
- Testobservator $$t = \frac{b_1}{SE_{b_1}}$$
- Når H_0 gjelder er testobservatoren t-fordelt med n-2 frihetsgrader
- I eksempelet er $t = \frac{-0.655}{0.158} = -4.135$ noe som gir tosidig p<0.001
- R og annen statistisk programvare tester også: H_0: $\beta_0 = 0$
  - Sjelden av interesse

To test the hypothesis $H_0: \beta_1 = 0$, compute the **test statistic**

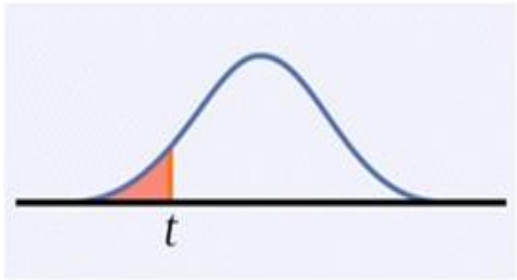
$$t = \frac{b_1}{\text{SE}_{b_1}}$$

The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

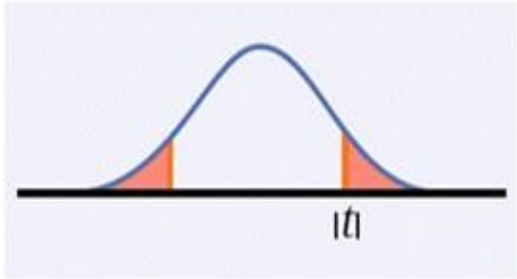
$H_a: \beta_1 > 0$ is $P(T \geq t)$



$H_a: \beta_1 < 0$ is $P(T \leq t)$



$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$



Konfidensintervall for β_1 og β_0

- Generelt: Estimat ± t* $SE_{estimat}$
- Basert på normalfordeling for b_0 og b_1
- t-fordeling fordi σ må estimeres ved *s*
  - n-2 frihetsgrader (= frihetsgradene til *s*)
- t* er verdien der t(n-2)-fordelingen har areal *C* mellom -t* og t*

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR REGRESSION SLOPE AND INTERCEPT

s.568

A **level C confidence interval for the intercept** β_0 is

$$b_0 \pm t^* \mathrm{SE}_{b_0}$$

A **level C confidence interval for the slope** β_1 is

$$b_1 \pm t^* \mathrm{SE}_{b_1}$$

In these expressions t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- SE_{b0} og SE_{b1} avhenger blant annet av *s*
- Dere finner dem ved bruk av R, eller annen programvare

95% Konfidensintervall - eksempel

For BMI-PA dataene er n=100, altså 98 frihetsgrader
97.5 persentilen i t(98) = 1.984 = t*

95% Konfidensintervall for β_1:

-0.655 ± 1.984 x 0.158 = (-0.97, -0.34)

```
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 ***
```

Konfidensintervall for forventa respons, og *Prediksjonsintervall* for en ny observasjon

Vi skal nå se på to utvidelser for statistisk inferens:

- Konfidensintervall for forventa respons $\mu_y = \beta_0 + \beta_1 x^*$ når forklaringsvariablen har en bestemt verdi x*

- Et prediksjonsintervall er et usikkerhetsintervall for responsen $y = \mu_y + \varepsilon = \beta_0 + \beta_1 x^* + \varepsilon$ til en ny observasjon når forklaringsvariablen har en gitt verdi x*.

Konfidensintervall for forventa respons

Med en angitt verdi av forklaringsvariablen x=x* blir forventa respons: $\mu_y = \beta_0 + \beta_1 x^*$. Den naturlige estimatoren for μ_y er dermed $\hat{\mu}_y = b_0 + b_1 x^*$, altså forventninga innsatt minste kvadraters estimatorene.

Denne $\hat{\mu}_y$ er en tilfeldig variabel med forventning μ_y, og standardavvik med litt komplisert formel (det varierer med x).
Men dersom observasjonene y_i er normalfordelt, er $\hat{\mu}_y$ også det.

Pga. at σ estimeres med s blir et nivå C konfidensintervall for μ_y:

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

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

Eksempel: PA og BMI – skritt-telling

Vil predikere forventa BMI μ_y med x = x* = 10, dvs. 10.000 skritt per dag.

Estimert forventa BMI: $\hat{\mu}_y = b_0 + b_1 \text{x}10 = 29.57 - 0.6547\text{x}10 = 23.03$

Men standardfeilen har en litt komplisert formel. Bruker kommandoen `predict` i `R`.

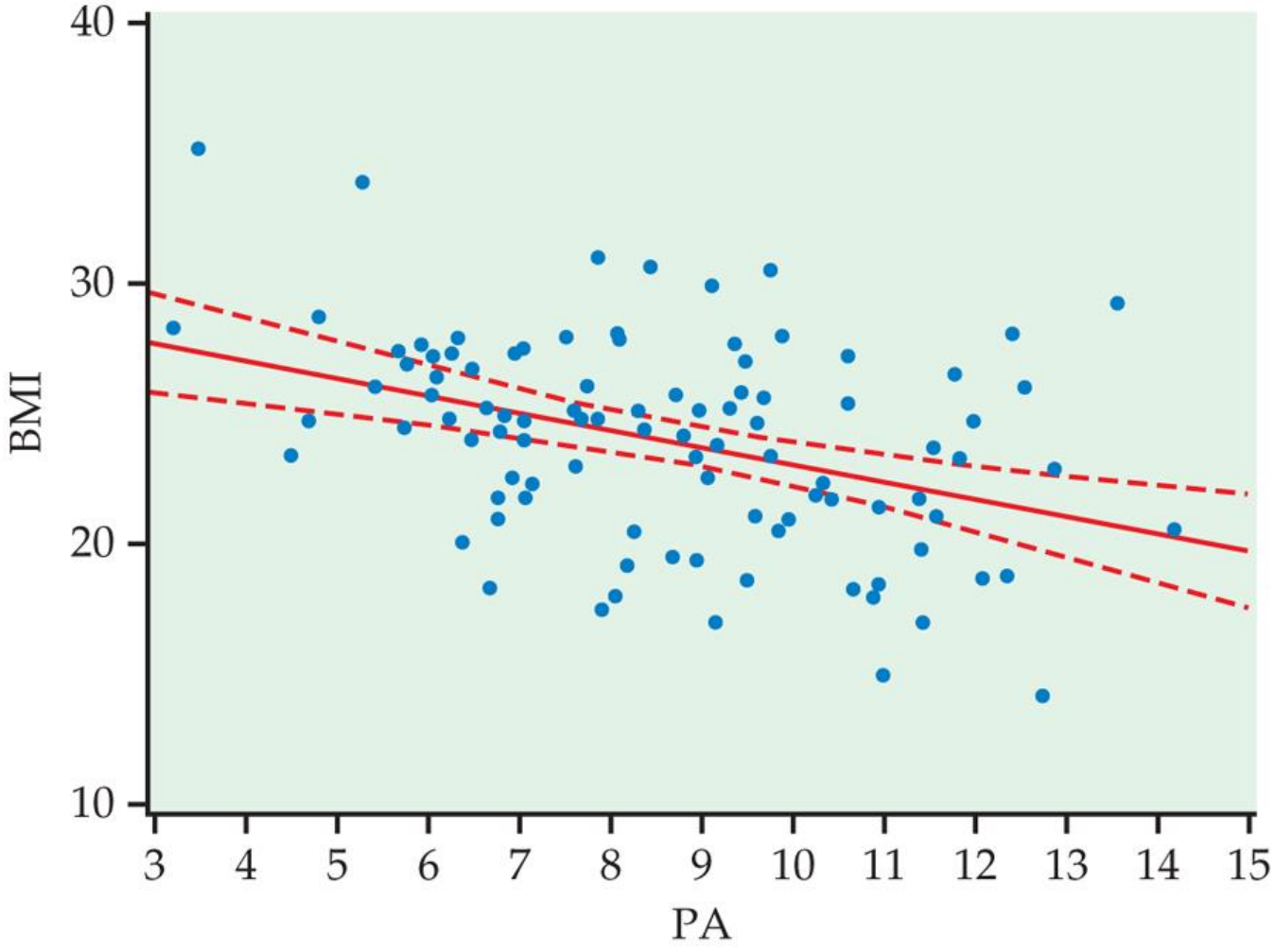
```
> lmbmipa=lm(BMI~PA)
> predict(lmbmipa,newdata=data.frame(PA=10),
                              interval="confidence")

       fit       lwr       upr
1 23.03139 22.18533 23.87744
```

Et 95% konfidensintervall for μ_y blir altså (22.19 , 23.88).

Eksempel: BMI og fysisk aktivitet.

Konfidensintervaller for forventa respons. Smalest for x* nær $\bar{x}$



Prediksjonsintervall for ny observasjon

Verdien for en ny observasjon $y = \mu_y + \varepsilon = \beta_0 + \beta_1 x^* + \varepsilon$ estimeres med $\hat{y} = b_0 + b_1 x^*$, altså samme punkt-estimat som forventninga.

Men for å ta høyde for variasjonen til nye observasjoner, 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 betrakter ny verdi som tilfeldig variabel har vi P(ny verdi i intervall)=C.

Eksempel: BMI og PA

Vil predikere BMI for en ny person med x = x* = 10,
dvs. 10.000 skritt per dag.

Estimat for ny verdi $\hat{y} = b_0 + b_1 \text{x}10 = 29.57 - 0.6547\text{x}10 = 23.03$
(likt som estimatet for forventa verdi)

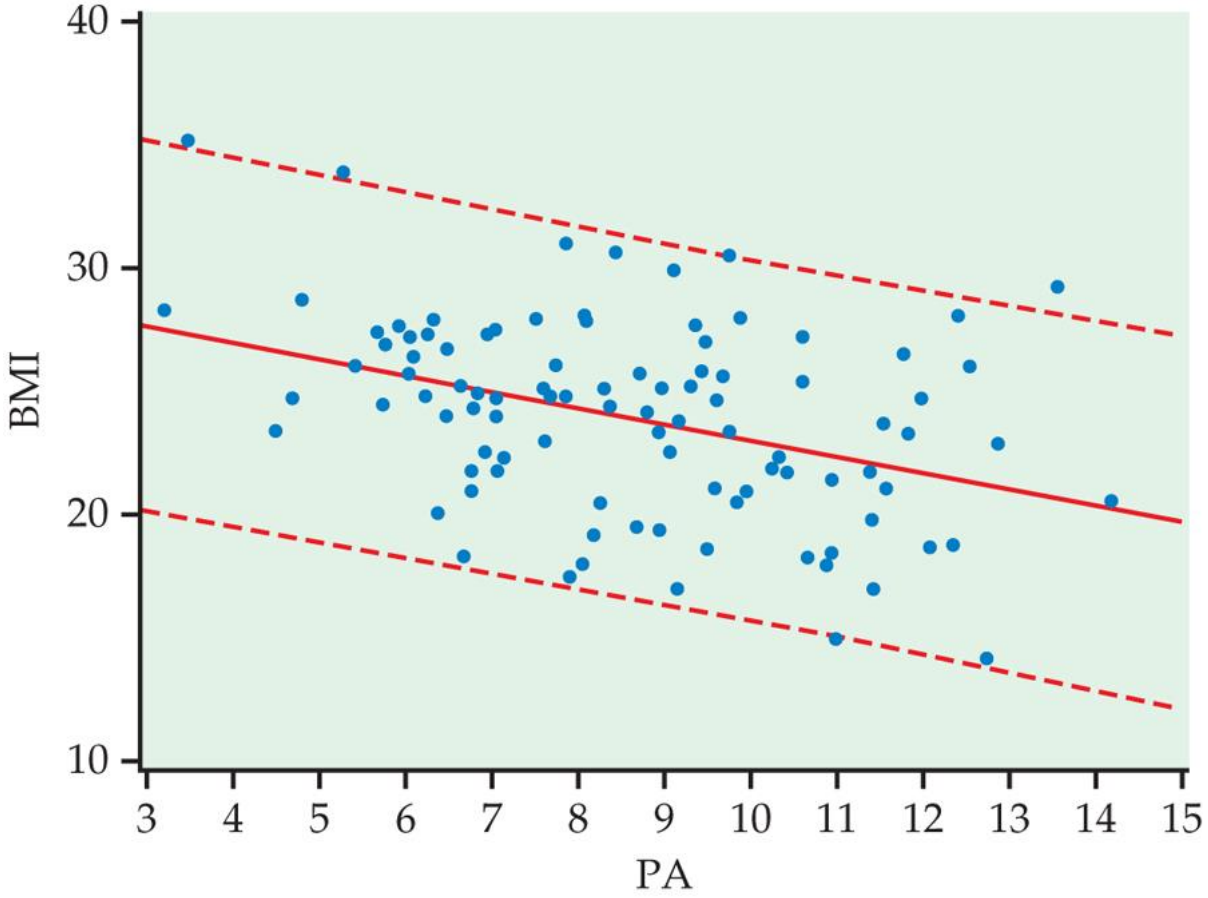
Igjen bruker vi kommandoen `predict` i `R`, men nå for å finne prediksjonsintervallet:

```
> predict(lmbmipa,newdata=data.frame(PA=10),
                  interval="prediction")
       fit      lwr      upr
1 23.03139 15.72921 30.33357
```

95% prediksjonsintervallet for $\hat{y}$ blir altså (15.73 , 30.33), og er altså vesentlig bredere enn konfidensintervallet for μ_y.

Eksempel: BMI og fysisk aktivitet.

Prediksjonsintervaller for BMI. Smalest for x* nær $\bar{x}$, men effekten er mindre synlig, fordi variasjonen ε i nye observasjoner dominerer



Modellsjekk - Residualplott

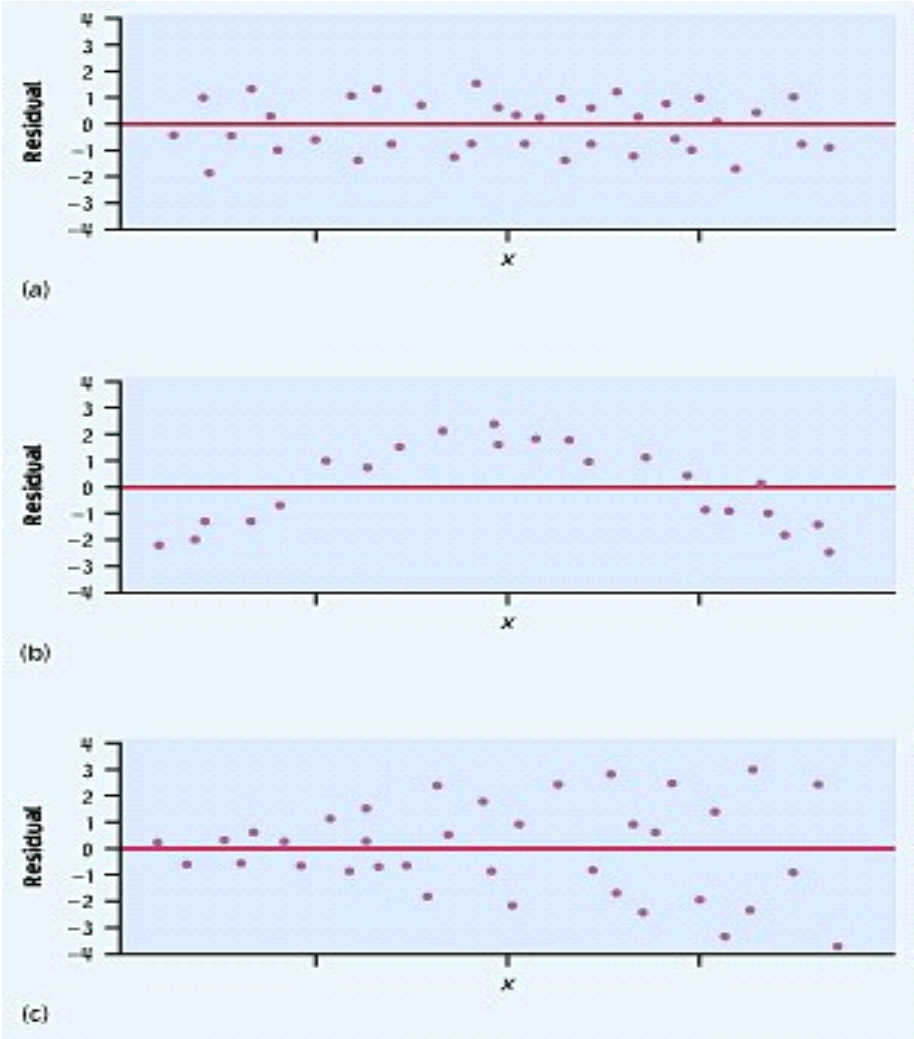
En viktig del av en regresjonsanalyse er å undersøke hvorvidt modellforutsetningene holder.

Det kan omfatte residualsjekk for å avdekke avvik fra

- 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 forklaringsvariablen (hvilket er helt tilsvarende å plotte mot predikert verdier).

Residualer mot forklaringsvariabel for ulike data

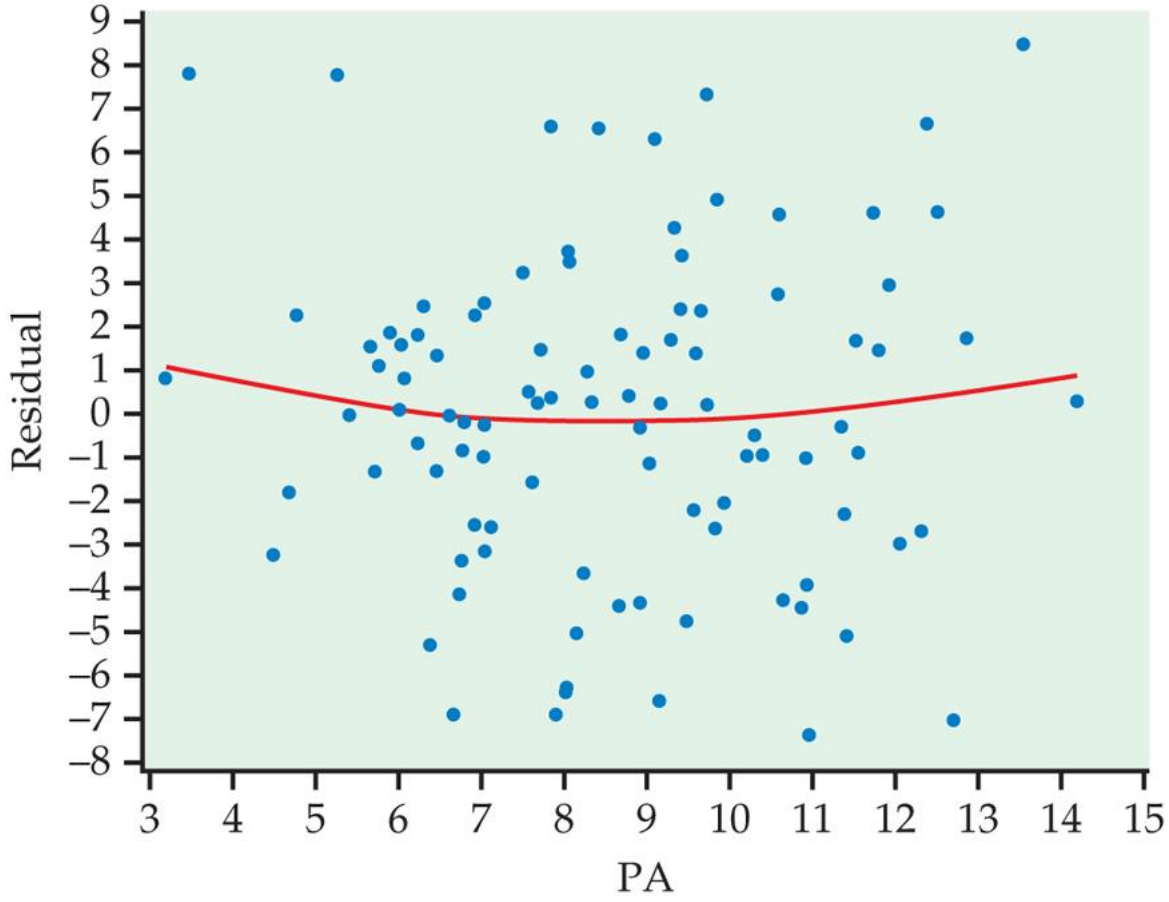


- I det første plottet er modellen OK. Ikke noe mønster i residualene. Linearitet og konstant varians holder.

- I det neste plottet ser vi en klar kurvatur. Dette innebærer at linearitetsantagelsen svikter.

- I det tredje plottet er lineariteten OK, men vi ser at spredningen øker med x. Dette viser at variansen ikke er konstant.

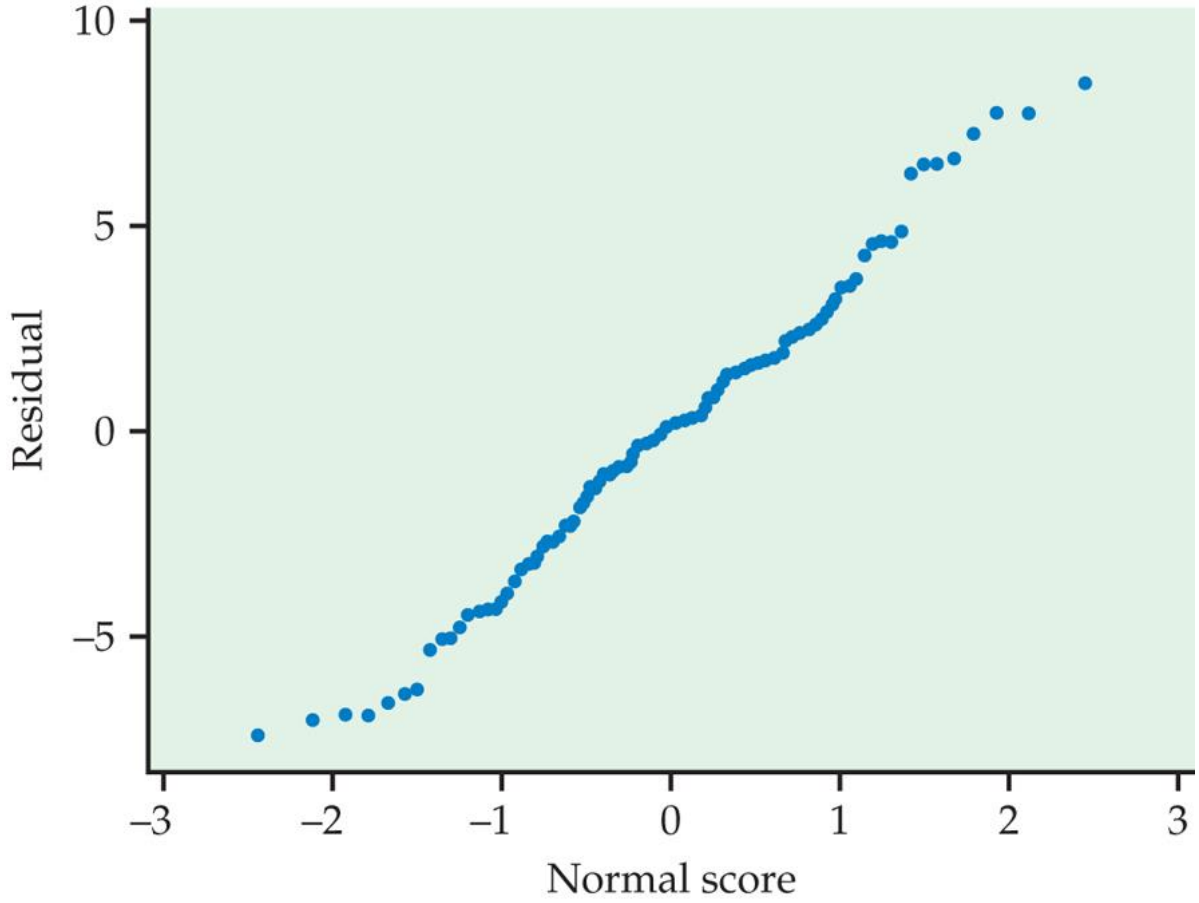
Eksempel: BMI og fysisk aktivitet.



Residualplott kan altså vise avvik fra linearitet og fra konstant varians.

Her: Avvik fra modellantagelsene er ubetydelige.

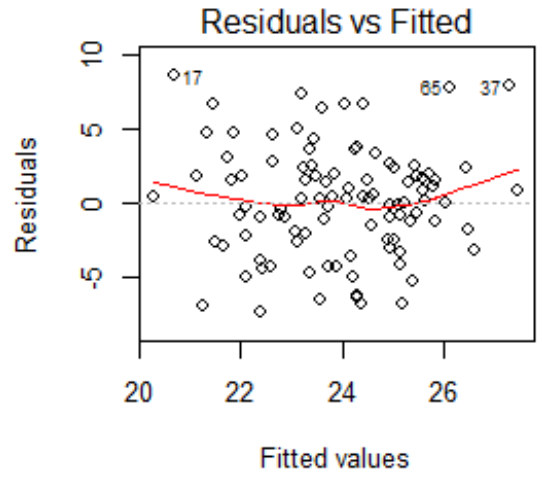
For å sjekke om feilleddene ε_i er normalfordelte kan vi lage et normalfordelingsplott over residualene e_i.

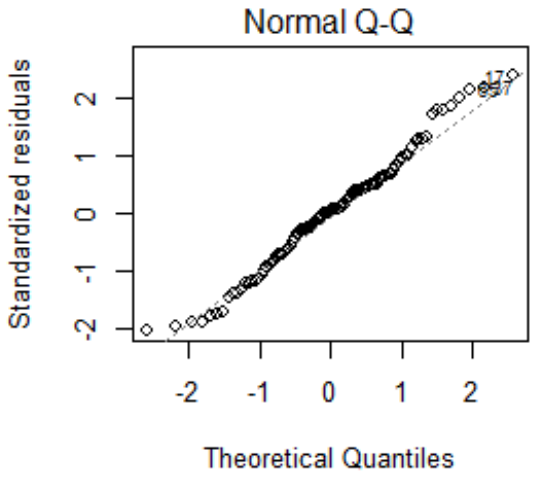


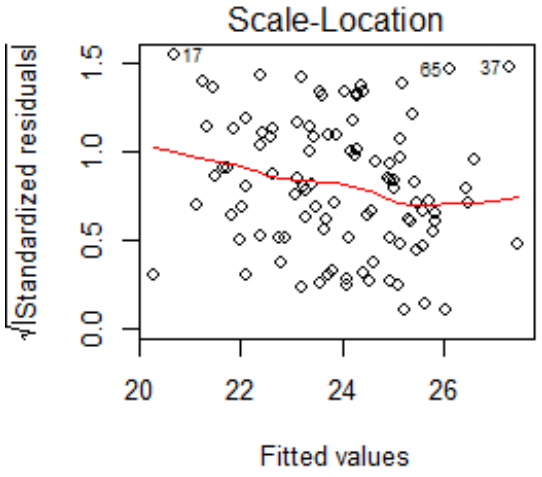
Alternativt kan vi se på et histogram over e_i-ene.

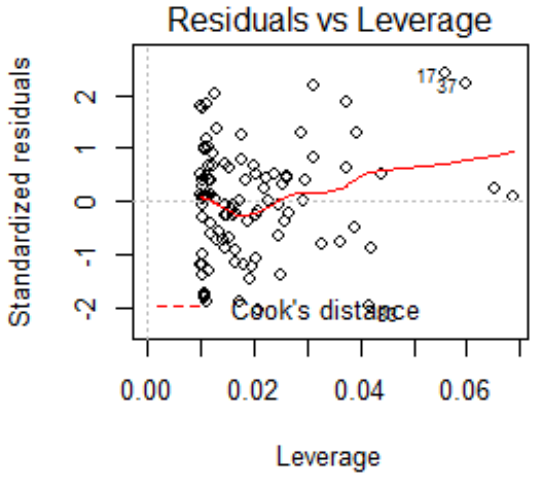
Men: (Små) avvik fra normalfordeling er ikke kritisk!

Residualplott i R: BMI og fysisk aktivitet.









`plot(lm(BMI~PA))` i R gir 4 plott:

- Residualer mot predikerte verdier
- Kvantilplott
- √|Residualer| mot predikerte verdier
- (Standardiserte) residualer mot "leverage" = potensiale til å påvirke estimatene

Eksempel der transformasjon av data ga forbedring av modellen

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.

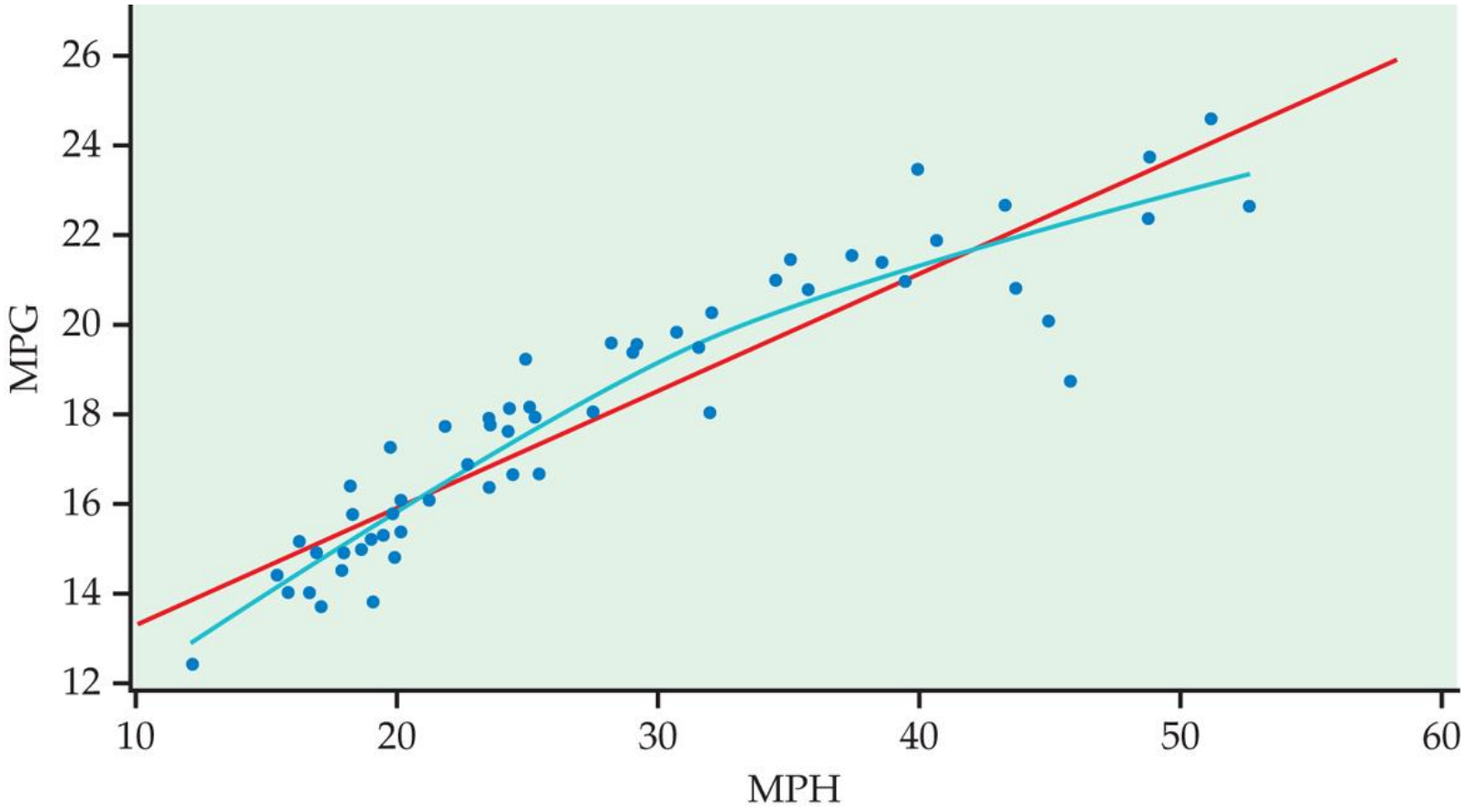
Eksempel: Bilers ytelse

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.

Avvik fra lineær modell:

Tilpasser ikke-lineær regresjonsmodell

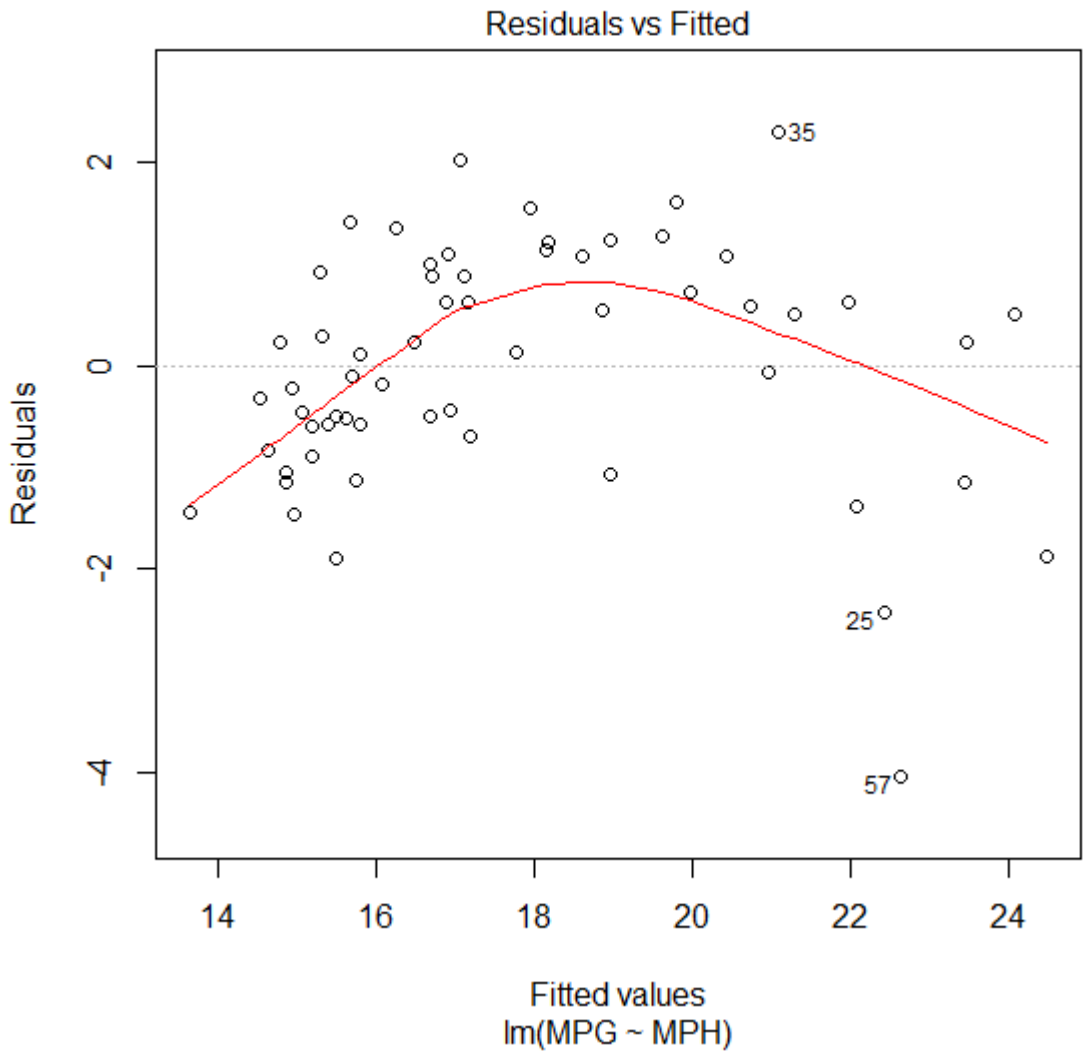


Rødt: Minste kvadrater

Blått: Glatting

Avvik fra lineær modell:

Residualplott fra lineær modell

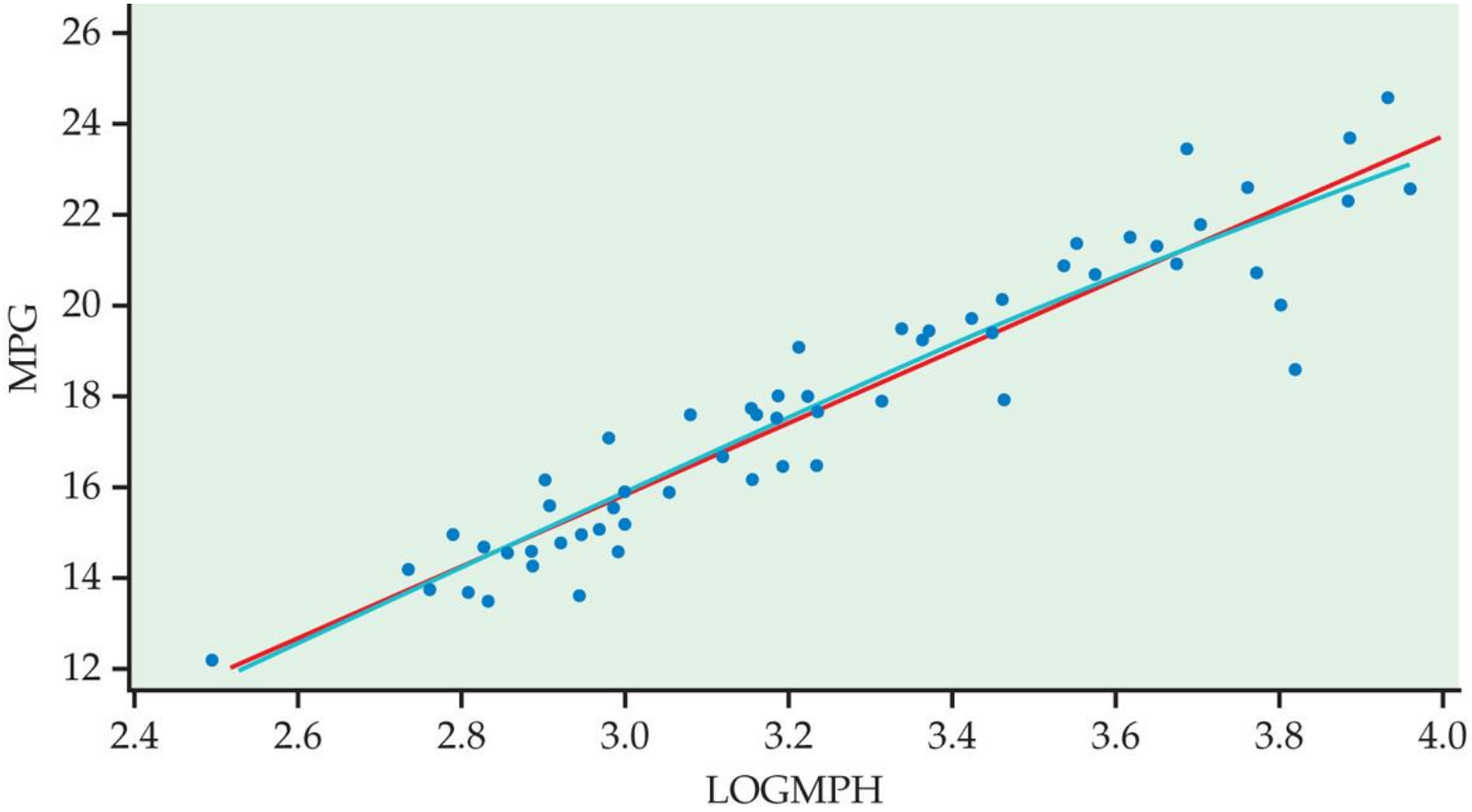



```
> lmmph=lm(MPG~MPH)
> plot(lmmph,which=1)
```

Klar kurvatur i plottet.

For tilpassede verdier nær senter i fordelingen er residualene gjennomgående positive.

Men MPG mot log MPH ser mye mer lineært ut.



Rødt: Minste kvadrater

Blått: Glatting

Oppsummering

- Modell - antagelser
- Estimering av parametre β_0 , β_1 og σ
- Fortolkning
- Testing og konfidensintervall
- Prediksjon og prediksjonsintervall
- Residualplott og modellsjekk
- Mulig forbedring av modell ved transformasjon