# Arbeidskrav 3

Sebastian, Henrik og Gard

## Beskrive sammenhenger

### Hvordan kan vi beskrive sammenhenger mellom to kontinuerlige variabler i en regresjonsmodell og hva er koblingen mellom en korrelasjonskoeffisient og estimatet i en regresjonsmodell?

Når vi skal beskrive sammenhenger mellom to kontinuerlige variabler i en regresjonsmodell, viser modellen hvordan endring i én variabel henger sammen med endringer i den andre. For eksempel kan vi undersøke hvordan den uavhengige variabelen vo2maks påvirker den avhengige variabelen Wmaks

En enkel lineær regresjon kan uttrykkes slik:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

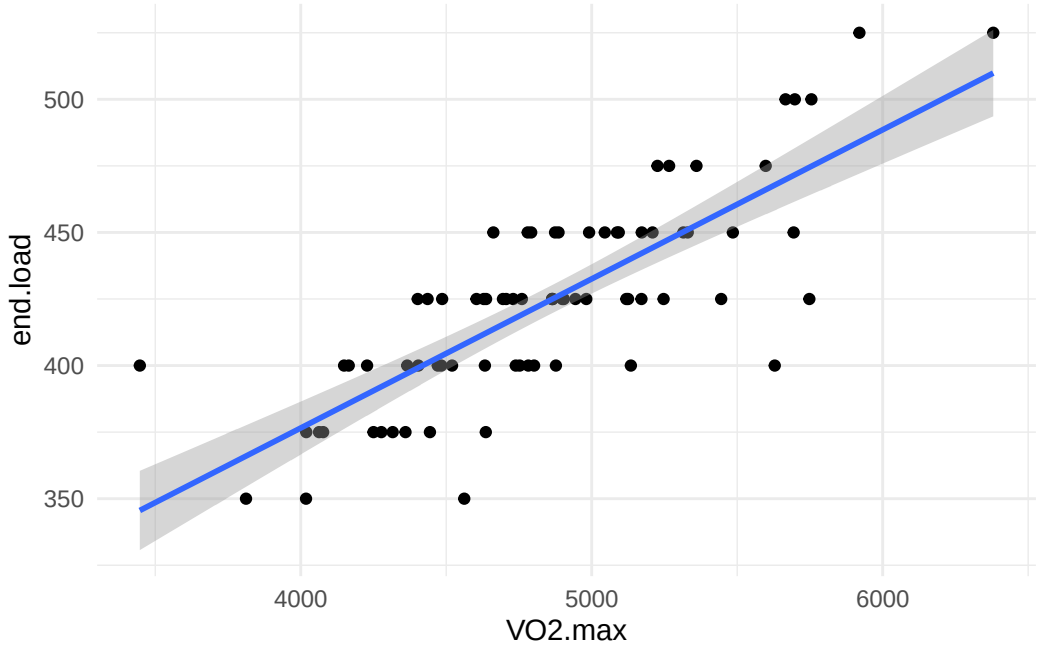
Her representerer $\beta_0$ skjæringspunktet, $\beta_1$ stigningstallet (den gjennomsnittlige endringen i $Y$ per enhet $X$), og $\varepsilon_i$ den tilfeldige variasjonen. Dersom $X = vo2maks$ og $Y = Wmaks$, vil $\beta_1$ vise hvor mye wattmaks i snitt øker per ekstra ml/min i vo2maks.

Sammenhengen kan også beskrives ved hjelp av korrelasjonskoeffisienten ($r$), som varierer fra -1 til 1 og uttrykker styrken og retningen på sammenhengen. En positiv korrelasjon betyr at begge variablene øker sammen, mens en negativ korrelasjon betyr at når én øker, går den andre ned. I enkel lineær regresjon henger korrelasjon og regresjon sammen: jo sterkere $r$, desto tydeligere blir sammenhengen i modellen. Forklaringsgraden ($R^2$) er lik $r^2$, og viser hvor stor andel av variasjonen i $Y$ som forklares av $X$.

Eksempelet viser at korrelasjonen mellom vo2maks og maksimal effekt er r = 0,79, som indikerer en sterk positiv sammenheng. Regresjonskoeffisienten  = 0,056 betyr at for hver økning på 1 ml/min i vo2maks, øker wattmaks med gjennomsnittlig 0,056 watt (56 watt per 1000 ml/min økning). $R^2$ viser at 63% av variasjonen i wattmaks kan forklares av forskjeller i vo2maks.

| Mål | Verdi |
|---|---|
| Regresjonskoeffisient ($\beta$ ) | 0.056 |
| Korrelasjon (r) | 0.792 |
| Forklaringsgrad ($R^2$) | 0.628 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 152.505 | 24.405 | 6.249 | 0 |
| VO2.max | 0.056 | 0.005 | 11.178 | 0 |




**Hvordan kan vi beskrive sammenhengen mellom en kontinuerlig avhengig variabel og en uavhengig kategorisk variable med en regresjonsmodell og hvordan tolker vi estimatet**

Når den avhengige variabelen er kontinuerlig, men den uavhengige er kategorisk, kan vi fortsatt bruke en lineær regresjonsmodell for å undersøke forskjeller mellom grupper. Vi kan undersøke

| r.squared | adj.r.squared | statistic | p.value | sigma |
|---|---|---|---|---|
| 0.628 | 0.623 | 124.944 | 0 | 23.096 |

[

Regresjonsestimat for modell: speed ~ gender ]
Regresjonsestimat for modell: speed ~ gender

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 9.580 | 0.019 | 510.669 | 0 |
| genderM | 0.862 | 0.025 | 35.122 | 0 |

[

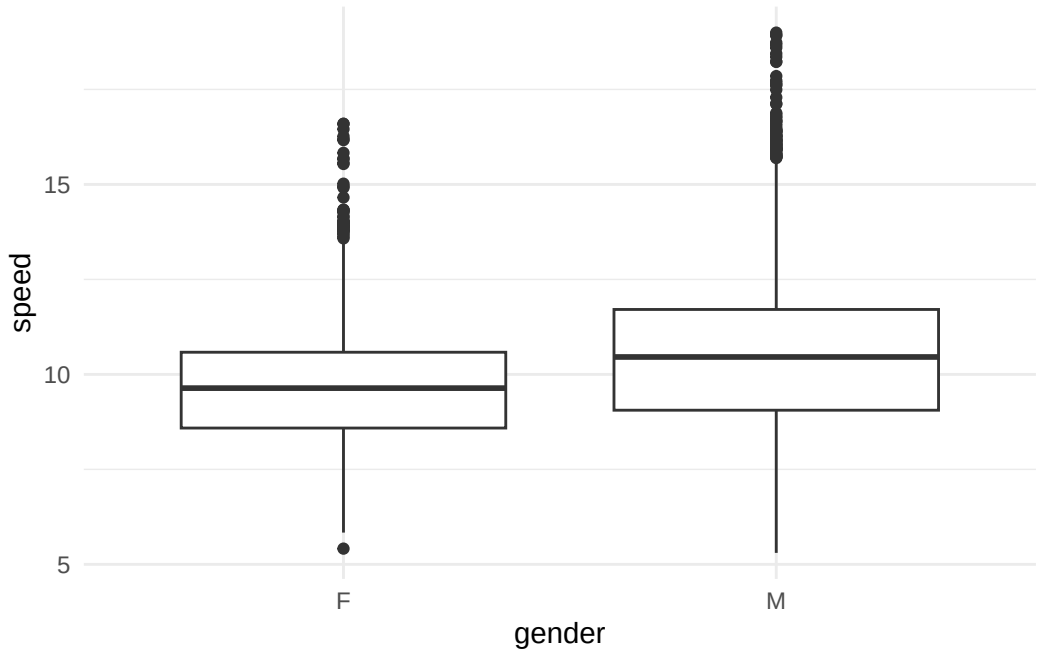
Modelloppsummering for speed ~ gender ]
Modelloppsummering for speed ~ gender

| r.squared | adj.r.squared | statistic | p.value | sigma |
|---|---|---|---|---|
| 0.054 | 0.054 | 1233.531 | 0 | 1.776 |

om menn og kvinner har forskjellig gjennomsnittlig løpshastighet i Boston Marathon. Modellen kan skrives som:

$$Y_i = \beta_0 + \beta_1(\text{genderM})_i + \varepsilon_i$$

Her representerer $\beta_0$ gjennomsnittet for referansegruppen (kvinner, kodet som 0), mens $\beta_1$ viser forskjellen i gjennomsnittlig hastighet mellom menn og kvinner. Dersom $\beta_1$ er positiv, betyr det at menn i gjennomsnitt har høyere hastighet enn kvinner med en verdi tilsvarende størrelsen på $\beta_1$. Resultatene fra regresjonsmodellen viste at gjennomsnittshastigheten for kvinner var = 9,58 km/t, mens menn i gjennomsnitt løp = 0,86 km/t raskere.




**Hvordan kan vi tolke estimatet i en generalisert lineær modell hvor den avhengige variabelen er enten 1 eller 0? Hva betyr «link-function» i denne sammenhengen og hva gjør den?**

Når den avhengige variabelen bare kan være 1 eller 0 (for eksempel skade = 1 og ingen skade = 0), bruker vi en logistisk regresjon. Det er en type generalisert lineær modell som gjør det mulig å analysere sannsynligheter. Siden sannsynligheter alltid må ligge mellom 0 og 1, brukes en link-funksjon kalt logit, som gjør om sannsynligheten p til log-odds. Dette lar oss bruke en lineær modell til å beskrive sammenhengen mellom variablene. Logit-modellen kan skrives som:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_i$$

I denne modellen er p sannsynligheten for utfallet (for eksempel skade). Forholdet p / (1 − p) kalles oddsen, altså hvor mye mer sannsynlig det er at utfallet skjer enn at det ikke skjer. Logit-funksjonen tar logaritmen av oddsen slik at vi kan bruke en vanlig lineær modell selv om sannsynligheten bare går fra 0 til 1.  er log-oddsen i referansegruppen (når alle variabler = 0), mens  viser hvor mye log-oddsen endrer seg når den uavhengige variabelen øker med én enhet. Hvis vi tar exp( ), får vi en odds ratio som viser hvordan oddsen for skade endres i praksis.

I eksempelet under undersøkes sammenhengen mellom oppvarmingsrutine og hamstringsfleksibilitet og risiko for skade. Modellen predikerer log-odds for skade, som deretter omregnes til

odds ved hjelp av exp()-funksjonen. Når hamstringsfleksibiliteten er 80, viser resultatene at oddsen for skade er 1,70 uten oppvarming og 0,55 med oppvarming. Dette betyr at spillere som følger oppvarmingsrutinen har lavere odds for skade sammenlignet med dem som ikke gjør det.

```
# A tibble: 2 x 3
  Warmup_Routine_Adherence Hamstring_Flexibility  odds
                     <dbl>                 <dbl> <dbl>
1                        0                    80 1.70
2                        1                    80 0.554
```

### *Predikere mulige observasjoner*

- Bruk data fra datasettet strengthvolume og lag en prediksjonsmodell for legext basert på legpress.
- Bruk data fra en tidspunkt (time) og et treningsvolum (sets)

```
Call:
lm(formula = legext ~ legpress, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-33.270 -17.122  -3.403  16.324  34.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.15202   15.04075   1.340     0.19
legpress     0.27032    0.05693   4.748 4.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.64 on 31 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.421, Adjusted R-squared:  0.4024
F-statistic: 22.54 on 1 and 31 DF,  p-value: 4.416e-05
```

```
Call:
lm(formula = legext ~ sex + legpress, data = dat)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-26.300  -7.317  -0.837   9.239  32.127

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.63303   11.70910   2.872  0.00741 **
sexmale     28.44486    5.79124   4.912 2.99e-05 ***
legpress     0.16431    0.04819   3.410  0.00188 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.87 on 30 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.6791,    Adjusted R-squared:  0.6577
F-statistic: 31.74 on 2 and 30 DF,  p-value: 3.941e-08


       fit      lwr      upr
1 74.21523 33.01092 115.4195


      fit      lwr      upr
1 94.9399 62.54697 127.3328


       fit      lwr     upr
1 66.49504 35.10477 97.8853
```

**Hvordan spiller kjønn (sex) inn på prediksjonen, hvordan kan du bruke kjønn for å si noe om prediksjoner innad kjønn og i gjennomsnitt i begge kjønn?**

Kjønn har en stor effekt. Modellen viser oss at menn har 28.4 kg høyere forventet legext enn kvinner, selv når begge har 200kg legpress. Dette betyr at for å gjøre en god prediksjon, må vi absolutt ta kjønn med i beregningen.

For prediksjoner innad i hvert kjønn bruker vi den samme stigningen på 0.164, men med forskjellig utgangspunkt. For en gjennomsnittlig prediksjon over begge kjønn får vi et mindre nøyaktig estimat som egentlig ikke passer for noen av gruppene.

**Modellen gir deg et estimat, men for en gitt verdi på legpress, hva sier modellen om i hvilket område vi kan forvente å finne nye observasjoner?**

For en gitt legpress verdi (200kg), gir modellen oss et prediksjonsintervall. For en ny kvinnelig utøver med 200 kg i leg press predikerer modellen et forventet gjennomsnitt (fit) på 66,4 kg i leg extension, med et 95 % prediksjonsintervall fra 35,1 (nedre grense) til 97,9 kg (øvre grense). Dette intervallet gjenspeiler den naturlige variasjonen som forekommer mellom individuelle utøvere.

***Trekke slutninger***

- Bruk datasettet strengthvolume og formuler en modell som gir oss et estimat på forskjell i gjennomsnitt mellom sets i forandring fra tidspunkt pre til tidspunkt post i legext. Gi begrunnelse til valg av modell og håndtering av data.

| | Parameter | Estimate | SE |
|---|---|---|---|
| 1 | (Intercept) | 56.86 | 3.12 |
| 2 | timepost | 32.47 | 1.93 |
| 3 | setssingle | 0.45 | 1.85 |
| 4 | timepost:setssingle | -4.39 | 2.72 |
| 5 | sd__(Intercept) | 17.68 | NA |
| 6 | sd__Observation | 8.19 | NA |

**Begrunnelse**

Den individuelle utgangspunktvariasjonen av leg extensions er på 17,68 enheter (SD_Intercept), i tillegg er pre -og postverdiene avhengige, de er fra samme individ. Dette betyr at det ikke er gunstig å anvende en tradisjonell lineær modell, siden den antar at pre- og postverdiene var uavhengige. Modellen hadde heller ikke tatt rede for variasjon i utgangspunkt. Derfor valgte vi å ta i bruk en mixed model for å finne den gjennomsnittlige forskjellen av endring fra pre til post mellom ett og flere sett. En mixed model vil anta at det er variasjoner imellom individer og innad i samme individ. I tillegg vil den anta at det er sammenheng mellom pre- og posttest, siden det er de samme individene som gjennomfører begge tester.

**Hvordan kan vi bruke regresjonsmodellen for å si noe om populasjonen som dataene kommer fra?**

Modellen estimerer at populasjonsgjennomsnittlig utgangspunkt for multiple sets er 56,86 kg (intercept), og 57,31 kg for single sets (intercept + setssingle). Den estimerte økningen for populasjonen fra pre til post i multiple sets er 32,47 kg (timepost), på single sets er det 4,39 kg mindre enn multiple sets (timepost:setssingle). Modellen estimerer at gjennomsnittlig forskjell

i utgangspunkt mellom multiple og single sets gruppen er liten (setssingle = 0.45 kg), men på individnivå er variasjonen i utgangspunktet stor(sd_(intercept) = 17,68 kg).