Arbeidskrav 3, Gard, Sebastian og Henrik

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:

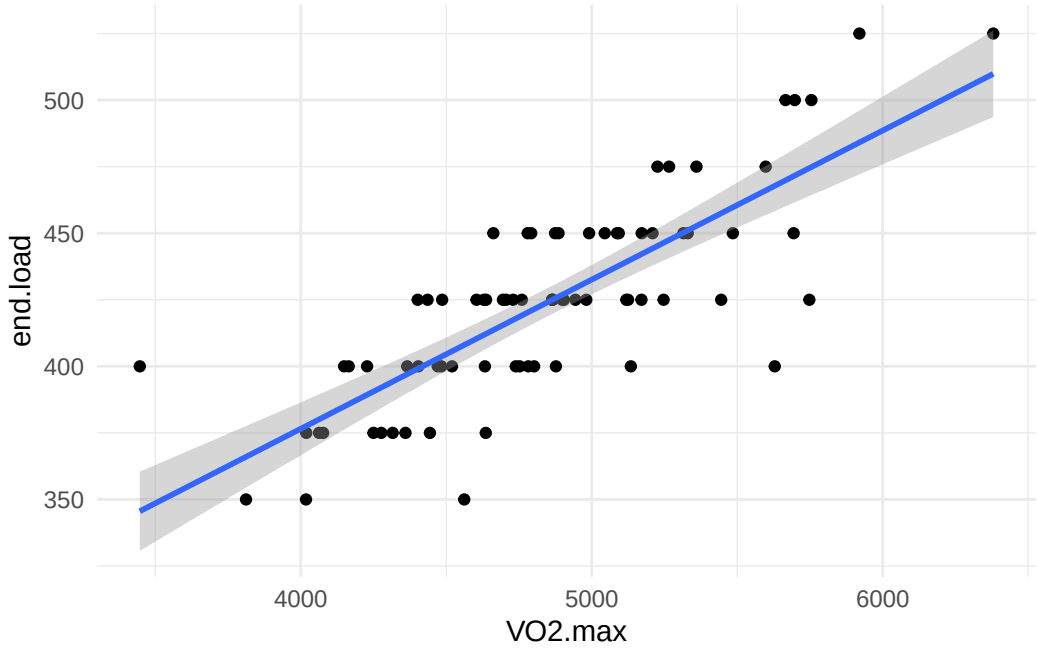
Yi= 0+ 1Xi+ i

Her representerer β_0 skjæringspunktet, β_1 stigningstallet (den gjennomsnittlige endringen i Y per enhet X), og ε_i den tilfeldige variasjonen. Dersom $X = vo2maks$ og $Y = Wmaks$, vil β_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.

```
Call:
lm(formula = end.load ~ VO2.max, data = cycling)

Residuals:
    Min      1Q  Median      3Q     Max
-67.786 -16.277   0.006  14.970  54.428

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.525e+02  2.440e+01   6.249 2.38e-08 ***
VO2.max     5.601e-02  5.011e-03  11.178  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.1 on 74 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.628, Adjusted R-squared:  0.623
F-statistic: 124.9 on 1 and 74 DF,  p-value: < 2.2e-16



    Pearson's product-moment correlation

data:  cycling$VO2.max and cycling$end.load
t = 11.178, df = 74, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6903832 0.8636365
sample estimates:
      cor
0.7924875
```



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

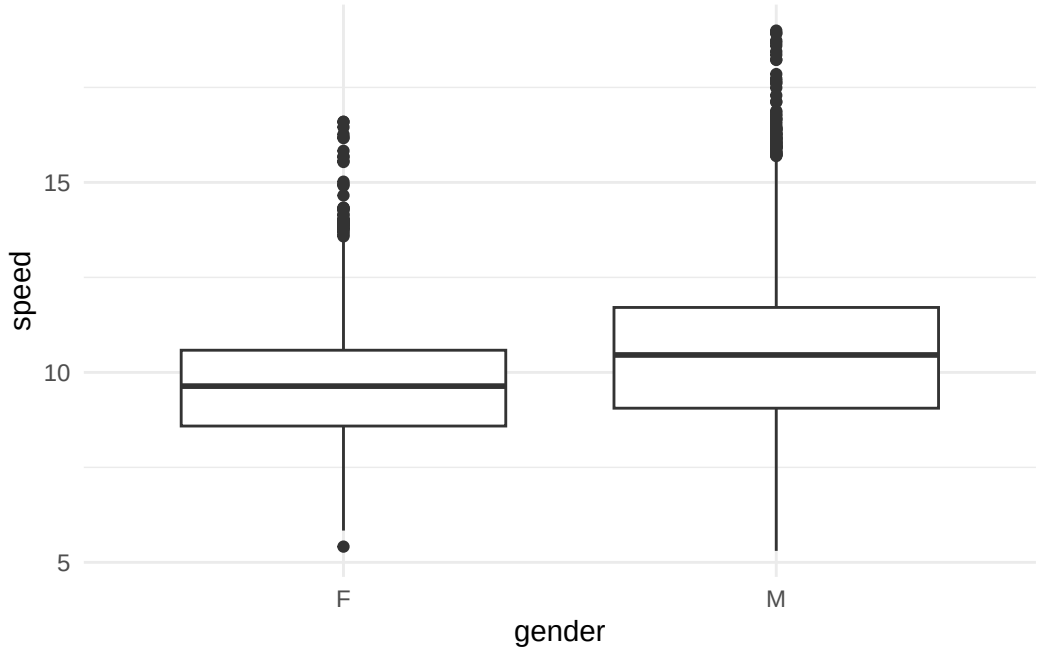
Når den avhengige variabelen er kontinuerlig, men den uavhengige er kategorisk, kan vi fortsatt bruke en lineær regresjonsmodell for å undersøke forskjeller mellom grupper. Vi kan undersøke om menn og kvinner har forskjellig gjennomsnittlig løpshastighet i Boston Marathon. Modellen kan skrives som:

$$Yi = \beta 0 + \beta 1(genderM)i + \varepsilon i$$

Her representerer β_0 gjennomsnittet for referansegruppen (kvinner, kodet som 0), mens β_1 viser forskjellen i gjennomsnittlig hastighet mellom menn og kvinner. Dersom β_1 er positiv, betyr det at menn i gjennomsnitt har høyere hastighet enn kvinner med en verdi tilsvarende størrelsen på β_1. Resultatene fra regresjonsmodellen viste at gjennomsnittshastigheten for kvinner var = 9,58 km/t, mens menn i gjennomsnitt løp = 0,86 km/t raskere.

```
Call:
lm(formula = speed ~ gender, data = boston_dat)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.1395 -1.2005  0.0377  1.1376  8.5530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.57976    0.01876  510.67   <2e-16 ***
genderM      0.86218    0.02455   35.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.776 on 21551 degrees of freedom
Multiple R-squared:  0.05414,   Adjusted R-squared:  0.0541
F-statistic:  1234 on 1 and 21551 DF,  p-value: < 2.2e-16
```



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(p)=log(p1−p)= 0+ 1Xi

Her viser 1 hvordan log-oddsen for skade endres når den uavhengige variabelen endres med én enhet. Når vi tar den naturlige eksponenten av estimatet (e 1), får vi odds-ratio, som forteller hvor mye oddsen for skade øker eller minker.

I eksempelet under undersøkes sammenhengen mellomvoppvarmingsrutine og hamtringsfleksibilitet 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 oppvaring 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 200kg i legpress, sier modellen at vi med 95% sikkerhet kan forvente at hennes faktiske legext verdi vil være mellom lwr 35.1 kg, fit på 66.4 og upr 97.9 kg. Intervallet tar høyde for den naturlige variasjonen som vil være mellom individuelle utøvere.

Trekke slutninger

- Bruk datasettet strengtvolume 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)	89.32	3.17
2	timepre	-32.47	1.93
3	setssingle	-3.94	2.00
4	timepre:setssingle	4.39	2.72
5	sd__(Intercept)	17.68	NA
6	sd__Observation	8.19	NA

Begrunnelse

Den individuelle utgangspunktvariasjonen av leg extentions 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 tar pre og "multiple sets" gruppen som utgangspunkt i analysen. Dette vil si at "Multiple sets" gruppen hadde 89,32 enheter belastning ved pre (intercept), som var 32,47 enheter lavere ved pre enn ved post (timepre). "Single sets" gruppen hadde et utgangspunkt 3,94 enheter lavere enn "multiple sets" gruppen(setssingle). "Multiple sets" gruppen økte med 4,39 enheter mer enn "Single sets" gruppen(timepre:setssingle). Derfor kan modellen indikere at flere sett leder til større økning av styrke i populasjonen.