

Arbeidskrav 3: Regresjonsanalyse

Ketil B., Laurits H., Herman E. og Theodor S.

2025-10-30

Innhold

1 Beskrive sammenhenger	2
1.1 Sammenhengen mellom to kontinuerlige variabler (Total og kroppsvekt)	2
1.2 Sammenhengen mellom total, kjønn og alder	4
1.2.1 Tolkning av regresjonsmodellen	5
1.2.2 Graf 2: Kurvilineær graf av sammenhengen mellom total, kjønn og alder	6
1.3 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?	7
2 Predikere observasjoner	9
2.0.1 Bruk data fra datasettet strengthvolume og lag en prediksjonsmodell for legext basert på legpress.	9
2.1 Bruk data fra et tidspunkt (time) og et treningsvolum (sets)	9
2.1.1 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?	10
2.1.2 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?	11
3 Trekke sluttninger	12
3.1 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.	12
3.1.1 Hvordan kan vi bruke regresjonsmodellen for å si noe om populasjonen som dataene kommer fra?	14

1 Beskrive sammenhenger

Første del av arbeidskravet tar utgangspunkt i “Openpowerlifting” datasettet.

Begreper som brukes videre i denne analysen:

Total: *Summen av beste løftet i knebøy, benkpress og markløft sammenlagt, uttrykt i antall kilogram (eksempel: 150kg knebøy, 100 kg benkpress, 200 kg markløft = total på 450 kg)*

Maksimalstyrke: Brukes synonymt med total i denne analysen

Utstyrsfritt: Det innebærer at man kun kan bruke et enkelt lag løftebelte (10 - 13mm bredde), myke knevarmere, håndleddsstøtter og drakt uten støtte. Alt annet utstyr er ikke tillatt.

1.1 Sammenhengen mellom to kontinuerlige variabler (Total og kroppsvekt)

Sammenhengen mellom to kontinuerlige variabler kan beskrives ved hjelp av korrelasjon og lineær regresjon. Korrelasjonskoeffisienten (r) viser styrken og retningen på den sammenhengen mellom de to variablene, og varierer mellom -1 og 1. I dette tilfellet er korrelasjonen mellom kroppsvekt og total løftet mengde $r=0.369$, noe som indikerer en moderat positiv sammenheng. Løftere med høyere kroppsvekt har en tendens til å løfte mer totalt.

Pearson's product-moment correlation

```
data: sample.powerlifting$TotalKg and sample.powerlifting$BodyweightKg
t = 25.48, df = 4998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3143008 0.3633702
sample estimates:
cor
0.3390661
```

Call:

```
lm(formula = TotalKg ~ BodyweightKg, data = sample.powerlifting)
```

Residuals:

Min	1Q	Median	3Q	Max
-492.42	-178.23	5.51	156.56	542.93

Coefficients:

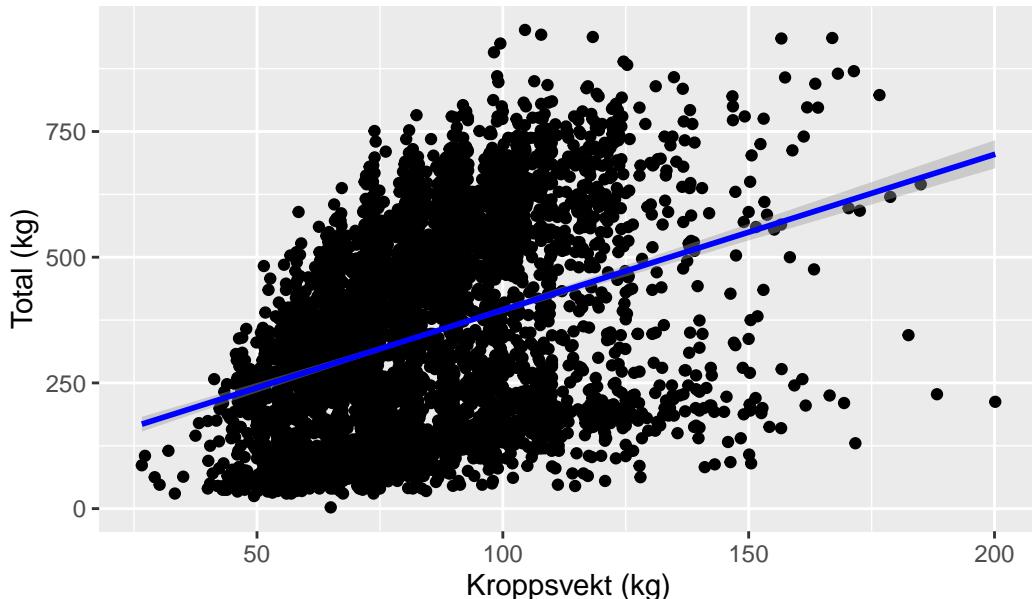
```

Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.7888    10.5539   8.129 5.43e-16 ***
BodyweightKg 3.0933     0.1214  25.480 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 186.7 on 4998 degrees of freedom
Multiple R-squared: 0.115, Adjusted R-squared: 0.1148
F-statistic: 649.2 on 1 and 4998 DF, p-value: < 2.2e-16

```

Graf 1: Forhold mellom kroppsvekt og total



I en lineær regresjonsmodell gir i tillegg et mål på hvor mye den avhengige variabelen (TotalKg) endrer seg når den uavhengige variabelen (BodyweightKg) øker.

I modellen $\text{TotalKg} = 79.93 + 3.35 \times \text{BodyweightKg}$ betyr dette at totalen i gjennomsnitt øker med omrent 3.35 kg for hver ekstra kilo kroppsvekt.

Det finnes en direkte kobling mellom korrelasjonskoeffisienten og regresjonsmodellen: $b_1 = r \times (s(y)/s(x))$, der $s(y)$ og $s(x)$ er standardavvikene til henholdsvis TotalKg og BodyweightKg. Når vi beregner denne forventede stigningstallet ut fra korrelasjonen, får vi nøyaktig samme verdi som i regresjonsmodellen ($b_{\text{expected}} = 3.35$). Dette viser at analysen er konsistent, og at korrelasjon og regresjon beskriver den samme lineære sammenhengen. Samlet viser analysen en signifikant, men moderat lineær sammenheng mellom kroppsvekt og total løfteprestasjon.

```

# Calculate correlation between bodyweight and total lift
r <- cor(sample.powerlifting$BodyweightKg, sample.powerlifting$TotalKg)

# Calculate standard deviations of both variables
sd_y <- sd(sample.powerlifting$TotalKg)           # standard deviation of Total (Y)
sd_x <- sd(sample.powerlifting$BodyweightKg)      # standard deviation of Bodyweight (X)

# Compute the expected slope (b) from the correlation
b_expected <- r * (sd_y / sd_x)

# Display the expected regression coefficient
b_expected

```

[1] 3.093326

```

mean_x <- mean(sample.powerlifting$BodyweightKg)
mean_y <- mean(sample.powerlifting$TotalKg)
b0_expected <- mean_y - b_expected * mean_x

```

1.2 Sammenhengen mellom total, kjønn og alder

Det er godt etablert at både kjønn og alder har en effekt på maksimalstyrke. Menn løfter i gjennomsnitt med kvinner, og alder kan ha både en negativ og positiv innvirkning på maksimalstyrke.

Fra et fysiologisk standpunkt kan vi anta at det vil være et kurvilineært forhold mellom total, kjønn og alder, ettersom man vil bli sterkere opp til en viss alder, før man senere vil bli svakere igjen. Vi vil også forvente at menn og kvinner vil registrere ulike totaler på bakgrunn av ulik styrke.

Siden “openpowerlifting” datasettet inneholder over tre millioner rader, må vi filtrere dataene slik at de er mer håndterbare å lage modeller og grafer med. Vi valgte dermed å filtrere et utvalg på 5000 observasjoner.

Vi testet både en lineær og en polynom regresjonsmodell. Det viste seg at den polynomiale modellen var mer passende, ettersom den lineære modellen viste en negativ korrelasjon mellom alder og total, som er fysiologisk svært usannsynlig (en 15-åring ville vært sterkere enn en 25 åring i gjennomsnitt med den lineære modellen). Ved bruk av den polynomiale modellen derimot, ser at vi totalen øker med økende alder til et visst punkt, før den igjen synker, og indikerer at det er et kurvilineært forhold.

```
#Filtrerer datasettet slik at det er mer håndterbart og gjelder en mer homogen målgruppe

powerlifting.clean <- openpowerlifting |>
  filter(
    !is.na(TotalKg), #Har en registrert total
    TotalKg > 0,
    Equipment == "Raw", #Utstyrssfritt
    !is.na(Tested), #Dopingtestet
    Sex %in% c("M", "F"), #Kun kvinner og menn
    !is.na(Dots),
    Event == "SBD"
  )

#Vi tar et tilfeldig utvalg "powerlifting.clean"

set.seed(123)
sample.powerlifting <- powerlifting.clean |>
  sample_n(5000)
```

Nå som vi har filtrert datasettet, kan vi lage en regresjonsmodell og en graf.

Merk: Med denne regresjonsmodellen gjør vi en antagelse at alder har en lik effekt på maksimalstyrke uavhengig av kjønn

1
38.18511

Variabel	Estimat	p-verdi
(Intercept)	38.185	<0.001
SexM	237.415	<0.001
Age	17.002	<0.001
I(Age^2)	-0.230	<0.001

1.2.1 Tolkning av regresjonsmodellen

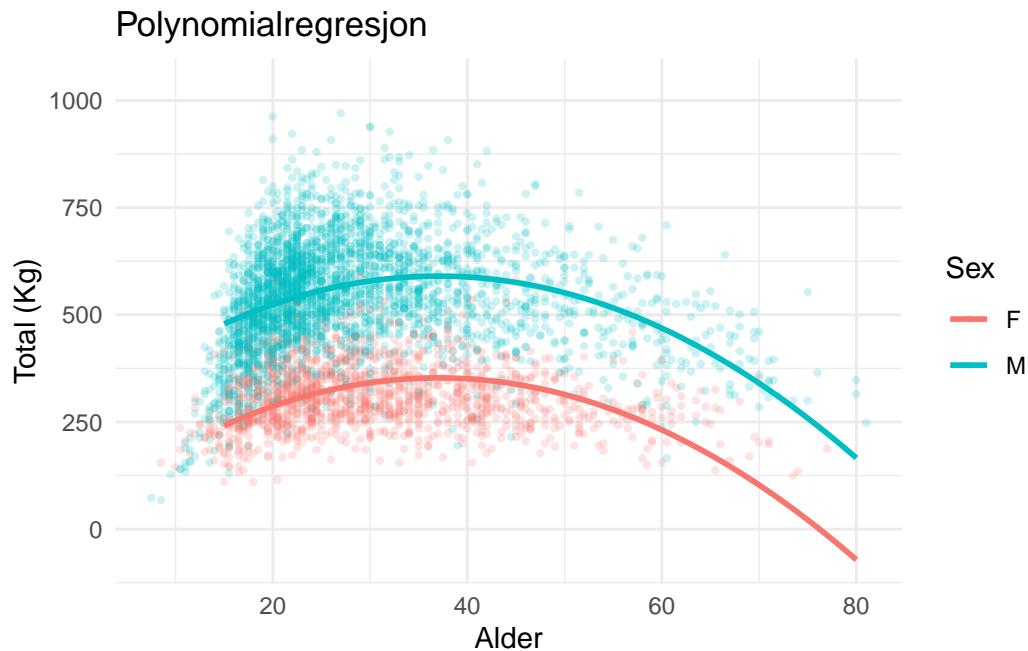
Vi ser fra modellen at skjæringspunktet er 124.5 kg, som betyr at gjennomsnittlig kvinne vil ha en total på 124.5 kg ved en alder av 0. Dette er det matematiske utgangspunktet til modellen, og i praksis ikke nyttig og unøyaktig siden datasettet ikke inneholder data på så unge utøvere. Vi ser også at menn i gjennomsnitt løfter 143.8kg mer enn kvinner, og at for hver økning i alder med antall år, vil vi få en økning på 9.6kg på totalen. Koeffisienten for Age^2 (-0.15)

viser at det er et ikke lineært forhold mellom alder og maksimalstyrke. Etter en viss alder, så vil styrken avta.

Alle koeffisientene har en p-verdi på <0.001 , som betyr at både kjønn og alder har en statistisk signifikant effekt på styrke. Modellen kan forklare omtrentlig 14% av variasjonen i total vekt løftet (Justert R-squared = 0.14), som tyder på at andre faktorer enn kjønn og alder spiller inn på maksimalstyrke.

1.2.2 Graf 2: Kurvilineær graf av sammenhengen mellom total, kjønn og alder

Graf 2: Figuren viser hvordan totalen varierer med alder for menn og kvinner, og når en topp rundt en alder av 30 -35 år, før den gradvis avtar.



Rettelser etter tilbakemeldingen: Den opprinnelig grafen viste ikke det samme som modellen vår. Modellen vår gjør en antagelse om at alder har samme effekt på total for begge kjønn, som tilsier at de to kurvene burde være parallele med ulike skjæringspunkt. I den opprinnelig grafen, så vi at linjene til menn og kvinner begynte å konvergere. Feilen var å bruke 'formula = $y \sim x + I(x^2)$ ', som fører til at vi får ulike koeffisienter for variabelene våre for menn og kvinner.

For å rette opp i dette, har vi lagd en ny dataramme med modellen vår, m, og produsert predikerte verdier for ulike totaler. Deretter har vi plottet dette, og fått en graf der vi får to parallele kurver for menn og kvinner.

Når det gjelder de to “clusterene” per kjønn, har vi oppdaget en kritisk mangel i filteringsprosessen av datasettet. Openpowerlifting-datasettet har en “event”-kolonne, som inkluderer komplette styrkeløftstevner (‘SBD’), men også markløft, (‘D’) og benkpress (‘B’) ekslusiveste stevner. Et komplett styrkeløftstevne vil ha 2 til 3 ganger høyere verdier sammenlignet med enkeltløftstevner, og gjør dermed den opprinnelig modellen vår ugyldig. Vi har dermed endret filtreringen til kun “SBD”, og dermed forsvinner “clusterene” i grafen vår.

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

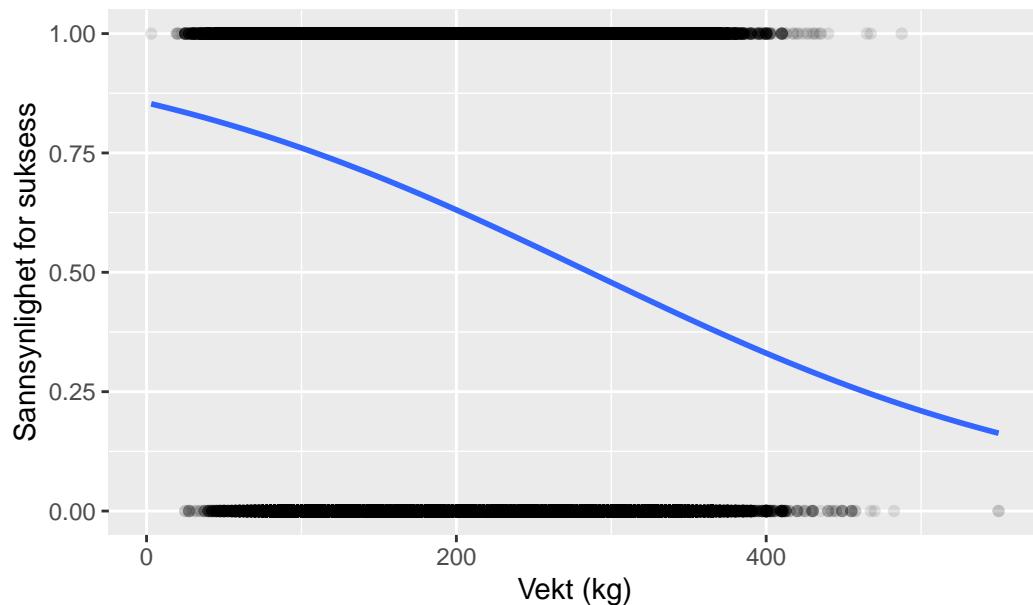
En generalisert lineær modell (GLM) er en type modell som lar oss jobbe med data som ikke er normalfordelt. I vårt tilfelle, så kan vi bruke en logistisk regresjonsmodell som er en underkategori av GLM, for å se på log-oddsen for et binomialt utfall, som godkjent og ikke-godkjent løft.

Vi tar utgangspunkt i om løfteren har fått det tredje løftet sitt i markløft godkjent eller ikke. Variabelen ”Deadlift3Kg” i datasettet blir brukt for å lage vår binære utfallsvariabel ”Deadlift3_success”. I openpowerlifting-datasettet, så har et godkjent løft en positiv verdi (eks: 120), og et ikke-godkjent løft en negativ verdi (eks: -120). Vi kan dermed definere utfallsvariablen som 1 for godkjent løft hvis variablen ”Deadlift3Kg” > 0, og som 0 for ikke-godkjent løft hvis ”Deadlift3Kg” < 0. Hvis verdien til variablen er ”NA”, betyr det at løftet ikke ble forsøkt, og disse verdiene blir ekskludert fra modellen.

I modellen vår blir den uavhengige variablene absolutt vekt forsøkt i markløft (Deadlift3attempt). Vi kan anta at med økende vekt, så vil sannsynligheten for å få et godkjent løft synke. Vi vil dermed forvente en negativ koeffisient.

term	estimate	std.error	statistic	p.value
(Intercept)	1.776	0.009	200.735	0
Deadlift3attempt	-0.006	0.000	-148.520	0

Graf 3: Sannsynlighet for suksess



Vi får følgende estimatorer i modellen vår, med skjæringspunkt = 1,7 og et stigningstall = -0.006. Skjæringspunktet forteller oss log-oddset for å få et godkjent løft når den uavhengige variablene er 0, altså når man løfter 0 kg. Dette er i praksis absurd, og ikke nyttig. Når vi ser på stigningstallet, så ser vi at den har en negativ verdi. Det betyr at for hver økning i x (antall kg), så vil log-oddset reduseres med 0.006. På odds-skalaen vil det bety at oddsen multipliseres med $\exp(-0.006) = 0.994$. Med andre ord, vil oddsen for å lykkes med tredje forsøket i markløft reduseres med ca. 0.6% per kg man øker vekten med.

Merk: Modellen predikrer kun basert på vekt forsøkt, og tar ikke hensyn til andre faktorer som treningserfaring, muskelmasse, teknikk og så videre.

I den logistiske regresjonsmodellen, så ønsker vi å se på sannsynligheten for et utfall ved hjelp av en lineær regresjonsmodell. Problemet med dette, er at sannsynligheten må alltid ha en verdi mellom 0 og 1, og en lineær regresjon kan gi oss verdier langt utenfor dette spennet. "Link-function" fungerer som en bro, ved at den transformerer sannsynligheter om til log-odds, som kan være verdier utenfor 0 til 1. Dermed kan vi fortsatt bruke den kraftige lineære regresjonsmodellen selv på binære utfall, ved hjelp av denne funksjonen.

2 Predikere observasjoner

2.0.1 Bruk data fra datasettet strengthvolume og lag en prediksjonsmodell for legext basert på legpress.

```
#Vi må først filtrere datasettet
Legext_modell <- strengthvolume |>
  filter(exercise %in% c("legpress", "legext")) |>
  pivot_wider(names_from = exercise,
  values_from = load)

#Prediksjonsmodell for legext basert på legpress
ext.m1 <- lm(legext ~ legpress, data = Legext_modell)
```

2.1 Bruk data fra et tidspunkt (time) og et treningsvolum (sets)

```
Legext_modell |>
  filter(
    time == "session1",
    sets == "single"
  )

# A tibble: 39 x 8
  participant sex   include time     sets   leg   legpress legext
  <chr>       <chr>  <chr>   <chr>   <chr>  <chr>  <dbl>   <dbl>
1 FP13        male   incl    session1 single  R      125    55
2 FP1          male   excl    session1 single  R      175    65
3 FP16         female incl    session1 single  R      95     35
4 FP17         male   incl    session1 single  L      295    75
5 FP12         female incl    session1 single  L      225    75
6 FP11         male   incl    session1 single  L      180    65
7 FP9          male   incl    session1 single  L      165    80
8 FP14         female incl    session1 single  L      87.5   50
9 FP4          female excl    session1 single  L      170    45
10 FP6         female incl   session1 single  R      110    40
# i 29 more rows
```

2.1.1 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?

Vi trenger to modeller for å svare på dette spørsmålet. Den første modellen, "ext.m2", har ingen interaksjonseffekt mellom kjønn og legpress på legeextension, mens "ext.m3" har en interaksjonseffekt mellom kjønn og legpress på legeextension. Vi tar utgangspunkt i en legpress på 150 kg, og predikerer legeextension verdier for både menn og kvinner i de tre modellene.

```
#Modell med kjønn som variabel, uten interaksjonseffekt på legpress
ext.m2 <- lm(legext ~ legpress + sex, data = Legext_modell)

#Modell med kjønn som variabel, med interaksjonseffekt på legpress
ext.m3 <- lm(legext ~ legpress + sex + sex:legpress, data = Legext_modell)

#Prediksjoner innad kjønn, vi bruker ext.m2 modellen. Utgangspunkt i en legpress = 150 kg

pred.m2female <- predict(ext.m2, newdata = data.frame(legpress = 150,
                                                       sex = "female"))
pred.m2male <- predict(ext.m2, newdata = data.frame(legpress = 150,
                                                       sex = "male"))

#Prediksjoner innad kjønn med ext.m3 modellen.

pred.m3female <- predict(ext.m3, newdata = data.frame(legpress = 150,
                                                       sex = "female"))
pred.m3male <- predict(ext.m3, newdata = data.frame(legpress = 150,
                                                       sex = "male"))

#Prediksjoner i gjennomsnitt i begge kjønn. Her bruker vi ext.m1 som gir oss en gjennomsnitt

pred.m1average <- predict(ext.m1, newdata = data.frame(legpress = 150))

Prediksjoner2 <- data.frame(
  Model = c("m1 (average)", "m2 (female)", "m2 (male)", "m3 (female)", "m3 (male)"),
  Prediction = c(pred.m1average, pred.m2female, pred.m2male, pred.m3female, pred.m3male)
)

knitr::kable(Prediksjoner2, digits = 2, caption = "Prediksjoner fra de ulike modellene ved 150 kg legpress")
```

term	df.residual	rss	df	sumsq	statistic	p.value
legext ~ legpress + sex	387	57862.69	NA	NA	NA	NA
legext ~ legpress + sex + sex:legpress	386	57508.60	1	354.0907	2.376671	0.1239792

Table 3: Prediksjoner fra de ulike modellene ved legpress = 150

Model	Prediction
m1 (average)	56.21
m2 (female)	51.12
m2 (male)	67.54
m3 (female)	50.48
m3 (male)	68.64

```
ANOVA <- anova(ext.m2, ext.m3)
tidy(ANOVA) |>
  gt()
```

I tabell 3 ser vi at m2 og m3 viser tydelige forskjeller i styrke i legextension mellom menn og kvinner. Dette er forventet, ettersom menn er i gjennomsnitt sterkere enn kvinner. Likevel ser det ut til at m3 og m2 gir oss nesten identiske verdier. Vi utførte dermed en ANOVA-test mellom modell ext.m2 og ext.m3, som viste at det ikke var statistisk signifikant forskjell mellom modellene ($F = 2.38$, $p = 0.124$). Dette tyder på at effekten av legpress på legextension er tilnærmet lik mellom kjønnene, og dermed er den enklere modellen foretrukket basert på prinsippet om parsimoni (Okasha, 2016, s. 25 - 26). I modell m1, som ikke inkluderer kjønn som prediktor, får vi en prediksionsverdi for legextension basert på legpress, uavhengig av kjønn.

2.1.2 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?

Table 4: 95% prediksionsintervall for legextension ved legpress = 150 kg

Sex	Prediction	Lower	Upper
Female	51.12	27.01	75.23
Male	67.54	43.37	91.71

Prediksionsintervallet vårt forteller oss hvilke verdier vi kan forvente for legextension når en deltaker klarer 150 kg i legpress, både for kvinnelig og mannlige kjønn.

Hvis en kvinne tar 150kg i legpress, forventer vi at nye observasjoner forkvinne har en 95% sannsynlighet for å få en legextension verdi mellom 27 og 75kg. Modellen viser oss en gjennomsnittlig prediksjon på 51kg. Hvis en mann tar 150 kg i legpress, forventer vi at nye observasjoner for menn har en 95% sannsynlighet for å få en legextension verdi mellom 43 og 91 kg. Modellen viser oss en gjennomsnittlig prediksjon på 68 kg.

Prediksionsintervaller viser oss at det er et stort spenn av mulige verdier for legextension når legpress er 150kg, for begge kjønn. Dette kan tyde på store individuelle forskjeller, selv ved samme legpress verdier.

3 Trekke sluttninger

3.1 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.

Følgende del av oppgaven løses i JASP.

Vi begynte å redigere dataen og gi tallverdier. Først inkluderte vi kun "pre" og "post" i "time", samt ga de en tallverdi 1 – 2. i "exercise" fjernet vi alle øvelser unntatt legeext. Videre gikk vi inn på linear mixed model, satte "load" som dependent variable, "sets" og "time" i fixed effects variables, og "participants" i random effects grouping factors. Deretter satte vi inn "sets" og "time", og "sets x time" inn som ulike model components I random components valgte vi å se på intercept . I options huket vi av model summary og fixed effects estimates. I plots for å få selve modellen satte vi inn «time» i den horisontale aksen, «sets» i separate lines for å skille mellom pre og post, og «participants» inn i background data show. I tillegg førte vi inn i plots for å illustrere med en graf.

Fixed Effects Estimates

Term	Estimate	SE	df	t	p
Intercept	56.859	3.120	49.45	18.224	< .001
sets (single)	0.449	1.854	103.96	0.242	.809
time (post)	32.465	1.929	104.47	16.829	< .001
sets (single) * time (post)	-4.385	2.724	104.03	-1.610	.110

Note. The intercept corresponds to the mean in the default factor category (factor level parameters are estimated with treatment contrast coding). Consequently, the estimates cannot be easily interpreted in the presence of interactions. Use estimated marginal means for obtaining estimates for each factor level/design cell or their differences.

Figure 1: Mixed effect modellen

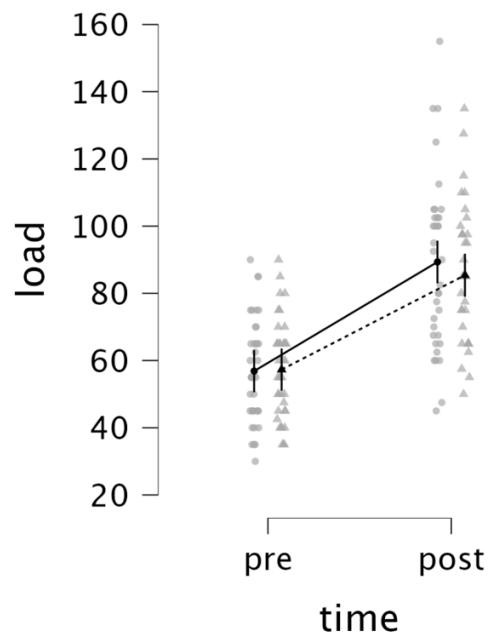


Figure 2: Linjediagram av mixed-effect modellen. Endring i belastning fra pre- til post. Stiplet linje er single sett, og heltrukket linje er multiple sets

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

Vi kan tolke de ulike variabelene våre i regresjonsmodellen som følger. Intercept kan tolkes som den gjennomsnittlig verdien til multiplesett gruppen ved pre-test (56.9, $p < 0.001$). Dette kan med andre ord, også tolkes som modellens beste estimat av den gjennomsnittlige verdien til populasjonen ved baseline. Forskjellen mellom single og multiple sets på baseline er ubetydelig som observert i `sets:single` ($p=0.809$). `time:post` viser oss endringen fra post og pre for multiple sets gruppen (32.5, $p < 0.001$). Den siste variablen `sets:single*time:post` viser oss gjennomsnittlig økning i single-sett sammenlignet med multiple-sett, med en verdi på -4.4 ($p = 0.110$). Dette antyder at single-sett gruppen fikk i gjennomsnitt en økning som var 4.4kg mindre enn multiple sett gruppen, men denne effekten er ikke statistisk signifikant og vi kan dermed ikke etablere om det var en reell forskjell i økning mellom gruppene.