Assignment 2: Regression models, predicting from data

Introduksjon / Bakgrunn

Denne oppgaven er delt inn tre separate deler som tar for seg konsepter innenfor analyse av data og regresjon. I del 1 kalkulerer vi laktat terskler, og ser nærmere på reliabiliteten mellom to ulike terskelnivåer. Del 2 bruker vi molekylær data til å predikere størrelsen på DNA-fragment ved hjelp av en veileder. I del 3 skal vi se nærmere på om det finnes en lineær sammenheng mellom to valgte variabler fra datasettet `hypertrophy`i datapakken `exscidata`.

Del 1: Laktat terskler

Introduksjon

Laktat terskel er en variabel som er godt brukt for å forutsi prestasjon innenfor utholdenhets idretter, til å styre intensiteten av treningsøkter og evaluere trenings effekt (Machado et al., 2012). Det finnes ulike metoder for å finne testpersonens laktat terskel. Machado et al. (2012) forteller oss at den "maximal-deviation method" (Dmax) anbefalt av Cheng et al. 1992, bidrar med å kunne evaluere de ulike mekanismene som virker bestemmende for prestasjon innenfor langdistanseløping og sykling (Machado et al., 2012). Videre hadde denne metoden en bedre korrelasjon med prestasjon og laktat terskel sammenliknet med andre metoder. I våres reliabilitets tester ble det ikke utført laktat målinger, på bakgrunn av dette benytter vi oss av data settet til "cyclingstudy". De representative tersklene som blir undersøkt er 2 mmol L-1 og 4 mmol L-1.

Metode

Som en kan se i den plotta grafen under, er de forskjellige grafene ikke så forskjellige rundt 2mmol og 4mmol L~-1. På den andre siden ser vi at den lineære modellen er feil ved 300w, den sekundærplynomiske modellen er feil ved 275w. Den tredje- og fjerdeplynomiske modellen derimot, varierer ikke mye fra hverandre.

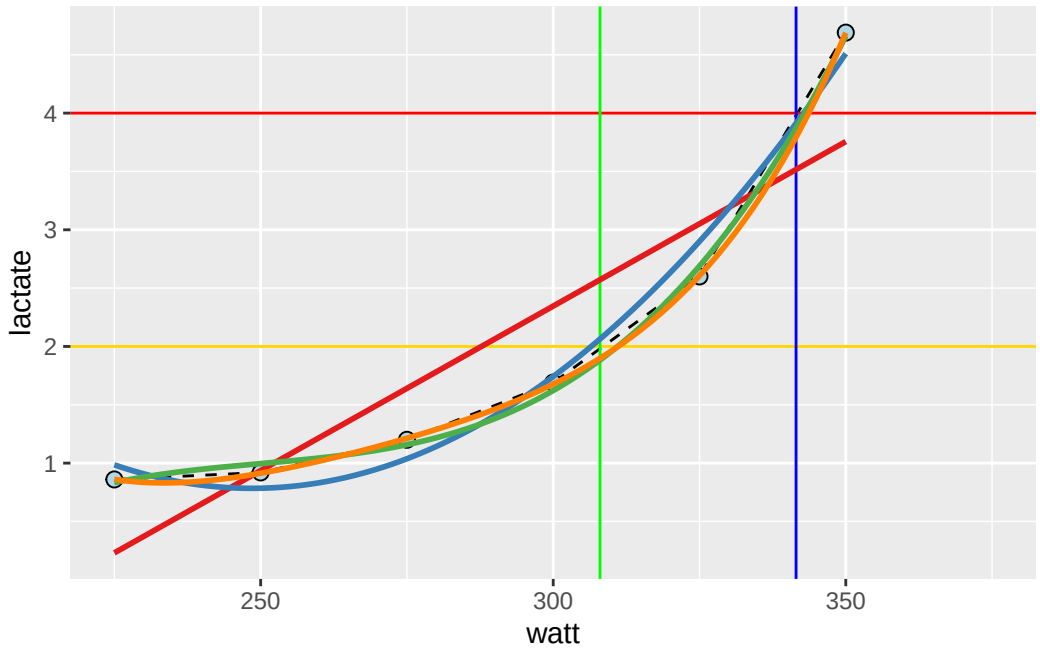
```r
### laste ned nødvendige packages
library(tidyr)
library(tidyverse)
library(ggplot2)
library(exscidata)


###laste inn data
data("cyclingstudy")


### Estimering av laktatterskelen og treningsintensiteten ved 4mmol L-1

cyclingstudy %>%
  # utvalg av nødvendige kolonner i analysen.
  select(subject, group, timepoint, lac.225:lac.375) %>%
  # Kun ein deltaker og ett tidspunkt.
  filter(timepoint == "pre", subject == 10) %>%
  # lang format ved å bruke laktatkolonnene.
  pivot_longer(names_to = "watt",
               values_to = "lactate",
               names_prefix = "lac.",
               names_transform = list(watt = as.numeric),
               cols = lac.225:lac.375) %>%
  # Plotte data, group = subject nødvendig for å sammenkoble punktene.
  ggplot(aes(watt, lactate, group = subject)) +
  geom_line(lty = 2) +
  geom_point(shape = 21, fill = "lightblue", size = 2.5) +
  # Linjer på spesifikke punktene for 2mmol og 4mmol, samt skjeringspunktet mellom linjene.
  geom_hline(yintercept = 4, color = "red") +
  geom_hline(yintercept = 2, color = "gold") +
  geom_vline(xintercept = 341.5, color = "blue") +
  geom_vline(xintercept = 308, color = "green") +
  # legge til en strak linje fra den lineære modelen.
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, color = "#e41a1c") +

  # poly(x, 2) Legger til en andregradsplynomisk modell.
  geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 2), color = "#377eb8") +
  # poly(x, 3) Legger til en tredjegradsplynomisk modell.
  geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 3), color = "#4daf4a") +
  # poly(x, 4) Legger til en fjerdegradsplynomisk modell.
  geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 4), color = "#ff7f00")
```




```
### vurdering av tilpasningen til de forskjellige lineære modellene på sammenhengen mellom t

lactate <- cyclingstudy %>%
  # utvalg av nødvendige kolonner i analysen.
  select(subject, group, timepoint, lac.225:lac.375) %>%
  # Kun ein deltaker og ett tidspunkt.
  filter(timepoint == "pre", subject == 10) %>%
  # lang format ved å bruke laktatkolonnene.
  pivot_longer(names_to = "watt",
               values_to = "lactate",
               names_prefix = "lac.",
               names_transform = list(watt = as.numeric),
               cols = lac.225:lac.375) %>%
  # Fjerne dei ugyldige veriene NA for å hindre feilmeldinger.
  filter(!is.na(lactate))

# Legger til en strak linje fra modelen.
m1 <- lm(lactate ~ watt, data = lactate)

# Legger til en andregradsplynomisk modell.
m2 <- lm(lactate ~ poly(watt, 2, raw = TRUE), data = lactate)

# Legger til en tredjegradsplynomisk modell.
```

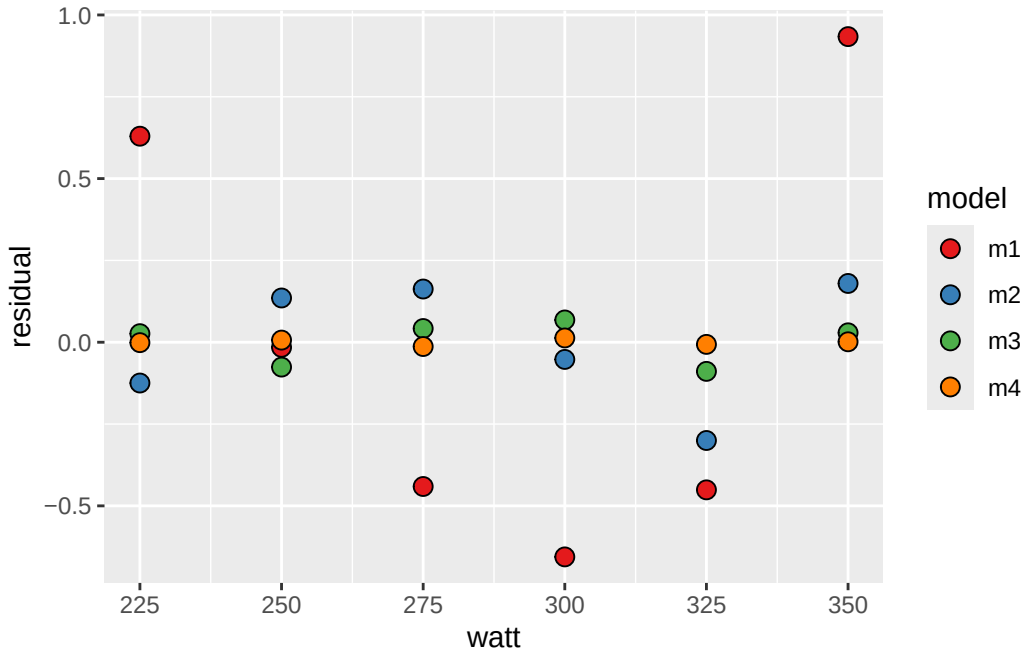
```r
m3 <- lm(lactate ~ poly(watt, 3, raw = TRUE), data = lactate)

# Legger til en fjerdegradsplynomisk modell.
m4 <- lm(lactate ~ poly(watt, 4, raw = TRUE), data = lactate)

# Lagre alle restverdiene som nye variabler.
lactate$resid.m1 <- resid(m1)
lactate$resid.m2 <- resid(m2)
lactate$resid.m3 <- resid(m3)
lactate$resid.m4 <- resid(m4)

lactate %>%
  # Samle all data fra modelleme.
  pivot_longer(names_to = "model",
               values_to = "residual",
               names_prefix = "resid.",
               names_transform = list(residual = as.numeric),
               cols = resid.m1:resid.m4) %>%
  # Plotte verdiene fra den observerte watten på x aksen og restverdiene på y aksen
  ggplot(aes(watt, residual, fill = model)) + geom_point(shape = 21, size = 3) +

  # For å ha samme farger som over, bruker me scale fill manual.
  scale_fill_manual(values = c("#e41a1c", "#377eb8", "#4daf4a", "#ff7f00"))
```



For å finne ut hva forutsatt wattverdi som er nærmest 2 og 4 mmol L-1, benytter vi koden under:

```
# Ny dataramme
ndf <- data.frame(watt = seq(from = 225, to = 350, by = 0.1))

ndf$predictions <- predict(m3, newdata = ndf)


# for å finne ut kva forutsatt Wattverdi som er nermest 2 og 4 mmol L-1
lactate_threshold <- ndf %>%
  filter(abs(predictions - 4) == min(abs(predictions - 4)))

summary(lactate_threshold)
```

```
      watt      predictions
 Min.   :343   Min.   :4
 1st Qu.:343   1st Qu.:4
 Median :343   Median :4
 Mean   :343   Mean   :4
 3rd Qu.:343   3rd Qu.:4
 Max.   :343   Max.   :4
```