StatMet eksamensopgave

Morten Tulstrup (gbz480), Gabriel Boeskov, Christian Adrian Jensen

2023-01-16

Om opgavefordeling

Vi har alle bidraget ligeligt til alle opgaver.

Opgave 1

Vi tager udgangspunkt i Theorem 2.1.5 (koncentrationsulighed baseret på Chebychevs ulighed):

$$P\left(|\hat{p}-p| > \delta\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{\delta^2}$$

Indsætter $p = 0.2$ og $n = 80$:

$$P\left(|\hat{p}-0.2| > \delta\sqrt{\frac{0.2(1-0.2)}{80}}\right) = P\left(|\hat{p}-0.2| > \delta\sqrt{0.002}\right) \leq \frac{1}{\delta^2}$$

Vi vælger δ så

$$\delta\sqrt{0.002} = 0.1 \Rightarrow \delta = \frac{0.1}{\sqrt{0.002}} = 2.236$$

Dermed er:

$$P\Big(|\hat{p}-0.2| > 0.1\Big) \leq \frac{1}{2.236^2} = 0.2 \Rightarrow$$
$$P(\hat{p} \in [0.1, 0.3]) \leq 0.2$$

Eftersom $0 \leq \hat{p} \leq 1$ og P er et sandsynlighedsmål er det ensbetydende med:

$$P(\hat{p} \in [0.1, 0.3]^C) = P(\hat{p} \in [0, 0.1) \cup (0.3, 1]) = P(\hat{p} \in [0.1)) + P(\hat{p} \in (0.3, 1])) > 0.8$$

Vi kan altså bruge koncentrationsuligheden til at sige, at sandsynligheden for at observere bivirkningen hos 10-30% af de vaccinerede er større end 0.8. Vi kan derimod ikke bruge koncentrationsuligheden til at sige, hvordan sandsynligheden fordeler sig mellem tilfældende hvor $\hat{p} < 0.1$ vs. $\hat{p} > 0.3$, så den egentlige sandsynlighed for at observere $\hat{p} > 0.1$ er altså større end 0.8.

Den faktiske sandsynlighed ud fra binomialfordelingen er 0.987:

```
1-pbinom(8, 80, 0.2)
```

```
## [1] 0.9869125
```

Som er et meget mere præcist tal, dvs.at koncentrationsuligheden er korrekt men meget upræcis.

Opgave 2

Opgave 3

Opgave 4

Opgave 5

Opgave 6

Vi betragter den lineære model for maxLA $\text{maxLA}_i = \mu_1(\text{race}_i) + \epsilon_i$ (race_i er racen for hund i, $i = 1, ..., n$), hvor $\mu_1(\text{race}) = \beta_0 + \beta_1 \cdot 1(\text{race} = \text{Petit Basset}) + \beta_2 \cdot 1(\text{race} = \text{Whippet})$ og det tilhørende underrum $L_1 \subseteq \mathbb{R}^n$ med $\dim(L_1) = 3$. Dvs. modellens designmatrix $\mathbf{X}$, som frembringer L_1 er en $n \times 3$ matrix med søjler:

$$\begin{pmatrix} \mathbf{1} & 1(\text{race}_i = \text{Petit Basset}) & 1(\text{race}_i = \text{Whippet}) \end{pmatrix}$$

Hvor $\mathbf{1} \in \mathbb{R}^n$ er en vektor med 1 i alle n indgange de to øvrige kolonner er dummyvariable for racerne Petit Basset og Whippet.

Vi betragter også $L_0 = \text{span}(\mathbf{1})$, og vi bemærker at $L_0 \subseteq L_1$ (fordi 1-vektoren indeholdt i begge modelmatricer).

For at teste om der er forskel på venstre forkammers volumen i de tre hunderacer tester vi, med udgangspunkt i modellen for L_1, nulhypotesen:

$$H_0 : \mu \in L_0$$

Dvs. vi foretager et F-test. Konkret gøres det ved resampling efter fremgangsmåden i eksemplet i filen RprogF62.Rmd. Først opretter vi en ny tibble i R, som kun indeholder de tre relevante hunderacer, og fitter en lineær model svarende til nulhypotesen ("nulmodellen") samt konstruerer den tilhørende modelmatrix og bestemmer n (antal rækker i datasættet):

```
hunde <- read.table("../hunde.txt", sep = "\t", header = T)
trehunde <- hunde %>% filter(race %in% c("Border_Terrier", "Petit_Basset", "Whippet"))
lm_0 <- lm(maxLA ~ 1, data = trehunde)
X0 <- model.matrix(lm_0)
n <- nrow(X0)
```

Så foretager vi resampling, dvs. vi genererer 10000 datasæt med hver n = 61 resamplede værdier af maxLA, hvor:

$$\text{maxLA}_i^* = \hat{\text{maxLA}}_i + \epsilon_i^*$$

Hvor $\hat{\text{maxLA}}_i$ er de faktiske fittede værdier i L_0, og ϵ_i^* er resamplede værdier fra den empiriske fordeling af residualer i nulmodellen, og hvor resamplingen er foretaget med tilbagelægning:

```
set.seed(2022)
B <- 10000
my_boot <- tibble(residuals = residuals(lm_0)) %>%
  rep_sample_n(size = n, replace = TRUE, reps = B) %>%
  mutate(y = fitted(lm_0) + residuals)
```

Efterfølgende fitter vi modellen svarende til L_1, og opstiller dens modelmatrix:

```
lm_fit <- lm(maxLA ~ race, data = trehunde)
X <- model.matrix(lm_fit)
```

For hvert af de 10000 resamplede datasæt foretager vi nu et F-test, som tester nulhypotesen $H_0 : \mu \in L_0$ under den større model L_1, hvor μ nu er den resamplede middelværdivektor. F teststørrelsen udregnes som defineret i NRHAT Theorem 4.3.15:

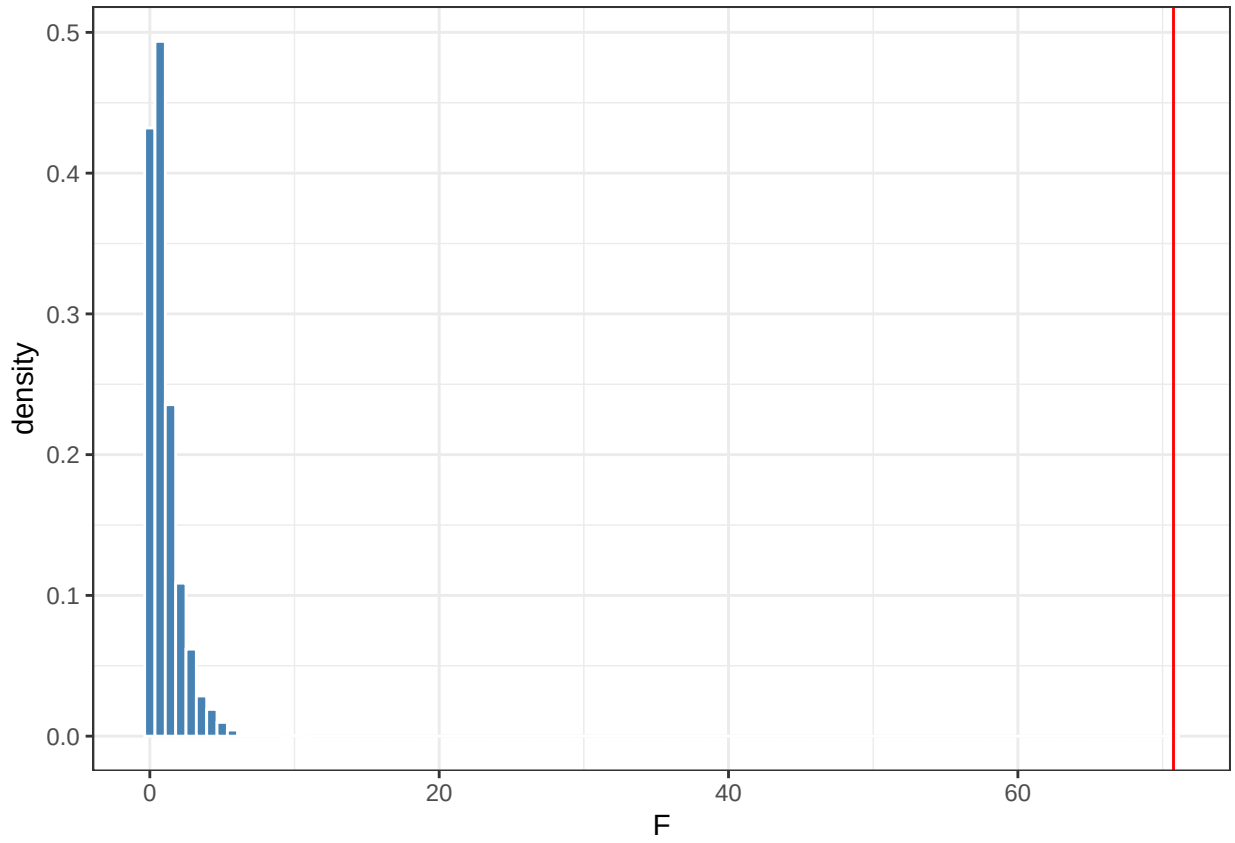
```
F_test <- function(lm_null, lm_full) {
  p <- lm_full$rank
  q <- lm_null$rank
  lm_full$df.residual * sum((lm_full$fitted.values - lm_null$fitted.values)^2)/
    (sum(lm_full$residuals^2) * (p - q))
}

my_res <- my_boot %>%
  summarize(F = F_test(lm.fit(X0, y), lm.fit(X, y)))
```

Den observerede F-teststørrelse udregnes på samme vis og P-værdien bestemmes:

```
F_obs <- F_test(lm.fit(X0, trehunde$maxLA), lm.fit(X, trehunde$maxLA))
p_value <- sum(my_res$F > F_obs) / B
p_value
```

```
## [1] 0
```

I dette tilfælde er $P = 0$, fordi der i de 10000 resamplede datasæt ikke forekom en eneste F-teststørrelse som var større end den observerede. Dette illustreres også i følgende figur, som viser bootstrapfordelingen for F-teststørrelsen (blå søjler) og den observerede F-teststørrelse (rød linje):



Vi forkaster altså nulhypotesen og konkluderer, at venstre forkammers volumen de tre hunderacer ikke er ens mellem de tre hunderacer.

Opgave 7:

Vi fokuserer på hunderacen Whippet.

Samplingfordelingen for $\hat{m_1}$ bestemmes ved ikke-parametrisk bootstrap::

```
whip <- hunde %>% filter(race == "Whippet")
set.seed(123)
m1_bootstrap <- whip %>%
      select(maxLA) %>%
      rep_sample_n(size = 50,
                   replace = TRUE,
                   reps = 10000) %>%
      summarise(stat = median(maxLA)) %>%
      mutate(est = "m1hat")
```

Samplingfordelingen for $\hat{m_2}$ bestemmes vha. parametrisk bootstrap, baseret på antagelsen om at $\log(Y_i)$ følger en normalfordeling med parametre μ og σ^2. Dvs. at vi først bestemmer $\hat{\mu}$ og $\hat{\sigma^2}$ for $\log(Y_i)$ ved: .

```
muhat <- mean(log(whip$maxLA))
sigmahat <- var(log(whip$maxLA))
```

Og dernæst resampler 10000 gange 17 realisationer fra en normalfordeling med parametre $\mu = \hat{\mu} = 10.54$ og $\sigma^2 = \hat{\sigma}^2 = 4.41$:

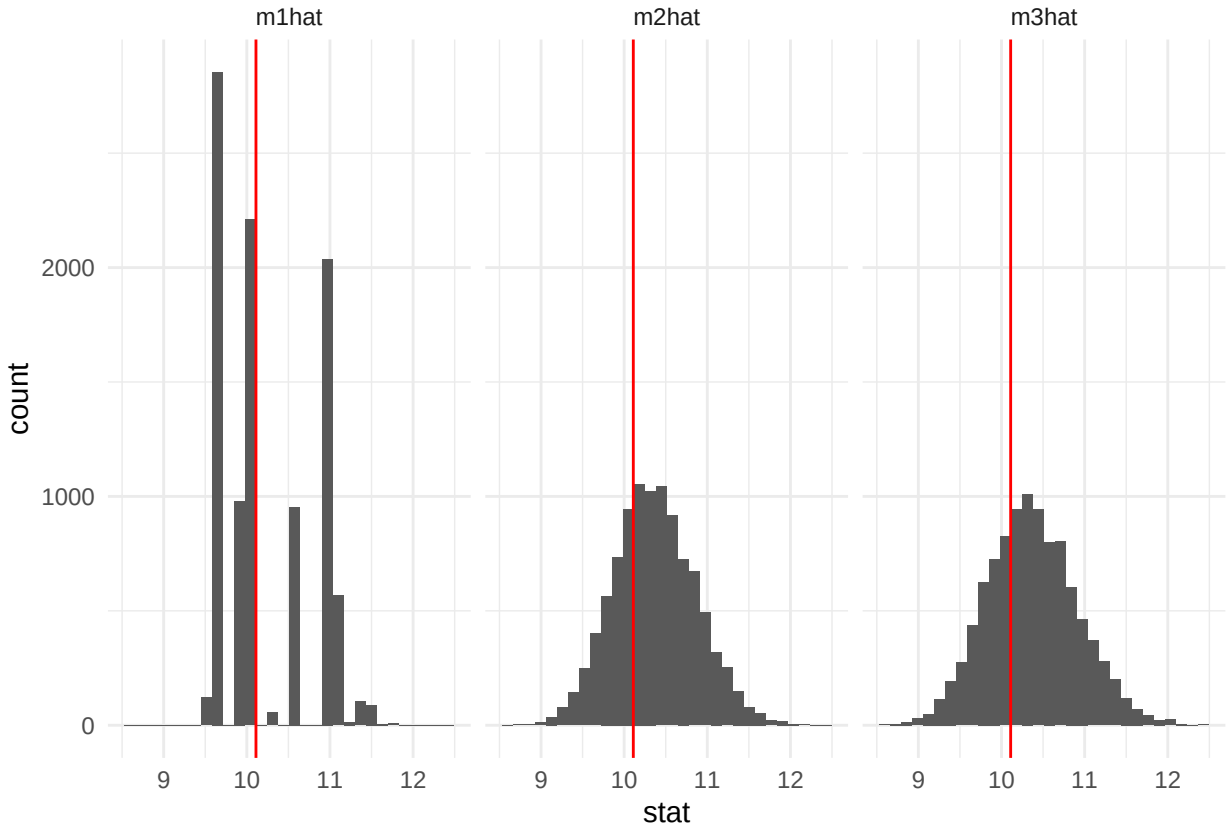
```
n <- nrow(whip)
B <- 10000
set.seed(123)
m2_bootstrap <- tibble(replicate = rep(1:B, each = n),
                       logyi = rnorm(B*n, mean = muhat, sd = sqrt(sigmahat))) %>%
  group_by(replicate) %>%
  summarise(stat = exp(mean(logyi))) %>%
  mutate(est = "m2hat")
```

$\hat{m}_3$ bestemmes også ved parametrisk boostrap, dvs. vi resampler fra den samme fordeling som for $\hat{m}_2$:

```
set.seed(123)
m3_bootstrap <- tibble(replicate = rep(1:B, each = n),
                       logyi = rnorm(B*n, mean = muhat, sd = sqrt(sigmahat))) %>%
  group_by(replicate) %>%
  summarise(stat = exp(mean(quantile(x = logyi, probs = c(0.25, 0.75))))) %>%
  mutate(est = "m3hat")
```

De tre samplingfordelinger visualiseres og sammenlignes med den observerede, empiriske median i datasættet:

```
d7 <- bind_rows(m1_bootstrap, m2_bootstrap, m3_bootstrap)
d7 %>% ggplot(aes(x = stat)) +
  geom_histogram(bins = 30) +
  facet_wrap(~factor(est)) +
  geom_vline(xintercept = median(whip$maxLA), col = "red") +
  theme_minimal()
```



En god estimator skal helst være både precise, accurate og unbiased. For at være "precise", skal alle estimaterne være samlet tæt om det samme punkt. En måde at vurdere dette på er den observerede varians af samplingfordelingen for hver estimator:

```
d7 %>% group_by(est) %>%
  summarise(var(stat))
```

```
## # A tibble: 3 x 2
##   est   `var(stat)`
## * <chr>       <dbl>
## 1 m1hat       0.322
## 2 m2hat       0.246
## 3 m3hat       0.296
```

$\hat{m_2}$ er altså den mest precise, mens $\hat{m_1}$ er mindst precise, hvilket også stemmer overens med samplingfordelingernes udseende i histogrammerne ovenfor. Visuelt ser især $\hat{m_1}$ ud til at være meget unprecise. Accuracy og bias er i sagens natur svære at vurdere her, da vi ikke kender den sande median for fordelingen af venstre forkammer-volumen i Whippet-hunde. Hvis vi for øvelsens skyld antog at den observerede median i vores (meget lille) datasæt var den sande median, kan vi se, at både $\hat{m_2}$ og $\hat{m_3}$ rammer en lille smule ved siden af, dvs. de er en anelse inaccurate og biased. I sidste ende kommer valget af estimator primært an på, hvilke antagelser vi tør gøre om Whippet-hundes hjerter: Hvis vi er overbeviste om, at antagelsen $\log(Y_i)$ er normalfordelt er en god antagelse, så er $\hat{m_2}$ formentlig den bedste estimator at bruge, da den er den mest precise, og teorien fortæller os, at medianen er lig gennemsnittet for en normalfordelt variabel.

Opgave 8:

Vi bemærker først at wgt > 0 og maxLA > 0 for alle hunde i datasættet (se tabel i opgave 4), og at vi derfor godt kan bruge logaritmen på de to variable.

Vi betragter en additive noise model for log(wgt) på formen:

$$\log(\text{maxLA}) = \mu(\log(\text{weight}), \text{race}) + \epsilon_i$$

Hvor race indgår som faktor-variabel og log(weight) indgår som numerisk variabel, dvs. middelværdifunktionen μ er en lineær funktion af den kontinuerte variabel log(wgt) og faktorvariablen race givet ved:

$$\mu(\log(\text{weight}), \text{race}) = \beta_0 + \beta_1 \cdot \log(\text{weight}) + \beta_2 \cdot 1(\text{race} = \text{Grand Danois}) +$$

$$\beta_3 \cdot 1(\text{race} = \text{Labrador}) + \beta_4 \cdot 1(\text{race} = \text{Petit Basset}) + beta_5 \cdot 1(\text{race} = \text{Whippet})$$

Dvs. dimensionen af det lineære underrum $L = L_{\log(\text{weight})} + L_{\text{race}}$ bliver (Jvf NRHAT lemma 3.2.14):

$$\dim(L) = \dim(L_{\log(\text{weight})}) + \dim(L_{\text{race}}) - \dim(L_{\log(\text{weight})} \cap L_{\text{race}}) =$$
$$\dim(L_{\log(\text{weight})}) + \dim(L_{\text{race}}) - \dim(L_1) = 2 + 5 - 1 = 6$$

Vores modelmatrix $\mathbf{X}$ ser dermed således ud (første seks rækker):

```
head(model.matrix(~ log(wgt) + race, data = hunde))
```

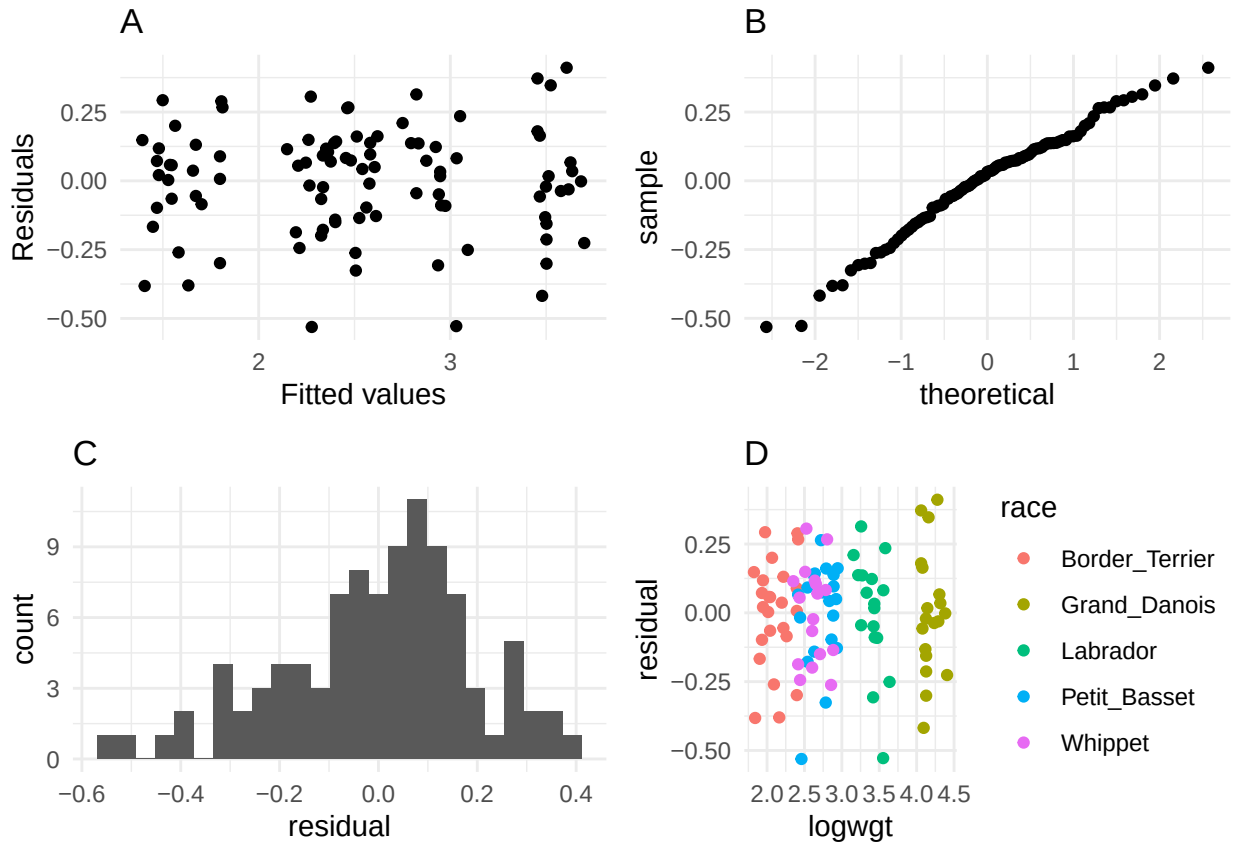
```
##   (Intercept) log(wgt) raceGrand_Danois raceLabrador racePetit_Basset
## 1           1 2.219203                0            0                0
## 2           1 1.931521                0            0                0
## 3           1 2.041220                0            0                0
## 4           1 2.397895                0            0                0
## 5           1 1.840550                0            0                0
## 6           1 1.902108                0            0                0
##   raceWhippet
## 1           0
## 2           0
## 3           0
## 4           0
## 5           0
## 6           0
```

Og den multiple lineære regressionsmodel kan fittes i R med lm:

```
hunde$logmaxLA <- log(hunde$maxLA)
hunde$logwgt <- log(hunde$wgt)
fit <- lm(logmaxLA ~ logwgt + race, data = hunde)
```

For at undersøge antagelsen om at fejlene/residualerne er normalfordelt visualiserer vi residualerne i nedenstående fire figurer. Figur A viser residualerne plottet mod de fittede værdier i modellen, og generelt ser der ikke ud til at være nogen sammenhæng mellem residualer og fittede værdier. Figur B er et QQ-plot, som viser at fraktilerne i den observerede fordeling af residualer plottet mod fraktilerne fra en normalfordeling generelt ligger på en lige linje. Figur C viser den observerede fordeling af residualer i et histogram, og selvom der er lidt flere observationer i den mest negative ende af

fordelingen, end der er i den positive ende, så ser normalfordelingsantagelsen alligevel ikke helt forfærdelig ud, da observationerne generelt er centreret omkring 0 og har en nogenlunde symmetrisk fordeling. Figur D viser residualerne plottet mod log(wgt) og farvet efter hunderace. Også her ser der ikke ud til at være den store forskel i fordeling af residualer på baggrund af de forklarende variable.



Konklusionen bliver, at vi vælger at tro på antagelsen om, at residualerne er uafhængige og identisk fordelte med en normalfordeling, og dermed kan vi tillade os at konstruere konfidensintervaller for parameterestimaterne ud fra t-fordelingen. Svarende til vores valgte parametrisering bliver parameterestimaterne med tilhørende 95%-konfidensintervaller:

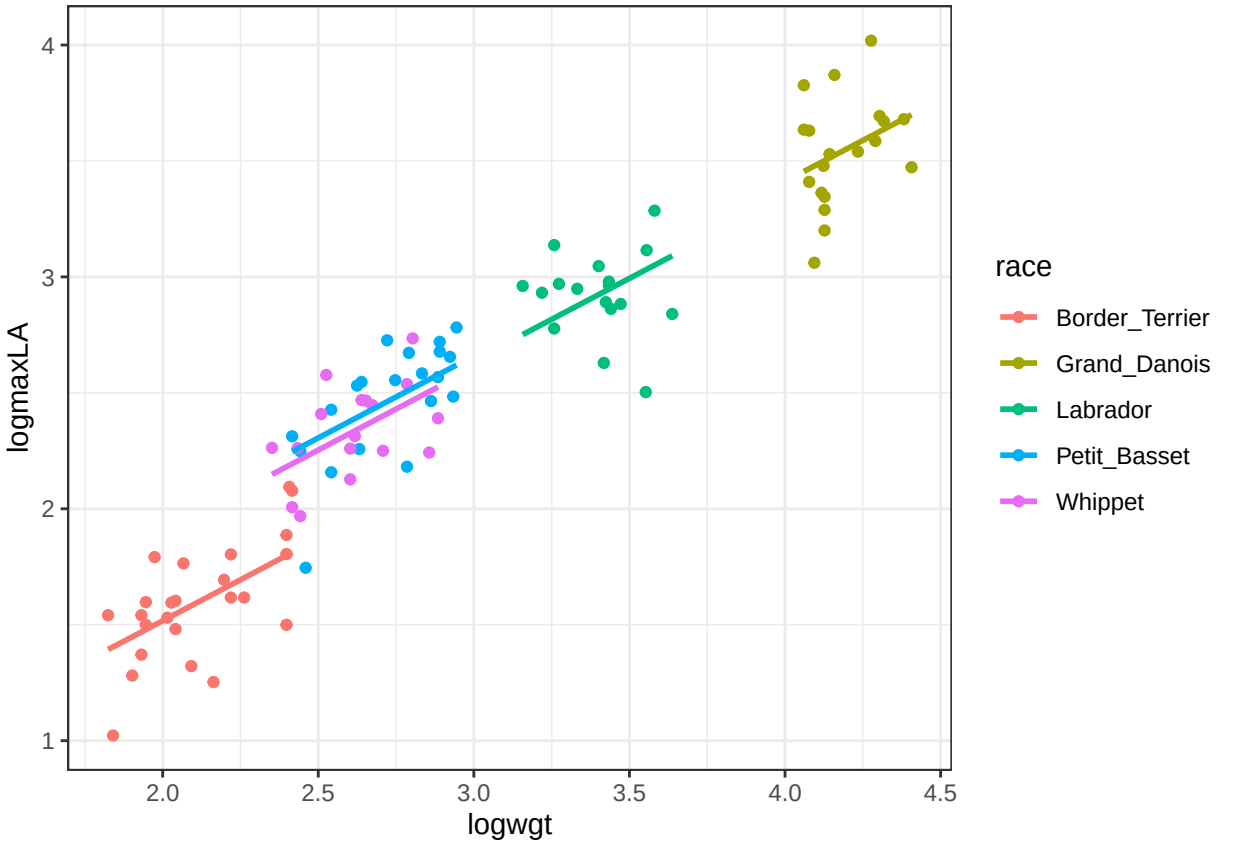
```
##     beta                term estimate lower_ci upper_ci
## 1 beta_0           intercept    0.103   -0.448    0.655
## 2 beta_1               logwgt    0.707    0.448    0.966
## 3 beta_2 race: Grand_Danois    0.481   -0.069    1.031
## 4 beta_3     race: Labrador    0.416    0.059    0.773
## 5 beta_4 race: Petit_Basset    0.435    0.236    0.634
## 6 beta_5      race: Whippet    0.383    0.201    0.565
```

Modellen er en additiv model, hvor $\beta_0 = 0.10$ (interceptet) er estimatet af log(maxLA) for Border Terriers med log(wgt) = 0. β_1 er koefficienten for "hældningen" i den affine funktion der beskriver sammenhængen mellem log(wgt) og log(maxLA), dvs. at for hver stigning i log(wgt) på 1 stiger log(maxLA) med 0.71 uanset hunderace. β_{2-5} er forskellen i log(maxLA) for de øvrige hunderacer sammenlignet med en Border Terrier. Hvis man f.eks. vil bruge modellen til at estimere venstre forkammervolumen i mL hos en Grand Danois med en vægt på 70 kg (dvs. log(wgt) = 4.25) skal man altså udregne:

$$\exp(\beta_0 + \beta_1 \cdot 4.25 + \beta_2) = \exp(0.10 + 0.71 \cdot 4.25 + 0.48) = 36.15 \text{ mL}$$

(Det angivne resultat er udregnet med flere decimaler i mellemregningerne end vist ovenfor)

Modellen kan også visualiseres således:



Vi bemærker desuden, at hvis vi ville teste om det var rimeligt at antage, at hældningen for de fem hunderacer var ens, kunne vi foretage et test for vekselvirkning, dvs. f.eks. ved at undersøge modellen med med parametriseringen $\mu(\log(\text{weight}), \text{race}) = \beta_0 + \beta_{\text{race},\log(\text{wgt})} \cdot \log(\text{wgt})$, dvs. ved at lade hver hunderace have sin egen hældning af linjen for sammenhængen med log(wgt), og man kunne sammenligne de to modeller, f.eks. med et F-test. Baseret på opgaveformuleringen betragter vi dog dette som uden for denne opgave.

Opgave 9

Vi betragter først underrummet svarende til vores model ovenfor, dvs. $L_{\log(\text{wgt})} + L_{\text{race}}$

Og det mindre underrum L_{race} svarende til en model, hvor vi kun bruger log(wgt) som forklarende variabel, og vi bemærker at:

$$L_{\text{race}} \subseteq L_{\text{race}} + L_{\log(\text{wgt})}$$

Vi bruger nu et F-test til at teste nulhypotesen $H_0 : \mu \in L_{\text{race}}$ under den større model givet ved $L_{\text{race}} + L_{\log(\text{wgt})}$. Eftersom vi har konkluderet at residualerne i vores model er normalfordelte kan vi gøre dette eksakt ud fra en teoretisk F-fordeling med (4,91) frihedsgrader. Dette gøres ved følgene R-kode:

```
fit0 <- lm(logmaxLA ~ logwgt, data = hunde)
a <- anova(fit0, fit)
```

Vi får en F-statistic på:

```
a$F[2]
```

```
## [1] 9.822703
```

Hvilket svarer til $P < 1.2 \cdot 10^{-6}$, og vi forkaster dermed nulhypotesen og konkluderer, at modellen som indeholder både hunderace og log(vægt) er signifikant bedre til at beskrive fordelingen af log(maxLA), end modellen som kun indeholder log(vægt), hvilket er ensbetydende med at datasættet *ikke* understøtter en hypotese om, at der ikke er forskel på venstre forkammervolumen for de forskellige racer, hvis man justerer for hundens kropsvægt. En mere mundret formulering er, at datasættet understøtter den videnskabelige hypotese om, at der *er* forskel på hunderacernes venstre forkammervolumen, selvom der justeres for kropsvægt.