Eksamen i Statistik 2, 22. juni 2017

Vejledende besvarelse

Opgave 1

1. Da $E[1(Y_1 = 2)] = P(Y_1 = 2) = p_2$ følger af Store Tals Lov, at $\hat{\theta}_n \xrightarrow{P} p_2$ for $n \to \infty$. Specielt er estimatoren konsistent for p_2. Da $V[1(Y_1 = 2)] = p_2(1-p_2)$ giver Den Centrale Grænseværdisætning, at $\hat{\theta}_n \overset{as}{\sim} \mathcal{N}(p_2, \frac{p_2(1-p_2)}{n})$.

2. Fordelingen af Y har tæthed (mht. tællemålet på $\{0,1,2\}$) givet ved

$$\begin{aligned} f_\theta(y) &= p_0^{1(y=0)} \cdot p_1^{1(y=1)} \cdot p_2^{1(y=2)} \\ &= (1-p_1-p_2)^{1-1(y=1)-1(y=2)} \cdot p_1^{1(y=1)} \cdot p_2^{1(y=2)} \\ &= \exp\left(\log\left(\frac{p_1}{1-p_1-p2}\right) \cdot 1(y=1) + \log\left(\frac{p_2}{1-p_1-p_2}\right) \cdot 1(y=2)\right) \\ &\times (1-p_1-p_2), \end{aligned}$$

   Dette er en eksponentiel familie med kanonisk stikprøvefunktion $t(y) = (1_{(y=1)}, 1_{(y=2)})^T$, parameter $\theta = (\theta_1, \theta_2)^T$ (se opgaveformuleringen) og normeringskonstant $c(\theta) = \frac{1}{1-p_1-p_2} = 1 + \exp(\theta_1) + \exp(\theta_2)$.

3. Fra delopgave 1 ved vi, at $\hat{\theta}_n$ er konsistent for $p_2 = p^2$. Det følger af EH s. 174 formel (5.3) og *Deltametoden* (EH: Sætning 5.15) med $f(x) = \sqrt{x}$, at $f(\sqrt{\hat{\theta}_n})$ er konsistent (for p) og asymptotisk normalfordelt med (asymptotisk) middelværdi $p = f(p^2)$ og asymptotisk varians $\frac{1}{n} f'(p^2) p^2 (1-p^2) f'(p^2)$. Den asymptotiske varians kan omskrives til $\frac{1-p^2}{4n}$.

4. Likelihoodfunktion

$$\begin{aligned} L_{Y_1,\ldots,Y_n}(p) &= \prod_{i=1}^n \binom{2}{Y_i} p^{Y_i} (1-p)^{2-Y_i} \\ &= \left(\prod_{i=1}^n \binom{2}{Y_i}\right) \cdot p^{\sum_i Y_i} (1-p)^{2n - \sum_i Y_i} \end{aligned}$$

   (Minus) loglikelihoodfunktion

$$l_{Y_1,\ldots,Y_n}(p) = c - \sum_i Y_i \cdot \log(p) - \log(1-p) \cdot (2n - \sum_i Y_i)$$

   Scorefunktion

$$l'_{Y_1,\ldots,Y_n}(p) = -\frac{\sum_i Y_i}{p} + \frac{2n - \sum_i Y_i}{1-p}$$

5. Den observerede information

$$l''_{Y_1,\ldots,Y_n}(p) \quad = \quad \frac{\sum_i Y_i}{p^2} + \frac{2n - \sum_i Y_i}{(1-p)^2}$$

er strengt positiv, hvorfor en eventuel løsning til likelihoodligningen vil være et globalt minimum for $l_{Y_1,\ldots,Y_n}(p)$. Løses likelihoodligningen fås følgende udtryk for maksimaliseringsestimatoren $\hat{p}_n = \frac{\sum_i Y_i}{2n}$.

Bemærk: Med de praktiske antagelser omkring problemstillingen er vi kun interesserede i parameterværdier i det åbne interval $0 < p < 1$. For $\sum_i Y_i = 0$ eller $\sum_i Y_i = 2n$ eksisterer MLE således ikke, men dette indtræffer med ssh 0 i grænsen, så MLE er asymptotisk veldefineret. Hvis man vælger at betragte likelihoodfunktionen over hele $[0,1]$ så eksisterer MLE altid, men ligger på randen for de to typer af *panikobservationer* indikeret ovenfor. Der gives fuldt point, selvom man ikke forholder sig til disse specialtilfælde.

Den asymptotiske fordeling af MLE kan bestemmes enten vha. Den Centrale Grænseværdisætning eller vha. Cramér's sætning (EH: Sætning 5.23). Da Y følger en binomialfordeling med antalsparameter 2 og sandsynlighedsparameter p, så er $EY = 2p$ og den forventede information (for een observation!) kan fx. findes ved at indsætte i udtrykket fra 4. (svarende til $n = 1$)

$$E_p[l''_Y(p)] = \frac{2p}{p^2} + \frac{2-2p}{(1-p)^2} = 2\frac{1-p+p}{p(1-p)}.$$

Vi konkluderer, at $\hat{p}_n \overset{as}{\sim} \mathcal{N}(p, \frac{p(1-p)}{2n})$.

Kommentar: Den asymptotiske varians på estimatoren fra delspørgsmål 3. er lig med $\frac{1+p}{2p}$ gange den asymptotiske varians for MLE. Det ses, at denne faktor altid er ≥ 1. Dette er en konsekvens af den asymptotiske optimalitet af MLE som diskuteret i EH kapitel 5.4.

6. Reparameteriseringsafbildningen fra p til B er givet ved afbildningen

$$B = \phi(p) \quad = \quad \frac{D}{p}$$

MLE i den alternative parametrisering bliver blot $\hat{B}_n = \phi(\hat{p}_n)$ (se f.x. EH eksempel 15.19), så den asymptotiske fordeling kan bestemmes vha. Deltametoden. For at udtrykke den asymptotiske varians i B-parametriseringen er det nyttigt at bemærke, at $p = \phi^{-1}(B) = \frac{D}{B}$. Opgaven løses nu ved at omregne følgende udtryk til B-parametriseringen

$$D\phi(p)\frac{p(1-p)}{2n}D\phi(p).$$

Kommentar: Den asymptotiske varians er en aftagende funktion af diameteren D. Det giver mening at en papskive med kendt diameter meget tæt på B gør det muligt at bestemme bredden af plankerne med stor præcision. En naiv estimator (med varians 0!) ville være D, men denne er ikke central og bliver udkonkurreret af MLE i det lange løb. Der gælder et tilsvarende resultat, hvis man benytter sig af papbrikker med andre former blot sandsynligheden for at berøre to planker vokser lineært med diameteren af den mindste cirkulære skive, som kan indeholde papbrikken. Et udartet tilfælde fås ved at betragte uendelig tynde pinde (=tændstikformede papbrikker), hvor $p = \frac{D}{B}(1/2 + 1/\pi)$. Varianter af denne situation omtales som *Buffons nåleproblem*.

Opgave 2

1. Udgangsmodellen (=vekselvirkningsmodellen) udtrykker, at $X = (X_i)_{i\in I}$ er regulært normalfordelt på $\mathbb{R}^I$ med middelværdi $\xi \in L_{G\times V}$ og varians $\sigma^2 I$. Den additive hypotese, $H_0 : \xi \in L_G + L_V$ kan testes ved F-teststørrelsen $F = 0.8124$, der under H_0 følger en F-fordeling med $(1,11)$-frihedsgrader. P-værdien ses at være 0.3867, hvorfor vi accepterer nulhypotesen. Det er ikke et krav at model og hypotese opskrives, men angivelse af teststørrelse, P-værdi og konklusion anses som minimum for en fuldstændig besvarelse. Resultaterne er aflæst i R-udskriften efter `anova(mod2, mod1)`.

2. Dimensionen af det additive underrum er her 3. Det er fint, blot at argumentere med, at der optræder 3 estimater for middelværdistrukturen, når man laver et summary af modellen `mod2`. Det demonstrerer dog et større overblik, hvis man benytter den generelle formel

$$dim(L_G + L_V) = dim(L_G) + dim(L_V) - dim(L_{G\wedge V}) = 2 + 2 - 1,$$

   hvor det benyttes at minimumsfaktoren $G \wedge V$ er identisk med den konstante faktor. Den størrelse der ønskes beregnet i opgaven er F-teststørrelsen for test af den additive hypotese i tosidet variansanalyse med *geometrisk ortogonale faktorer*. Det konstateres, at teststørrelsen bliver 1.046 som *ikke* er lig med teststørrelsen fra R-udskriften. Dette skyldes, at faktorerne G og V i dette tilfælde ikke er geometrisk ortogonale, netop fordi observationen fra et af forsøgsplottene mangler.

3. MLE for parametrene i middelværdistrukturen er givet ved formlen $\hat{\beta} = (A^T A)^{-1} A^T X$, hvor A er designmatricen for den valgte parametrisering af den additive model. Det følger af EH korollar 10.21, at $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$, hvor $A^T A$ og dens inverse er anført i R-udskriften. Vi har, at

$$\hat{\beta} = (9.777, -0.322, -2.280)^T.$$

   Her angiver β_1 middelværdien for kombinationen `G = hoj, V = II`, mens β_2 angiver forskellen mellem `V = I` og `V = II`, og β_3 angiver forskellen mellem `G = lav` og `G = hoj`.

4. Da hver kombination af gødning (`G`) og vanding (`V`) optræder i forsøgsplanen, og da sort (`S`) optræder for hver vandret række (gødning) og for hver lodret søjle (vanding) konkluderes udmiddelbart, at

$$\mathtt{G} \wedge \mathtt{V} = \mathtt{G} \wedge \mathtt{S} = \mathtt{V} \wedge \mathtt{S} = \mathtt{1}.$$

   Da hver sort optræder netop en gang inden for hver jordtype haves desuden $\mathtt{J} \wedge \mathtt{S} = \mathtt{1}$. Derimod viser antalstabellerne nedenfor, at $\mathtt{G} \wedge \mathtt{J}$ og $\mathtt{V} \wedge \mathtt{J}$ er (to forskellige) faktorer med to niveauer.

```
table(g, j)
```

```
##     j
## g    I II III IV
##   G1 2  2   0  0
##   G2 2  2   0  0
##   G3 0  0   2  2
##   G4 0  0   2  2
```

```
table(v, j)

##    j
## v   I II III IV
##   A 2  0   2  0
##   B 2  0   2  0
##   C 0  2   0  2
##   D 0  2   0  2
```

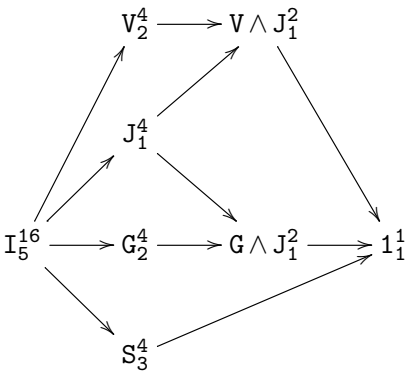
5. Tilføjes de to ikke-trivielle minima fra delspørgsmål 4.konstateres, at mængden

$$\mathbb{G} = \{V, G, S, J, V \wedge J, G \wedge J, 1\}$$

udgør et geometrisk ortogonalt design, som er afsluttet over for dannelse af minimum. Dimensionerne af V_G-rummene, $G \in \mathbb{G}$, fra sætningen om den ortogonale dekomposition (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

$$\begin{aligned} dim(L_V + L_G + L_J + L_S) &= dim(V_V) + dim(V_G) + dim(V_J) \\ &+ dim(V_S) + dim(V_{G\wedge J}) + dim(V_{V\wedge J}) + dim(V_1) \\ &= 2+2+1+3+1+1+1 = 11. \end{aligned}$$

Det kan her være nyttigt at støtte sig op ad et faktorstrukturdiagram, for at holde styr på ordningen af faktorerne



En **alternativ løsningsmetode** består i at indtaste et fiktivt datasæt (fx. med simulerede målinger) svarende til det angivne forsøgsdesign. Fittes den additive model til dette datasæt i R, så vil antallet af parameterestimater i den additive model angive dimensionen af det ønskede additive underrum.

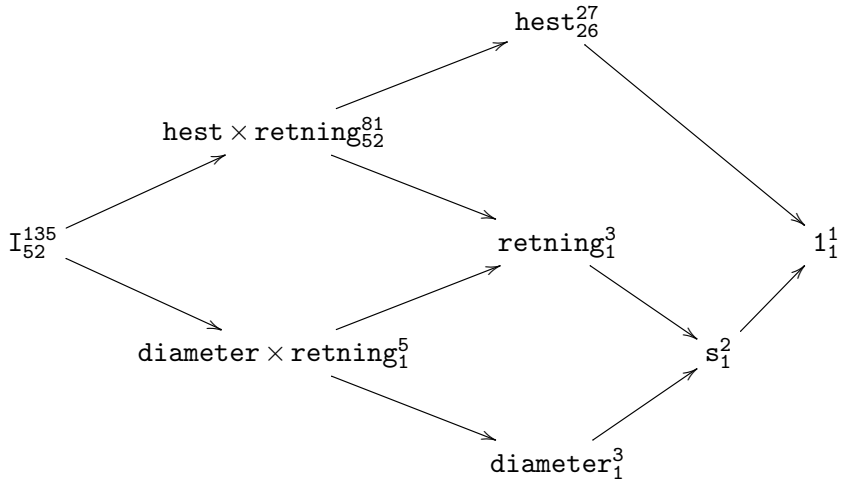
Opgave 3

1. Antalstabellen for faktorerne `diameter` og `retning` viser, at minimum $\mathtt{s} = \mathtt{diameter} \wedge \mathtt{retning}$ er en faktor med 2 labels, hvor hver "blok"opfylder balanceligningen (EH: Sætning 14.8).

```
##        diameter
## retning  0  8 16
##       M 27  0  0
##       H  0 27 27
##       V  0 27 27
```

   Dermed er disse to faktorer geometrisk ortogonale. Minimumsfaktoren angiver om symmetrimålingen stammer fra en måleserie, hvor hesten går lige ud eller bevæger sig i cirkler.

   Produktfaktoren mellem `diameter` og `retning` er reelt kun en faktor med 5 labels. Da datasættet består af netop een måling fra hver hest for hvert niveau af produktfaktoren, så vil alle antalstabeller konstrueret ud fra faktorer i designet opfylde balanceligningen. Specielt er der tale om et geometrisk ortogonalt design og dimensionerne af underrummene som indgår i den ortogonale dekomposition fra EH sætning 14.21 kan beregnes rekursivt ved *håndkraft*.



2. Formålet med forsøget vedrører ikke de konkrete 27 heste der indgår i eksperimentet, så alle faktorer indeholdende `hest` bør indgå som tilfældige effekter. Der bliver således to oplagte tilfældige effekter svarende til effektparrene $(\mathtt{hest}, 1)$ og $(\mathtt{hest} \times \mathtt{retning}, 1)$. Lader vi B_1 og B_2 betegne effektmatricerne hørende til de to effektpar, så kan modellen udtrykkes ved at $X = (X_i)_{i \in I}$ er normalfordelt på $\mathbb{R}^{135}$ med $\xi = EX \in L_{\mathtt{diameter} \times \mathtt{retning}}$ og $VX = \sigma^2 I + \nu_1^2 B_1 B_1^T + \nu_2^2 B_2 B_2^T$.

   Der gives et fradrag, hvis man (uden nogen form for argumentation) vælger *ikke* at inddrage en tilfældig effekt af $\mathtt{hest} \times \mathtt{retning}$ i modellen.

3. Parameterestimatorerne for variansparametrene (baseret på REML-estimation) bliver

$$\hat{\sigma}^2 = 0.3632^2 \quad \hat{\nu}_1^2 = 0.2739^2 \quad \hat{\nu}_2^2 = 0.5166^2.$$

Den totale varians på en symmetriscore bliver $\sigma^2 + \nu_1^2 + \nu_2^2$ og kovariansmatricen for de 5 målinger på en given hest kan udtrykkes som

$$\begin{pmatrix} \sigma^2+\nu_1^2+\nu_2^2 & \nu_1^2 & \nu_1^2 & \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \sigma^2+\nu_1^2+\nu_2^2 & \nu_1^2+\nu_2^2 & \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \nu_1^2+\nu_2^2 & \sigma^2+\nu_1^2+\nu_2^2 & \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \nu_1^2 & \nu_1^2 & \sigma^2+\nu_1^2+\nu_2^2 & \nu_1^2+\nu_2^2 \\ \nu_1^2 & \nu_1^2 & \nu_1^2 & \nu_1^2+\nu_2^2 & \sigma^2+\nu_1^2+\nu_2^2 \end{pmatrix}$$

(her er målingerne organiseret inden for hver hest som anført i datasættet).

4. Test for reduktion i middelværdistrukturen foretages ved brug af likelihoodratioteststørrelsen. Vi benytter her en χ^2-fordeling som approksimation ved beregning af p-værdien.

   Vekselsvirkningen $\texttt{diameter} \times \texttt{retning}$ ($H_0 : \xi \in L_{\texttt{diameter}} + L_{\texttt{retning}}, LRT = 0.9237, p = 0.3365$) og hovedeffekten af `retning` ($H_0 : \xi \in L_{\texttt{diameter}}, LRT = 0.8605, p = 0.3536$) lader ikke til at bidrage væsentligt til at forklare variationen i symmetriscorerne.

   Yderligere reduktion af modellen svarende til hypotesen $H_0 : \xi \in L_1$ forkastes ($LRT = 75.266, p < 0.0001$), hvilket indikerer at diameteren har betydning for symmetriscoren.

   Baseret på faktorstrukturdiagrammet kan man vælge at indskyde hypotesen $H_0 : \xi \in L_s$ om, at det kun betyder noget for symmetriscoren, om hesten løber ligeud eller i cirkler. Denne hypotese forkastes dog ($LRT = 33.433, p < 0.0001$).

5. Vi tilføjer en ny variable `invdiam` til datasættet, og tester hypotesen om at middelværdistrukturen er givet ved $\xi_i = EX_i = \alpha + \beta \cdot \texttt{invdiam}_\texttt{i}$ mod modellen $\xi \in L_{\texttt{diameter}}$. Et likelihoodratiotest for denne hypotese giver teststørrelsen $LRT = 5.166$ der ved opslag i en χ^2-fordeling med 1 frihedsgrad svarer til en p-værdi på 0.023. Der er således (svag) evidens imod hypotesen om, at symmetriscoren afhænger lineært af den reciprokke diameter.

Eksempel på R-kode som kunne være brugt til løsning af opgave 3

```
acc_data <- read.table(file = "stat2juni2017opg3.txt", header = T)
head(acc_data)

##   hest retning diameter         S
## 1  G01       M        0 -6.257128
## 2  G01       H        8 -4.723692
## 3  G01       H       16 -5.000378
## 4  G01       V        8 -4.297942
## 5  G01       V       16 -4.589719
## 6  G02       M        0 -5.348065

table(acc_data$retning, acc_data$diameter)

##
##      0  8 16
##   H  0 27 27
##   M 27  0  0
##   V  0 27 27

### tilfoej minimum af retning og diameter
acc_data$cirkel <- acc_data$diameter != 0

library(lme4)
modelfull <- lmer(S ~ factor(diameter) * retning +
                    (1 | hest) + (1 | hest : retning)
                  , data = acc_data, REML = TRUE)
modelfull

## Linear mixed model fit by REML ['lmerMod']
## Formula: S ~ factor(diameter) * retning + (1 | hest) + (1 | hest:retning)
##    Data: acc_data
## REML criterion at convergence: 247.835
## Random effects:
##  Groups       Name        Std.Dev.
##  hest:retning (Intercept) 0.5166
##  hest         (Intercept) 0.2739
##  Residual                 0.3632
## Number of obs: 135, groups:  hest:retning, 81; hest, 27
## Fixed Effects:
##                (Intercept)           factor(diameter)8
##                   -6.09429                     1.24956
##         factor(diameter)16                   retningV
##                    0.84203                     0.07725
## factor(diameter)8:retningV
##                    0.13242
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients
```

```r
model0 <- lmer(S ~ factor(diameter) * retning +
                 (1 | hest) + (1 | hest : retning)
               , data = acc_data, REML = FALSE)
model1 <- lmer(S ~ factor(diameter) + retning +
                 (1 | hest) + (1 | hest : retning)
               , data = acc_data, REML = FALSE)
anova(model1, model0)
```

```
## Data: acc_data
## Models:
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
## model0: S ~ factor(diameter) * retning + (1 | hest) + (1 | hest:retning)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model1  7 250.08 270.42 -118.04   236.08
## model0  8 251.16 274.40 -117.58   235.16 0.9237      1     0.3365
```

```r
model2a <- lmer(S ~ retning + (1 | hest) + (1 | hest : retning)
                , data = acc_data, REML = FALSE)
model2b <- lmer(S ~ factor(diameter) + (1 | hest) + (1 | hest : retning)
                , data = acc_data, REML = FALSE)
anova(model2a, model1)
```

```
## Data: acc_data
## Models:
## model2a: S ~ retning + (1 | hest) + (1 | hest:retning)
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model2a  6 281.45 298.88 -134.72   269.45
## model1   7 250.08 270.42 -118.04   236.08 33.364      1  7.643e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model2b, model1)
```

```
## Data: acc_data
## Models:
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model2b  6 248.94 266.38 -118.47   236.94
## model1   7 250.08 270.42 -118.04   236.08 0.8605      1     0.3536
```

```r
model3 <- lmer(S ~ 1 + (1 | hest) + (1 | hest : retning)
               , data = acc_data, REML = FALSE)
anova(model3, model2b)
```

```
## Data: acc_data
## Models:
## model3: S ~ 1 + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model3   4 320.21 331.83 -156.10   312.21
## model2b  6 248.94 266.38 -118.47   236.94 75.266      2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model3a <- lmer(S ~ cirkel + (1 | hest) + (1 | hest : retning)
                , data = acc_data, REML = FALSE)
anova(model3a, model2b)
```

```
## Data: acc_data
## Models:
## model3a: S ~ cirkel + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model3a  5 280.38 294.90 -135.19   270.38
## model2b  6 248.94 266.38 -118.47   236.94 33.433      1  7.375e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
### genfit model med REML-estimation
model2bfinal <- lmer(S ~ factor(diameter) + (1 | hest : retning) +(1 | hest)
                     , data = acc_data, REML = TRUE)
VarCorr(model2bfinal)
```

```
##  Groups       Name        Std.Dev.
##  hest:retning (Intercept) 0.51557
##  hest         (Intercept) 0.27462
##  Residual                 0.36286
```

```r
confint(model2bfinal)
```

```
## Computing profile confidence intervals ...
```

```
##                          2.5 %     97.5 %
## .sig01                0.3851479  0.6615867
## .sig02                0.0000000  0.4706491
## .sigma                0.2999166  0.4413148
## (Intercept)          -6.3530665 -5.8355101
## factor(diameter)8     1.0615006  1.6472870
## factor(diameter)16    0.5877646  1.1735511
```

```r
acc_data$invdiam <- 1/acc_data$diameter
acc_data$invdiam[acc_data$diameter == 0] <- 0
head(acc_data)
```

```
##   hest retning diameter         S cirkel invdiam
## 1  G01       M        0 -6.257128  FALSE  0.0000
## 2  G01       H        8 -4.723692   TRUE  0.1250
## 3  G01       H       16 -5.000378   TRUE  0.0625
## 4  G01       V        8 -4.297942   TRUE  0.1250
## 5  G01       V       16 -4.589719   TRUE  0.0625
## 6  G02       M        0 -5.348065  FALSE  0.0000
```

```r
model3b <- lmer(S ~ invdiam + (1 | hest) + (1 | hest : retning)
                , data = acc_data, REML = FALSE)
anova(model3b, model2b)
```

```
## Data: acc_data
## Models:
## model3b: S ~ invdiam + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model3b  5 252.11 266.63 -121.06   242.11
## model2b  6 248.94 266.38 -118.47   236.94 5.1656      1    0.02304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```