Complex data - lab2

Stanisław Wilczyński

22 May 2018

Task 1

1. $(0, 0, 1, 0)$
2. $(0, 0, 1, -1)$ - this is a contrast

```
lead <- read.table(file = "../data/lead.txt", header = FALSE)
## Give names to variables
names(lead) <- c("id", paste("y", 1:4, sep=""))
lead.uni <- data.frame(id=rep(lead$id, each=4),
  y=as.numeric(t(as.matrix(lead[,2:5]))),
  time=rep(c(0,1,4,6)),
  time.cat=rep(1:4))
```

3. Below is the output for the test. The answer is 3.

```
lead.cat.ml <- gls(y~factor(time),
correlation=corSymm(form= ~1 | id),
weights=varIdent(form= ~1 | factor(time)),
method = "ML",
data=lead.uni)

lead.cat.no.ml <- gls(y~1,
correlation=corSymm(form= ~1 | id),
weights=varIdent(form= ~1 | factor(time)),
method = "ML",
data=lead.uni)

anova(lead.cat.ml, lead.cat.no.ml)
```

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio
## lead.cat.ml        1 14 1314.459 1360.635 -643.2294
## lead.cat.no.ml     2 11 1381.866 1418.148 -679.9331 1 vs 2 73.40745
##                p-value
## lead.cat.ml
## lead.cat.no.ml  <.0001
```

4. REML can't be used to compare nested models for the means in likelihood ratio tests. The reason is that REML estimates the random effects by considering linear combinations of the data that remove the fixed effects. If the fixed effects are changed two models are not directly comparable anymore. For example in case of simple linear regression the restricted maximum likelihood estimator is $\hat{\sigma}^2 = \frac{RSS}{n-p}$, which is clearly dependent on the number of regression coefficients.

5. We used multivariate Wald test for model parameters. We conclude that there is no group by time effect because p-value for group by time is quite large (0.3265) - we do not reject null hypothesis.

6. We conclude that both time and diet are significant - for both covariates the p-value is below standard significance level of 0.05. Score test, likelihood ratio test and Wald test can all be used for testing if some models' parameters are zeros. In fact, Wald and score tests are asymptotically equivalent to the likelihood ratio test. Therefore these 3 tests can be used exchangeably. The main difference is that for Wald and score test you just have to fit one model. In comparison in LRT you need to fit two models.

Therefore when fitting model is computationally expensive it may be more reasonable to use Wald or score tests. If we were to use LRT in to test these two hypothesis we would have to create two *reduced* models: one without **factor(diet)** as explanatory variable, the other one without **factor(time)**.

Task 2

a) Here we define our covariates:

$X_{1ij} = 1$ for all measurements
$X_{2ij} = 1$ if jth measurement was taken at $time = 2$ weeks, 0 otherwise
$X_{3ij} = 1$ if jth measurement was taken at $time = 3$ weeks, 0 otherwise
$X_{4ij} = 1$ if jth measurement was taken at $time = 4$ weeks, 0 otherwise
$X_{5ij} = 1$ if ith cow ate barley and lupins, 0 otherwise
$X_{6ij} = 1$ if ith cow ate only lupins, 0 otherwise
$X_{7ij} = 1$ if ith cow ate barley and lupins and the jth measurement is at $time = 2$, 0 otherwise
$X_{8ij} = 1$ if ith cow ate barley and lupins and the jth measurement is at $time = 3$, 0 otherwise
$X_{9ij} = 1$ if ith cow ate barley and lupins and the jth measurement is at $time = 4$, 0 otherwise
$X_{10ij} = 1$ if ith cow ate only lupins and the jth measurement is at $time = 2$, 0 otherwise
$X_{11ij} = 1$ if ith cow ate only lupins and the jth measurement is at $time = 3$, 0 otherwise
$X_{12ij} = 1$ if ith cow ate only lupins and the jth measurement is at $time = 4$, 0 otherwise

Then the model is:

$$Y_{ij} = \epsilon_{ij} + \beta_1 + \sum_{k=2}^{12} \beta_k X_{kij}$$

and

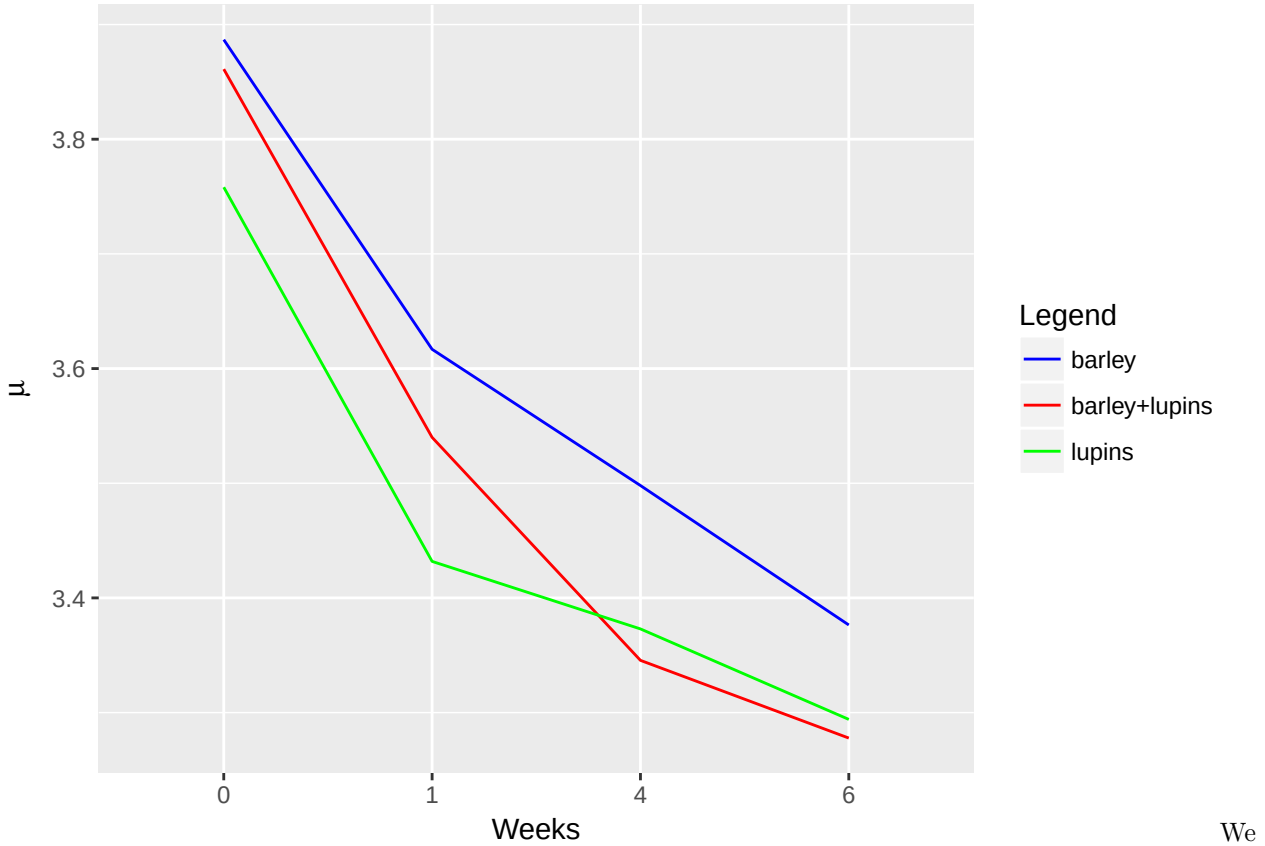
$$\begin{aligned}
\mu_{b1} &= \beta_1 \\
\mu_{b2} &= \beta_1 + \beta_2 \\
\mu_{b3} &= \beta_1 + \beta_3 \\
\mu_{b4} &= \beta_1 + \beta_4 \\
\mu_{lb1} &= \beta_1 + \beta_5 \\
\mu_{lb2} &= \beta_1 + \beta_2 + \beta_5 + \beta_7 \\
\mu_{lb3} &= \beta_1 + \beta_3 + \beta_5 + \beta_8 \\
\mu_{lb4} &= \beta_1 + \beta_4 + \beta_5 + \beta_9 \\
\mu_{l1} &= \beta_1 + \beta_6 \\
\mu_{l2} &= \beta_1 + \beta_2 + \beta_6 + \beta_{10} \\
\mu_{l3} &= \beta_1 + \beta_3 + \beta_6 + \beta_{11} \\
\mu_{l4} &= \beta_1 + \beta_4 + \beta_6 + \beta_{12}
\end{aligned}$$

First we will try to get some intuition about possible results of the tests for parallelism and main effects based on data visualization. Here we provide the plot of means vs time grouped by the cows' diet.

```
moo <- read.table(file = "../data/mooAll.txt", header = TRUE)
colnames(moo) <- c("protein", "week", "cow", "diet")
moo.multi <- NULL
for(cow in unique(moo$cow)){
  p1 <- moo$protein[which(moo$cow==cow & moo$week==1)]
```

```
  p2 <- moo$protein[which(moo$cow==cow & moo$week==2)]
  p3 <- moo$protein[which(moo$cow==cow & moo$week==3)]
  p4 <- moo$protein[which(moo$cow==cow & moo$week==4)]
  diet <- unique(moo$diet[which(moo$cow==cow & moo$week==1)])[1]
  moo.multi <- rbind(moo.multi, c(p1,max(p2,3),p3,p4,diet))
}
moo.barley.means <- apply(moo.multi[which(moo.multi[,5]==1),], mean, MARGIN = 2, na.rm=TRUE)
moo.mixed.means <-  apply(moo.multi[which(moo.multi[,5]==2),], mean, MARGIN = 2, na.rm=TRUE)
moo.lupins.means <- apply(moo.multi[which(moo.multi[,5]==3),], mean, MARGIN = 2, na.rm=TRUE)
moo.means <- data.frame(rbind(moo.barley.means, moo.mixed.means, moo.lupins.means))
colnames(moo.means) <- c("0", "1", "4", "6", "diet")
moo.means <- melt(moo.means, id.vars = c("diet"))


ggplot(moo.means, aes(x=variable, y=value, group=diet, color=factor(diet))) +
       geom_line() +
       scale_color_manual(labels=c("barley", "barley+lupins", "lupins"), values=c("blue", "red", "green
       labs(x="Weeks", y=TeX('$\\mu$'), colour="Legend")
```



We can clearly see that mean value of **protein** variable decreases with time, so we expect that the influence of time variable to be high. Although the plots look quite similar (in terms of slopes) based purely on visualization it is hard to predict the result of test for parallelism.

b) Now it's time for a real test. We fit the model which takes into account group by time interactions, which means $H_0 : \beta_7 = \ldots = \beta_{12} = 0$ vs H_a : at least one is non-zero.

```
moo.gls.interaction <- gls(protein~factor(week)*factor(diet),
correlation=corSymm(form= ~1 | cow),
```

```r
weights=varIdent(form= ~1 | factor(week)),
data=moo)
summary(moo.gls.interaction)
```

```
## Generalized least squares fit by REML
##   Model: protein ~ factor(week) * factor(diet)
##   Data: moo
##        AIC      BIC    logLik
##   108.8282 190.5304 -32.41412
##
## Correlation Structure: General
##  Formula: ~1 | cow
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.440
## 3 0.474 0.485
## 4 0.321 0.515 0.600
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | factor(week)
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.6789112 0.6399626 0.6305705
##
## Coefficients:
##                                            Value  Std.Error  t-value
## (Intercept)                                3.885532 0.08052206 48.25425
## factor(week)2                             -0.246580 0.07526628 -3.27611
## factor(week)3                             -0.387148 0.07211824 -5.36824
## factor(week)4                             -0.510722 0.07988426 -6.39327
## factor(diet)barley+lupins                 -0.024420 0.11178982 -0.21845
## factor(diet)lupins                        -0.127383 0.11178982 -1.13949
## factor(week)2:factor(diet)barley+lupins -0.074531 0.10420488 -0.71523
## factor(week)3:factor(diet)barley+lupins -0.128408 0.10012548 -1.28247
## factor(week)4:factor(diet)barley+lupins -0.072611 0.11113362 -0.65337
## factor(week)2:factor(diet)lupins        -0.083790 0.10420488 -0.80409
## factor(week)3:factor(diet)lupins         0.001963 0.10012548  0.01960
## factor(week)4:factor(diet)lupins         0.046648 0.11113362  0.41975
##                                           p-value
## (Intercept)                                0.0000
## factor(week)2                              0.0012
## factor(week)3                              0.0000
## factor(week)4                              0.0000
## factor(diet)barley+lupins                  0.8272
## factor(diet)lupins                         0.2554
## factor(week)2:factor(diet)barley+lupins    0.4750
## factor(week)3:factor(diet)barley+lupins    0.2007
## factor(week)4:factor(diet)barley+lupins    0.5140
## factor(week)2:factor(diet)lupins           0.4220
## factor(week)3:factor(diet)lupins           0.9844
## factor(week)4:factor(diet)lupins           0.6750
##
##  Correlation:
```

```
##                                       (Intr) fct()2 fct()3 fct()4 fct()+
## factor(week)2                         -0.749
## factor(week)3                         -0.777  0.724
## factor(week)4                         -0.800  0.770  0.822
## factor(diet)barley+lupins             -0.720  0.539  0.560  0.576
## factor(diet)lupins                    -0.720  0.539  0.560  0.576  0.519
## factor(week)2:factor(diet)barley+lupins  0.541 -0.722 -0.523 -0.556 -0.752
## factor(week)3:factor(diet)barley+lupins  0.560 -0.522 -0.720 -0.592 -0.777
## factor(week)4:factor(diet)barley+lupins  0.575 -0.554 -0.591 -0.719 -0.800
## factor(week)2:factor(diet)lupins         0.541 -0.722 -0.523 -0.556 -0.390
## factor(week)3:factor(diet)lupins         0.560 -0.522 -0.720 -0.592 -0.403
## factor(week)4:factor(diet)lupins         0.575 -0.554 -0.591 -0.719 -0.414
##                                       fctr() f()2:()+ f()3:()+ f()4:()+
## factor(week)2
## factor(week)3
## factor(week)4
## factor(diet)barley+lupins
## factor(diet)lupins
## factor(week)2:factor(diet)barley+lupins -0.390
## factor(week)3:factor(diet)barley+lupins -0.403  0.728
## factor(week)4:factor(diet)barley+lupins -0.414  0.774    0.823
## factor(week)2:factor(diet)lupins        -0.752  0.522    0.377    0.400
## factor(week)3:factor(diet)lupins        -0.777  0.377    0.519    0.426
## factor(week)4:factor(diet)lupins        -0.800  0.400    0.426    0.517
##                                       fc()2:() fc()3:()
## factor(week)2
## factor(week)3
## factor(week)4
## factor(diet)barley+lupins
## factor(diet)lupins
## factor(week)2:factor(diet)barley+lupins
## factor(week)3:factor(diet)barley+lupins
## factor(week)4:factor(diet)barley+lupins
## factor(week)2:factor(diet)lupins
## factor(week)3:factor(diet)lupins         0.728
## factor(week)4:factor(diet)lupins         0.774    0.823
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.65092977 -0.66554849  0.04504866  0.69691506  2.77838379
##
## Residual standard error: 0.4029334
## Degrees of freedom: 315 total; 303 residual
```

```
anova(moo.gls.interaction)
```

```
## Denom. DF: 303
##                            numDF   F-value p-value
## (Intercept)                    1 19681.141  <.0001
## factor(week)                   3    49.423  <.0001
## factor(diet)                   2     2.840  0.0600
## factor(week):factor(diet)      6     1.152  0.3322
```

As we can see from the anova output the p-value (0.3322) for $factor(week) : factor(diet)$ from the multivariate Wald test (so for group by time interaction) is quite high. Therefore we don't reject the H_0.

c) Now we test the main effect. We change the model slightly (not to include interactions) and run anova function once again. We test $H_0^1 : \beta_5 = \beta_6 = 0$ and $H_0^2 : \beta_2 = \beta_3 = \beta_4 = 0$

```
moo.gls.main <- gls(protein~factor(week)+factor(diet),
correlation=corSymm(form= ~1 | cow),
weights=varIdent(form= ~1 | factor(week)),
data=moo)
summary(moo.gls.main)
```

```
## Generalized least squares fit by REML
##   Model: protein ~ factor(week) + factor(diet)
##   Data: moo
##       AIC      BIC    logLik
##   82.4687 142.2022 -25.23435
##
## Correlation Structure: General
##  Formula: ~1 | cow
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.448
## 3 0.467 0.479
## 4 0.324 0.505 0.605
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | factor(week)
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.6865158 0.6423457 0.6332044
##
## Coefficients:
##                                Value  Std.Error   t-value p-value
## (Intercept)                 3.918225 0.05769065  67.91784  0.0000
## factor(week)2              -0.301282 0.04183379  -7.20189  0.0000
## factor(week)3              -0.430633 0.04068023 -10.58581  0.0000
## factor(week)4              -0.519752 0.04487239 -11.58288  0.0000
## factor(diet)barley+lupins -0.112933 0.06020380  -1.87585  0.0616
## factor(diet)lupins        -0.133910 0.06020380  -2.22427  0.0269
##
##  Correlation:
##                            (Intr) fct()2 fct()3 fct()4 fct()+
## factor(week)2              -0.583
## factor(week)3              -0.609  0.722
## factor(week)4              -0.626  0.765  0.825
## factor(diet)barley+lupins -0.543 -0.001  0.002  0.002
## factor(diet)lupins        -0.543 -0.001  0.002  0.002  0.520
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.72722990 -0.70856392  0.08694074  0.70834695  2.89454872
##
## Residual standard error: 0.4012551
## Degrees of freedom: 315 total; 309 residual
```

```r
anova(moo.gls.main)
```

```
## Denom. DF: 309
##              numDF   F-value p-value
## (Intercept)      1 19672.510  <.0001
## factor(week)     3    49.027  <.0001
## factor(diet)     2     2.829  0.0606
```

We can clearly see that p-value for time covaraites are very low. This means that these variables are significant in our model and we reject H_0^2. On the other hand the p-value for group factor is 0.0606 - still above the standard significance level of 0.05. Therefore we conclude that the influence of the group on our data is negligible and we do not reject H_0^1.