Knee surgery Exercise - solution

Question 1

We first load the data and visualize the first lines.

```
load(url("http://paulblanche.com/files/kneeSurgery.rda"))
d <- kneeSurgery
head(d)
```

```
##   arm site age    sex Oxford.pre Oxford.01 Oxford.06 Oxford.12 Oxford.24
## 1   1    E  69   male         20        35        45        44        47
## 2   2    B  50 female         16        26        43        46        42
## 3   1    A  61 female         20        29        46        46        NA
## 4   1    D  65   male         18        28        40        39        47
## 5   2    A  73   male         28        22        22        27        26
## 6   2    D  73   male         32        33        27        38        43
```

We then create a baseline table to summarize the distribution of the important variables, per arm.

```
library(Publish)
Tab1 <- univariateTable(arm~site+Q(age)+sex+Oxford.pre,
                        data=d,
                        compare.groups = FALSE,
                        show.totals = FALSE)
Tab1
```

```
##     Variable        Level arm = 1 (n=174) arm = 2 (n=172)
## 1       site            A       41 (23.6)       39 (22.7)
## 2                       B       35 (20.1)       38 (22.1)
## 3                       C       41 (23.6)       37 (21.5)
## 4                       D       20 (11.5)       28 (16.3)
## 5                       E       37 (21.3)       30 (17.4)
## 6        age median [iqr] 68 [62.2, 74.0]     67 [60, 74]
## 7        sex       female       82 (47.1)       83 (48.3)
## 8                    male       92 (52.9)       89 (51.7)
## 9 Oxford.pre    mean (sd)        23.1 (6)        22 (6.3)
```

There are no substantial difference between the two arms. This is a consequence of the randomization. We can notice that the baseline Oxford score (i.e., pre-surgery score) was slighty better in arm 1 than in arm 2, on average (1 point). There are aproximately as many men and women and about half of the patients are aged between 60 and 75. As expected, the baseline Oxford scores are rather low (indication for surgey), but there is substantial variablity from patient to patient (SD=6).

Question 2

Let's now have a quick look at the evolution of the scores, for a few random patients.

Question 2.a

We first select 20 random patients, to have approximately 10 of each arm. We select just a few patients because the plots are often difficult to read with too many patients and because a few patients is usually sufficient to get a feel of the data.

```
d20 <- d[101:120,]
```

Question 2.b

To use the `xyplot()` function of the lattice" package to produce a spaghetti plot, we first need to create a long format version of the data. We can do it with the `reshape()` function.

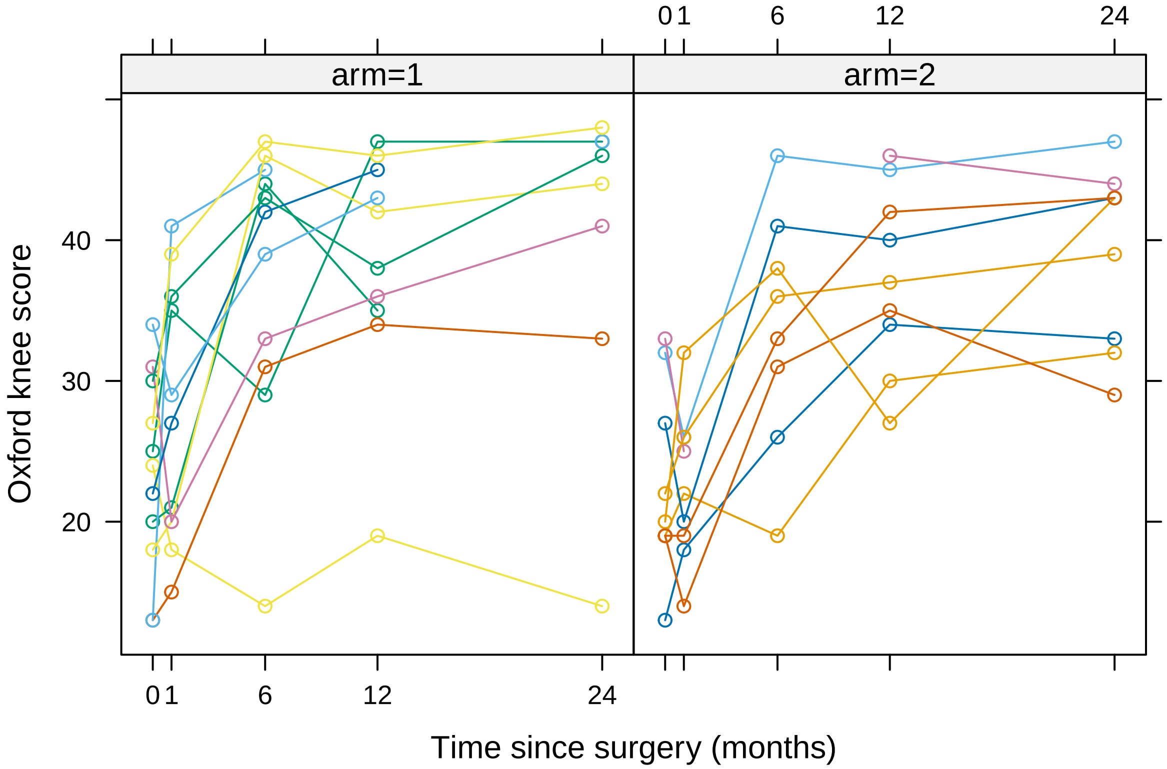
```
thetimes <- c(0,1,6,12,24) # times of repeated measures
long20 <- reshape(d20,
                varying = c( "Oxford.pre", "Oxford.01", "Oxford.06",
                            "Oxford.12", "Oxford.24"),
                v.names = "Oxford",
                timevar = "time",
                times=thetimes,
                direction = "long")
long20 <- long20[order(long20$id),] # reorder by subject id
rownames(long20) <- NULL            # delete row names
head(long20,n=20)                   # quick check
```

```
##    arm site age    sex time Oxford id
## 1    2    A  65 female    0     13  1
## 2    2    A  65 female    1     18  1
## 3    2    A  65 female    6     26  1
## 4    2    A  65 female   12     34  1
## 5    2    A  65 female   24     33  1
## 6    2    E  76 female    0     19  2
## 7    2    E  76 female    1     22  2
## 8    2    E  76 female    6     19  2
## 9    2    E  76 female   12     30  2
## 10   2    E  76 female   24      32  2
## 11   1    C  72 female    0      25  3
## 12   1    C  72 female    1      35  3
## 13   1    C  72 female    6      29  3
## 14   1    C  72 female   12      47  3
## 15   1    C  72 female   24      47  3
## 16   2    B  52   male    0      19  4
## 17   2    B  52   male    1      14  4
## 18   2    B  52   male    6      31  4
## 19   2    B  52   male   12      35  4
## 20   2    B  52   male   24      29  4
```

Question 2.c

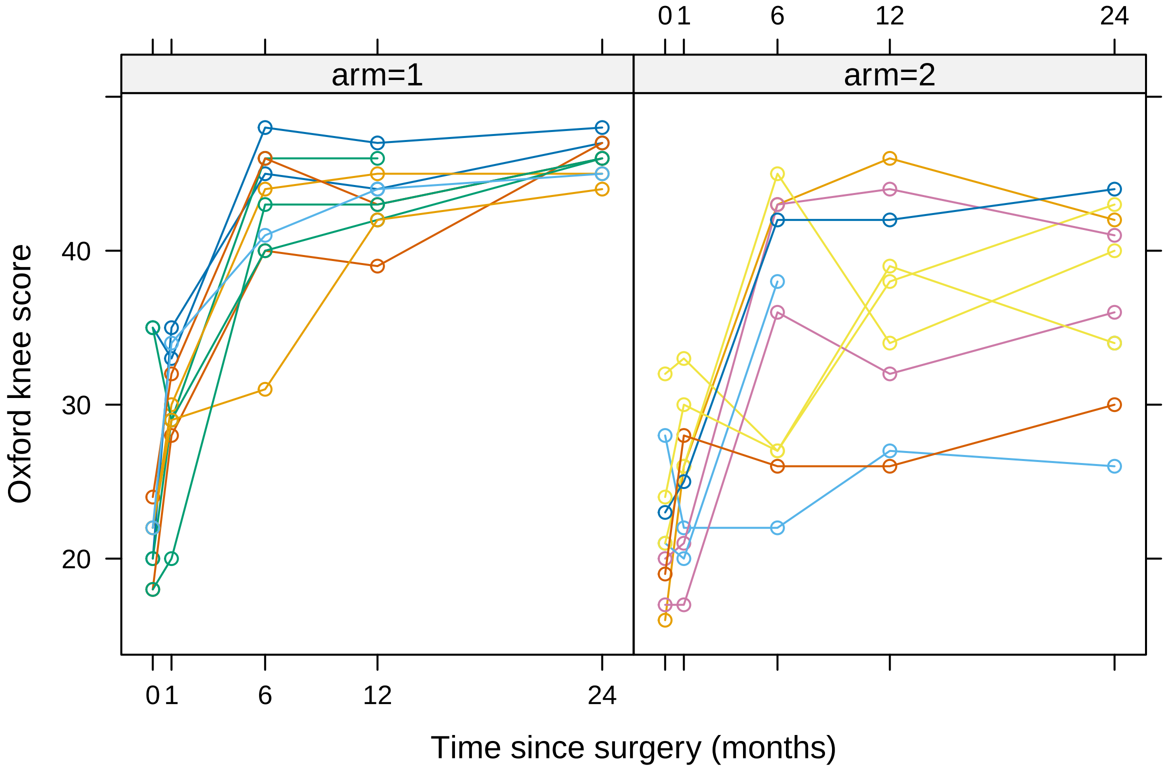
We are now almost ready to call the `xyplot()` function of the "lattice" package to produce a spaghetti plot. We just need to make the variable `arm` a factor variable before.

```r
library(lattice)
long20$arm <- factor(long20$arm,levels=1:2,
                     labels=c("arm=1","arm=2"))
xyplot(Oxford~time | arm,         # show score per time, for each arm
       data=long20,               # data in long format
       group=id,                  # patient id
       type='b',                  # both points and lines are drawn
       xlab="Time since surgery (months)",
       ylab="Oxford knee score",
       scales=list(x=list(at=thetimes))) # set values to show on x-axis
```



Question 3

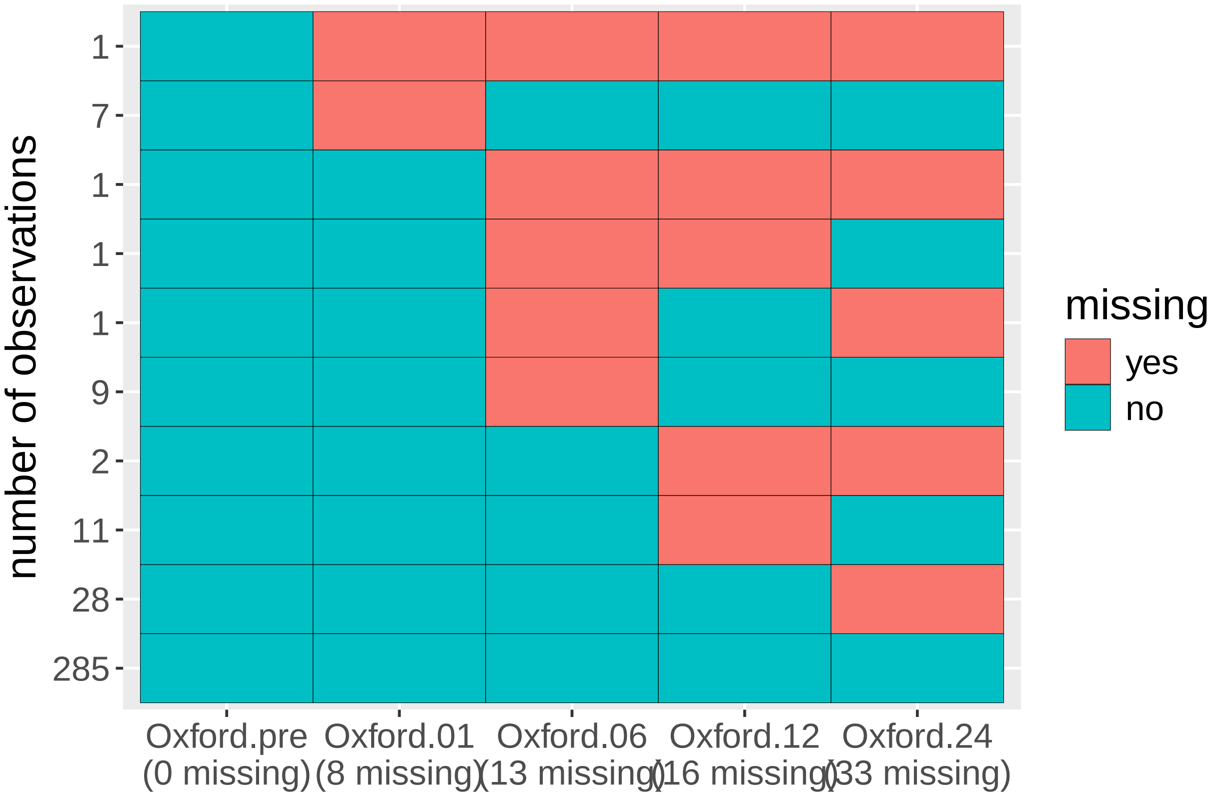
Overall, the score of each patient is improving over time. However, sometimes is goes down a bit before going up again. This will typically happen if e.g., the health of the patient (pain and knee function) is stable and patients answers slightly differently at the same question in a random way, e.g, randomly switch from the answer "With little difficulty" to "With moderate difficulty" (leading to 1 point difference in the score) when replying to the question "During the past 4 weeks… Could you kneel down and get up again afterwards?". Maybe more importantly, we see than for most patients, the score changes much more within the 12 months than from 12 to 24 months (i.e., 1-year after surgery, the score does no longer change much). Finally, based on these 20 patients, the outcome of the surgery does not look very different, in average, at 2 years after surgery. Of course, we should not draw any strong conclusion based on 20 patients and it can be useful to redo the plot for 20 other randomly selected patients. See e.g., the plot below obtained using `d20 <- d[1:20,]`. **Note**: I can say that these patients are randomly selected here because the rows of the dataset are not ordered in any specific way. This would be a questionable statement if , e.g., we had sorted the dataset by date of inclusion. Arm 1 looks so much better when we randomly chose these patients ! This might be a good incentive to redo the plot for 20 other random patients (again, just to get a feel of the data).



Question 4

The next descriptive plot of interest with this kind of repeated measurements data is the plot of the missing data pattern. We could already see that we have missing data from the spaghetti plots, because some subjects did not have dots at all times (and/or no lines between the dots).

```
library(LMMstar)
MissPat <- summarizeNA(d[,c( "Oxford.pre",
                             "Oxford.01",
                             "Oxford.06",
                             "Oxford.12",
                             "Oxford.24")])
plot(MissPat)
```



Question 4.a

The first line (1 patient), third line (1 patient), seventh line (2 patients) and ninth line (28 patients) are compatible with patients being "lost to follow-up". That is, as soon as we have a missing data because a patient does not answer a questionnaire, then the patient does not reply either at questionnaires sent later. In practice, we usually can (and should!) collect data about the reasons why we have missing data. In this exercise, we have no such data.

Question 4.b

Some patients did not reply to the questionnaire at some follow-up times but replied later. E.g., 7 patients did not reply at the questionnaire sent after 1 month, but replied to all questionnaires sent later. In total, we have 7+1+1+9+11=29 with this kind of intermittent missing data.

Question 4.c

We can see that 285 out of 346, i.e. 82%, replied to all questionnaires.

Question 4.d

For the analysis of the change score at 24 months, which is our primary outcome of interest here, we can read from the x-axis than 33 patients have missing data (i.e., did not reply to the questionnaire). That is, 9.5% missing data.

Question 5

Informally, Missing Completely At Random (MCAR) means that the missingness mechanism is unrelated to the outcome and covariates. In this context, it means that the missingness mechanism is unrelated to age, sex, study site and pre-surgery Oxford score (i.e., the baseline covariates that we will use in the model for the analysis) and also unrelated to previously collected Oxford knee scores (i.e., to the answer to the previous questionnaires). This can be realistic if the main reason for not answering is that the patients simply forgot to answer or that we failed to reach out to them. This might be unrealistic if patients who are doing bad have a much stronger incentive to answer the questionnaire than those who are doing well. It could happen if, e.g., there is a free text question at the end of the questionnaire that patient can use to further communicate their worries to their doctors. In that case, missing data would be less common among patients having a "good" score than among patients having a "poor" score.

Informally, Missing At Random (MAR) means that the missingness may depend on covariates and previous measures of the outcome. In this context, it means that the missingness mechanism can be related to age, sex, study site and any Oxford score obtained via previous questionnaires. This is more realistic (less restrictive). Although not perfect, because e.g., the missingness cannot depend on the current score (unobserved, because missing), the more correlated the scores over time and the closer we are from the situation in which we could assume that it can depend on this current value.

Question 6

We now perform the main analysis, by fitting a Mixed Model for Repeated Measurements (MMRM). Before calling the `lmm()` function, we first need to create a data set in the long format. Additionally, we will *center* the covariates `age` and Oxford score pre-surgey and choose 24 months as the reference level for the factor variable. This is just to facilitate the interpretation of the default output of the software.

```
long <- reshape(d,
                varying = c("Oxford.01", "Oxford.06","Oxford.12", "Oxford.24"),
                v.names = "Oxford",
                timevar = "time",
                times=c(1,6,12,24),
                direction = "long")
long <- long[order(long$id),] # reorder by subject ID
```

```
rownames(long) <- NULL # delete row names
## head(long,n=5)      # quick check
#--- data management steps ---
long$time <- factor(long$time)
long$time <- relevel(long$time,ref="24")
long$arm <- factor(long$arm)
long$change <- long$Oxford-long$Oxford.pre
long$age67 <- long$age-67
long$Oxford.pre22 <- long$Oxford.pre-22
#--- fit MMRM --------------------
lmmfit <- lmm(change~Oxford.pre22*time + site*time
              + sex*time + arm*time + age67*time,
              repetition = ~time|id,
              structure = "UN", data = long)
```

```
## Warning in .lmmNormalizeData(as.data.frame(data)[unique(stats::na.omit(var.all))], :
```

```
summary(lmmfit)
```

```
##      Linear Mixed Model
##
## Dataset: long
##
##   - 345 clusters were analyzed, 1 were excluded because of missing values
##   - 1314 observations were analyzed, 70 were excluded because of missing values
##   - between 1 and 4 observations per cluster
##
## Summary of the outcome and covariates:
##
##     $ change      : num  15 25 24 27 10 27 30 26 9 26 ...
##     $ Oxford.pre22: num  -2 -2 -2 -2 -6 -6 -6 -6 -2 -2 ...
##     $ time        : Factor w/ 4 levels "24","1","6","12": 2 3 4 1 2 3 4 1 2 3 ...
##     $ site        : Factor w/ 5 levels "A","B","C","D",..: 5 5 5 5 2 2 2 2 1 1 ...
##     $ sex         : Factor w/ 2 levels "female","male": 2 2 2 2 1 1 1 1 1 1 ...
##     $ arm         : Factor w/ 2 levels "1","2": 1 1 1 1 2 2 2 2 1 1 ...
##     $ age67       : num  2 2 2 2 -17 -17 -17 -17 -6 -6 ...
##     reference level: time=24;site=A;sex=female;arm=1
##
## Estimation procedure
##
##   - Restricted Maximum Likelihood (REML)
##   - log-likelihood :-4078.991
##   - parameters: mean = 36, variance = 4, correlation = 6
##   - convergence: TRUE (6 iterations)
##     largest |score| = 9.257865e-05 for k.12
##             |change|= 6.40005142749089e-06 for sigma
##
## Residual variance-covariance: unstructured
##
##   - correlation structure: ~0 + time
##           24     1     6    12
##     24 1.000 0.374 0.604 0.762
##     1  0.374 1.000 0.432 0.431
##     6  0.604 0.432 1.000 0.706
##     12 0.762 0.431 0.706 1.000
##
##   - variance structure: ~time
##              standard.deviation ratio
##     sigma.24               6.67  1.00
##     sigma.1                6.69  1.00
##     sigma.6                7.09  1.06
##     sigma.12               6.70  1.00
##
## Fixed effects: change ~ Oxford.pre22 * time + site * time + sex * time + arm *      t
##
##                         estimate    se    df   lower  upper p.value
##    (Intercept)           20.656 0.935 323.1  18.817 22.495 < 1e-04 ***
##    Oxford.pre22           -0.68  0.06 320.1  -0.799 -0.561 < 1e-04 ***
##    time1                -14.369 1.047 334.1 -16.429 -12.31 < 1e-04 ***
##    time6                 -2.572 0.881 333.3  -4.304 -0.839 0.00373  **
##    time12                -0.665 0.682   306  -2.007  0.677 0.33048
##    siteB                 -0.135 1.103 316.8  -2.305  2.036 0.90287
##    siteC                 -0.309 1.091 319.4  -2.455  1.837 0.77721
##    siteD                 -1.673  1.26 318.3  -4.152  0.806 0.18527
##    siteE                  0.398 1.128 314.2  -1.821  2.616 0.72466
##    sexmale               -0.511  0.74 317.5  -1.967  0.945 0.49047
##    arm2                  -2.267 0.741 316.4  -3.724  -0.81 0.00240  **
##    age67                 -0.083 0.046   318  -0.174  0.008 0.07413   .
##    Oxford.pre22:time1    -0.039 0.068 331.5  -0.172  0.094 0.56131
##    Oxford.pre22:time6    -0.099 0.057 331.5   -0.21  0.013 0.08187   .
##    Oxford.pre22:time12   -0.078 0.043 300.2  -0.163  0.008 0.07452   .
##    time1:siteB           -1.002 1.233 325.5  -3.428  1.425 0.41725
##    time6:siteB            0.763 1.033 325.1   -1.27  2.796 0.46091
##    time12:siteB           0.712 0.795   302  -0.852  2.277 0.37090
##    time1:siteC            0.154 1.226 332.7  -2.258  2.566 0.90007
##    time6:siteC            2.442 1.019 324.8   0.437  4.447 0.01716   *
##    time12:siteC           0.009 0.788 299.7  -1.541  1.559 0.99073
##    time1:siteD           -0.844 1.407 326.8  -3.613  1.925 0.54914
##    time6:siteD           -0.253 1.183 328.2   -2.58  2.075 0.83102
##    time12:siteD          -1.401 0.902 296.3  -3.176  0.374 0.12138
##    time1:siteE           -1.128 1.271   328  -3.628  1.372 0.37533    
##    time6:siteE            0.295 1.056 321.9  -1.783  2.374 0.78007    
##    time12:siteE          -0.098 0.808   300  -1.689  1.492 0.90355    
##    time1:sexmale          0.642 0.831 329.1  -0.992  2.276 0.43988    
##    time6:sexmale          0.301 0.692 325.7  -1.061  1.662 0.66431    
##    time12:sexmale         0.267 0.531 299.1  -0.777  1.311 0.61526    
##    time1:arm2            -1.588 0.832 328.4  -3.224  0.049 0.05722   .
##    time6:arm2            -2.107 0.692 323.6  -3.468 -0.746 0.00251  **
##    time12:arm2           -0.856  0.53 299.7    -1.9  0.188 0.10757    
##    time1:age67            0.257 0.052 331.8   0.155   0.36 < 1e-04 ***
##    time6:age67            0.037 0.043   327  -0.048  0.122 0.39239    
##    time12:age67           0.075 0.033 299.8    0.01  0.141 0.02442   *
##    ---------------------------------------------------------------
##   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
##   Columns lower and upper contain 95% pointwise confidence intervals for each coeffic
##   Model-based standard errors are derived from the observed information (column se).
##   Degrees of freedom were computed using a Satterthwaite approximation (column df).
```

Question 7

Yes! The results suggest that this trial brings sufficient evidence that one type of surgery is better than the other, for the mean Oxford score at 2 years. We estimate that, in average, the change in Oxford score at 2 years is 2.267 (95-CI=[0.81; 3.724], p=0.002) points lower for patients who receive the surgery of arm 2 than for those receiving the surgery of arm 1. Hence, there is evidence that arm 1 is the better. Because, of randomization, this mean difference of 2.267 has two interpretations; either a difference "in average" (as in the previous sentence, the so-called *marginal* interpretation) or when comparing patients similar for age, sex, study side and baseline Oxford score (the so-called *conditional* interpretation). The *conditional* interpretation is valid under the assumption that the model is correct (e.g., no interaction between arm and age, as no interaction was assumed) whereas the marginal interpretation is valid even if the model is not correct, to a large extent (e.g., if the model is incorrect because an interaction term between arm and age exists and was not included in the model).

Question 8

The 95% confidence interval is [0.81,3.724]. So, we cannot rule out that the difference is less than one point. This is very small. Even the estimated value 2.267 is not very large, when looking at the patient to patient variability in the change score at 24 months, when considering patients of similar age, study site, gender, surgery arm and baseline score. The standard deviation that quantifies this variability is estimated as 6.67 (see `sigma.24` in the output of the software). This means that, in average, patients who receive the surgery of arm 1 are doing better than those receiving the surgery of arm 2, but that a large proportion of patients receiving the surgery of arm 2 will anyway have a better score at 24 months than many patients receiving the surgery of arm 1.

Question 9

Question 9.a

The interpretation of the default output seen above was easy because the timepoint of interest was the reference level for the factor variable time. This was important because of the interaction terms between `time` and `arm`. If we want an equally easy read of the results to estimate the between-arm difference in mean change score at earlier timepoints, we can simply change the reference level and re-fit the model.

```
long$time6 <- relevel(long$time,ref="6")
#--- fit MMRM --------------------
lmmfit6 <- lmm(change~Oxford.pre22*time6 + site*time6
               + sex*time6 + arm*time6 + age67*time6,
               repetition = ~time6|id,
               structure = "UN", data = long)
```

```
## Warning in .lmmNormalizeData(as.data.frame(data)[unique(stats::na.omit(var.all))], :
```

```
summary(lmmfit6, print=FALSE)$mean["arm2",] # print only the line of interest
```

```
##      estimate    se statistic  df lower upper null  p.value
## arm2    -4.37 0.778     -5.62 332  -5.9 -2.84    0 4.04e-08
```

So, we estimate that, in average, the change in Oxford score at 6 months is 4.37 (95-CI=[2.84; 5.9], p=0.002) points lower for patients who receive the surgery of arm 2 than for those receiving the surgery of arm 1. Hence we estimate that the between-arm difference is larger at the earlier timepoint of 6 months. Although changing the reference level is our *favorite trick*, we could have read this from the previous output. Indeed, -2.267 –2.107 = 4.37 (see lines with `arm2` and `time6:arm2` in the output above). However, we could not have read the confidence interval and p-value from the previous output. But, from the previous output, we could instead see that the treatment effect, i.e., the between-arm difference in change score, is estimated significantly larger at 6 months than at 24 months (difference is 2.107, 95-CI=[0.746;3,468], p=0.0025). In other words, there is evidence that one surgery is better than the other and that the superiority of this surgery is larger at 6 months than at 24 months. We can proceed similarly with other time points (e.g., 1 or 12 months).

Question 9.b

The model did not show evidence that age was associated with the change score at 24 months (p=0.074). We can even say that if an association exists, we are confident that it is rather small. Indeed, the confidence interval for the difference in mean change score at 24 months, when comparing two patients, one 10 years older than the other, both patients being similar for sex, arm, etc…, is [-1.74,0.08]. Because the results suggest that the association between the outcome and age is not very large, the gain in power obtained by adjusting on age was probably very modest. A similar conclusion applies to sex.

Question 10

A simpler analysis would have been to use a complete case analysis with a simple ANCOVA model.

Question 10.a

```
nrow(d)
```

```
## [1] 346
```

```
dCCA <- d[!is.na(d$Oxford.24),]
dCCA$change <- dCCA$Oxford.24 - dCCA$Oxford.pre
nrow(d)-nrow(dCCA)
```

```
## [1] 33
```

This analysis would excludes 33 patients (as already seen in question 4)

Question 10.a

```
fitANCOVA <- lm(change~Oxford.pre + site + sex + arm + age,data=dCCA)
summary(fitANCOVA)$coef["arm",] # print only the line of interest
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -2.294516560  0.744607827 -3.081510126  0.002248068
```

```
confint(fitANCOVA)["arm",]      # print only the line of interest
```

```
##      2.5 %     97.5 %
## -3.7597545 -0.8292787
```

The result of this simple complete case ANCOVA analysis is very similar! This is somewhat reassuring. We estimate that, in average, the change in Oxford score at 2 years is 2.295 (95-CI=[0.829; 3.760], p=0.002) points lower for patients who receive the surgery of arm 2 than for those receiving the surgery of arm 1. **Reminder**: the result of the main analysis seen at Question 7 was 2.267 (95-CI=[0.81; 3.724], p=0.002).

This is not very surprising because:

- not so many patients were excluded using the complete case analysis (only 9.5%)
- the missing completely at random assumption seems reasonable in the context of this trial
- the two methods of analysis are equivalent with complete data (i.e., we would have had the exact same results, if we had run the analysis with a data set without any missing data)

Question 11

An even simpler analysis would have been to use a complete case analysis with a t-test.

```
t.test(change~arm,data=dCCA)
```

```
##
##  Welch Two Sample t-test
##
## data:  change by arm
## t = 1.9088, df = 307.3, p-value = 0.05723
## alternative hypothesis: true difference in means between group 1 and group 2 is not e
## 95 percent confidence interval:
##  -0.05182871  3.40764741
## sample estimates:
## mean in group 1 mean in group 2
##        19.70323        18.02532
```

Here we estimate that, in average, the change in Oxford score at 2 years is of 1.68 points lower for patients who receive the surgery of arm 2 than for those receiving the surgery of arm 1 (95%-CI=[-0.05, 3.41], p=0.057). The results is not statistically significant.

However, this is not the recommended approach, as it does not leverage the information contained in the baseline variables (hence it is less powerful; note the wider confidence interval, width is 3.41-(-0.05)=3.46 vs 2.91=3.724-0.81 using the MMRM of the main analysis).

Question 12

Many researchers wonder whether they should define their primary outcome as the score at end of follow-up or as the change score at end of follow-up. The previous questions consider the later. We now perform the MMRM analysis using the former.

```
lmmfitOx <- lmm(Oxford~Oxford.pre22*time + site*time
              + sex*time + arm*time + age67*time,
              repetition = ~time|id,
              structure = "UN", data = long)
```

```
## Warning in .lmmNormalizeData(as.data.frame(data)[unique(stats::na.omit(var.all))], :
```

```r
summary(lmmfitOx, print=FALSE)$mean["arm2",]
```

```
##      estimate    se statistic  df lower upper null p.value
## arm2    -2.27 0.741     -3.06 316 -3.72 -0.81    0  0.0024
```

We get the exact same results! This is not so surprising if we think about it, because we adjust for the baseline score. Comparing the score at 24 months or the change score at 24 months is equivalent, when comparing patients who have the same baseline score. It does not matter whether we fit the model using the score or the change score at 24 months, as long as we adjust for the baseline score.

Using the score or the change score at 24 moths is however different, when we do not adjust for the baseline score, as e.g., when using a simple t-test with a complete case analysis. That is another reason to not like an analysis unadjusted for the baseline score (on top of the power gain argument).

```r
t.test(Oxford.24~arm,data=dCCA)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Oxford.24 by arm
## t = 3.3505, df = 286.43, p-value = 0.000915
## alternative hypothesis: true difference in means between group 1 and group 2 is not e
## 95 percent confidence interval:
##  1.064864 4.097652
## sample estimates:
## mean in group 1 mean in group 2 
##        42.61290        40.03165
```

Here the result is different from that of question 11. It is now significant!