Final Projects for “Introduction to Statistical Modelling in Life Sciences” (ISMLS)

Søren Højsgaard Created: November 11, 2011

The final project report has to be delivered no later than [1]

Monday, December 5. at 9:00 AM

- There will be no grades of your report; it is either pass or fail.
- There will be an external examiner evaluating the reports.
- The project reports must be handed in on an individual basis.
- The project reports must be no more than 30 pages long (but preferably shorter).
- Please write the report in a 12pt font.
- Please remember that this is a statistics course; not an R-course. This means that we are interested in that you document that you have understood the ideas of statistical modelling and inference. We are not particularly interested in your R-code.
- If you wish to include R-code in your report (and you are welcome to do so), please put the R-code in an appendix.
- Please think carefully about the amount of R-output that you present in the report.

[1] The document should be delivered either as a Word document or a PDF-file to Søren Højsgaard (e-mail: sorenh@agrsci.dk). Make sure that your name appears on the document.

Contents

TOPICS: random regression model; variance component model; non–linear model

1 Milk yield of dairy cows

Use the `milkman` dataset in the `doBy` package for this exercise. The overall purpose of the exercise is to model milk yield curves and to see if there are significant differences in various aspects of cows milk yield depending on their race and lactation number.

```
library(doBy)
data(milkman)
```

It is imperative that you treate lactation number as a factor:

```
milkman$lactno <- factor(milkman$lactno)
```

1.1 Describing data and the overall questions

1. Provide a description of data and the overall questions (see above and the documentation of the dataset) in your own words.

1.2 Manipulating and summarizing data

1. Create a dataframe `mm2` which contains the total daily milk yield for each cow-lactation. For the analyses that follows, the dataframe must also contain information about race and lactation number.

Solution:

```
mm2 <- summaryBy(my~cowlact+dfc, data=milkman, FUN=sum, id=~race+lactno,keep=T)
head(mm2)
```

```
  cowlact dfc     my race lactno
1  0263.3   1     NA  RDM      3
2  0263.3   2     NA  RDM      3
3  0263.3   3 21.595  RDM      3
4  0263.3   4 25.613  RDM      3
5  0263.3   5 29.600  RDM      3
6  0263.3   6 28.695  RDM      3
```

2. One common way of summarizing the milkyield throughout a lacatation is by calculating the 305-day yield which is the total milkyield from the first 305 days of a lactation. Notice: Not all cows in `milkman` have been milked for 305 days.

   Therefore you should create a dataframe `mm3` consisting of those cow-lactations from `mm2` for there are there are recodings of milk yield at least up to day 305. Recordings later than day 305 should be deleted from `mm3`.

Solution:

```
ss  <- summaryBy(dfc~cowlact, data=mm2, FUN=max)
cc  <- subset(ss, dfc.max>=305, select=cowlact)$cowlact
mm3 <- subset(mm2, cowlact %in% cc & dfc<=305)
```

3. Produce a 3-by-3 table which shows the number of cow-lactations in `mm3` for each combination of race and lactation number. (Hint: The functions `summaryBy()` and `xtabs()` could be your friends.)

Solution:

```
xx <- summaryBy(cowlact~race+lactno, data=mm3, FUN=function(x){c(N=length(unique(x)))})
xtabs(cowlact.N~race+lactno, data=xx)
```

```
          lactno
race        1  2  3
  RDM      34 12  6
  Holstein 43 15  3
  Jersey   35 18  9
```

4. For each cow–lactation in `mm3`, calculate the total milkyield over the first 305 days. Put this information into a dataframe `mm4` which should also contain information on the race and lactation number of each cow–lactation.

Solution:

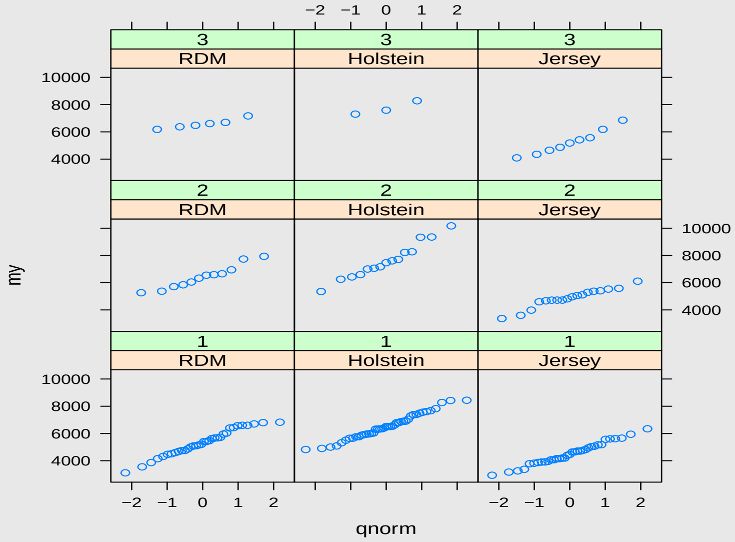
```
mm4 <- summaryBy(my~cowlact, data=mm3, id=~race+lactno,keep=T, FUN=sum, na.rm=T)
head(mm4)
```

```
  cowlact       my     race lactno
1  0263.3 7171.874      RDM      3
2  0266.3 7592.683 Holstein      3
3  0287.3 6608.071      RDM      3
4  0297.2 7735.381      RDM      2
5  0301.2 6652.673      RDM      2
6  0302.2 6943.766      RDM      2
```

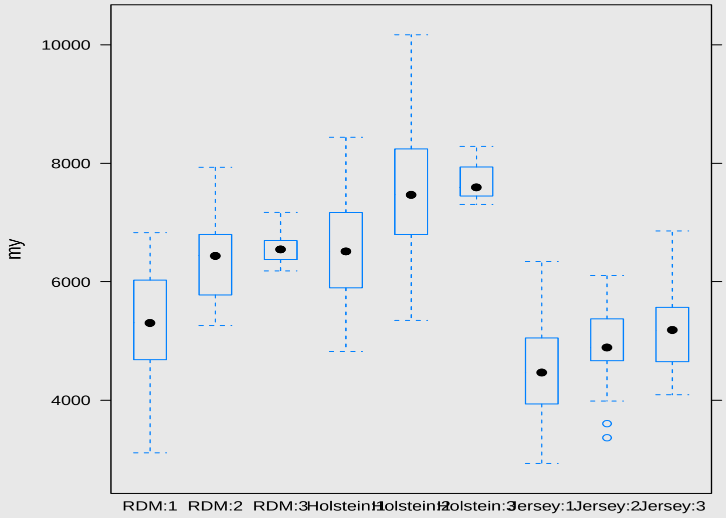
5. Create graphical display of the distribution of the 305-day yield for each combination of race and lactation number. (Hint: Your friends could be the functions `histogram()`, `densityplot()`, `bwplot()` and `qqmath()` from the `lattice` package.

Solution:

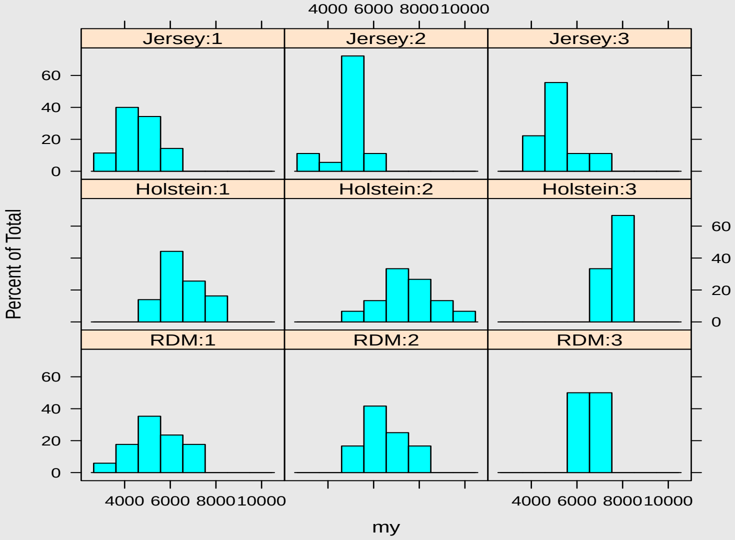
```
print(qqmath(~my|race+lactno, data=mm4))
```



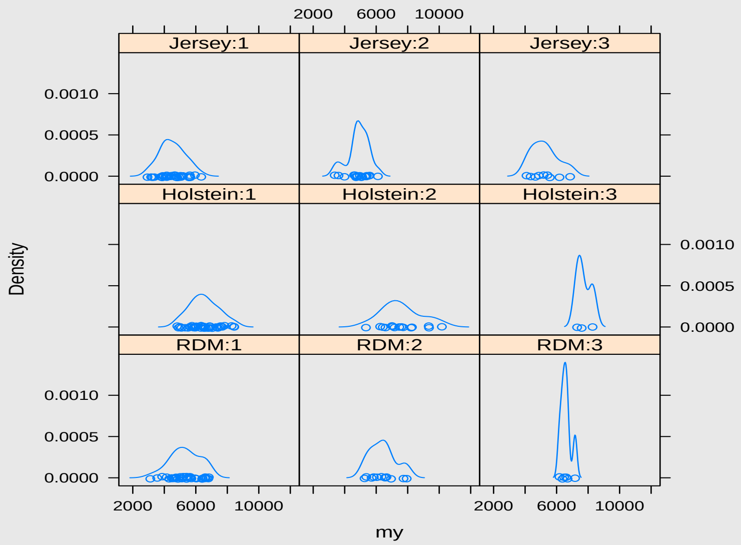
```
print(bwplot(my~race:lactno, data=mm4))
```



```
print(histogram(~my|race:lactno, data=mm4))
```



```
print(densityplot(~my|race:lactno, data=mm4))
```



6. Based on the dataframe `mm4`, create a dataframe `mm5` which contains the mean and standard deviation of the total milkyield for each combination of race and lactation number.

Solution:

```
mm5 <- summaryBy(my~race+lactno, data=mm4, FUN=c(mean,sd))
mm5
```

```
      race lactno  my.mean     my.sd
1      RDM      1 5325.897  968.6476
2      RDM      2 6413.914  842.7440
3      RDM      3 6584.641  339.0723
4 Holstein      1 6519.848  933.8506
5 Holstein      2 7597.565 1300.0076
6 Holstein      3 7726.765  503.8193
7   Jersey      1 4517.323  816.6155
8   Jersey      2 4867.849  691.2815
9   Jersey      3 5241.450  883.4170
```

1.3 Analzying the 305–day yield

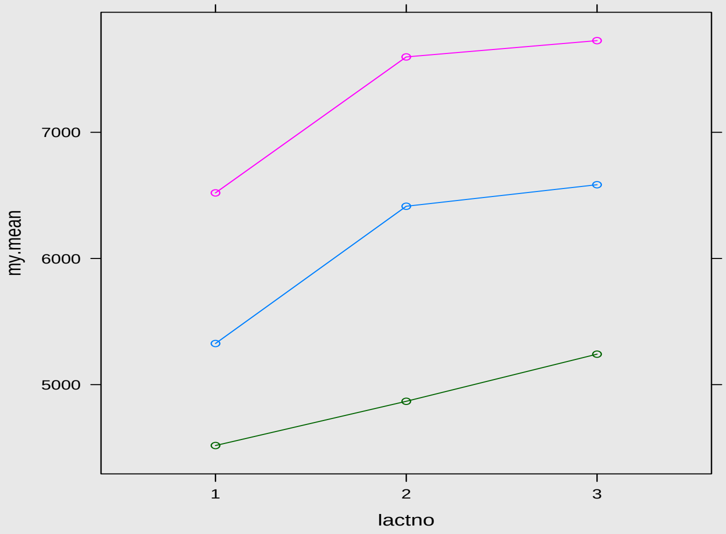
1. One way of analyzing lactation data is by considering the total 305-day yield as response and the factors race and lactno as explanatory variables. Initially we consider the possibility of an interaction between lactation number and race.

   Write down such a model in precise mathematical terms (writing the model in R-syntax is not valid). Write down in your own words the assumptions behind such a model

2. The question of an interaction can be investigated graphically: Based on the data `mm5` plot the mean yield against lactation number for each level of race. Does this plot suggest suggest an interaction. (Hint: Your friend is `xyplot()`).

Solution:

```
print(xyplot(my.mean~lactno, groups=race, data=mm5, type='b'))
```

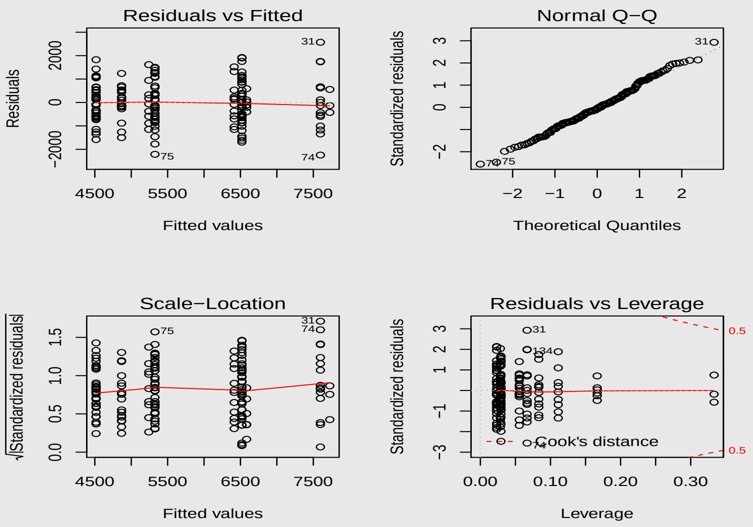


3. Fit this model and investigate graphically whether there is evidence that the model does not fit well to data. What is your conclusion?

Solution:

```
lm1 <- lm(my~race*lactno, data=mm4)
```

```
par(mfrow=c(2,2))
plot(lm1)
```



4. Test whether there is a significant interaction between lactation number and race

Solution:

```
anova(lm1)
```

```
Analysis of Variance Table

Response: my
             Df    Sum Sq  Mean Sq F value    Pr(>F)
race          2 138299492 69149746  83.674 < 2.2e-16 ***
lactno        2  30868350 15434175  18.676 4.835e-08 ***
race:lactno   4   4327237  1081809   1.309    0.2687
Residuals   166 137186063   826422
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Create a model without the interaction

Solution:

```
lm2 <- update(lm1, .˜.-race:lactno)
summary(lm2)
```

```
Call:
lm(formula = my ˜ race + lactno, data = mm4)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2304.1  -554.0   -72.3   575.2  2763.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5416.0      136.2  39.773  < 2e-16 ***
raceHolstein   1177.9      172.9   6.813 1.58e-10 ***
raceJersey    -1077.0      172.1  -6.260 3.04e-09 ***
lactno2         812.3      161.5   5.028 1.25e-06 ***
lactno3        1029.6      234.2   4.397 1.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 912.4 on 170 degrees of freedom
Multiple R-squared: 0.5445, Adjusted R-squared: 0.5338
F-statistic: 50.81 on 4 and 170 DF,  p-value: < 2.2e-16
```

6. Create all pairwise comparisons of lactation number and of race. Are all pairwise differences statistically significantly different from zero? Provide confidence intervals for the pairwise differences. Comment on your finding! (Hint: Your friend is the `glht()` function in the `multcomp` package).

Solution:

```
library(multcomp)
ddd1 <- glht(lm2, mcp(race='Tukey'))
summary(ddd1, test=univariate())
```

```
 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = my ˜ race + lactno, data = mm4)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Holstein - RDM == 0      1177.9      172.9   6.813 1.58e-10 ***
Jersey - RDM == 0       -1077.0      172.1  -6.260 3.04e-09 ***
Jersey - Holstein == 0  -2254.9      166.4 -13.550  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p values reported)
```

```
confint(ddd1)
```

```
 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = my ˜ race + lactno, data = mm4)
```

```
Quantile = 2.3641
95% family-wise confidence level


Linear Hypotheses:
                       Estimate   lwr        upr
Holstein - RDM == 0     1177.8749   769.1890  1586.5609
Jersey - RDM == 0      -1077.0459 -1483.7991  -670.2927
Jersey - Holstein == 0 -2254.9208 -2648.3507 -1861.4910
```

```
ddd2 <- glht(lm2, mcp(lactno='Tukey'))
summary(ddd2, test=univariate())
```

```
 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = my ~ race + lactno, data = mm4)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    812.3      161.5   5.028 1.25e-06 ***
3 - 1 == 0   1029.6      234.2   4.397 1.93e-05 ***
3 - 2 == 0    217.4      255.6   0.850    0.396
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p values reported)
```

```
confint(ddd2)
```

```
 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = my ~ race + lactno, data = mm4)

Quantile = 2.3474
95% family-wise confidence level


Linear Hypotheses:
           Estimate  lwr       upr
2 - 1 == 0  812.2653  433.0825 1191.4481
3 - 1 == 0 1029.6434  479.9600 1579.3267
3 - 2 == 0  217.3781 -382.6573  817.4134
```

7. Estimate the population means of the milk yield for each race. (Hint: Your friend is `popMeans()` in the `doBy` package). Comment on the result.

Solution:

```
popMeans(lm2, effect='race')
```

```
  beta0 Estimate Std.Error  t.value  DF Pr(>|t|)    Lower    Upper
1     0 6029.936  138.6938 43.47662 170        0 5756.152 6303.720
2     0 7207.811  136.6040 52.76427 170        0 6938.152 7477.469
3     0 4952.890  124.6690 39.72831 170        0 4706.791 5198.989
```

1.4 Modelling with Woods curves – by extracting features

In the following we shall take a different approach, namely modelling the individual milk yield curves using a Woods function

$$y_t = a_0 t^b e^{ct}$$

This is by no means the best possible curve for lactation data, but the curve has the property of linearity on a log–scale:

$$z_t = \log y_t = a + b \log t + ct; \quad a = \log a_0$$

We shall call the second form the "log–Woods curve". For later use we notice that the first and second derivatives are

$$z'_t = \frac{b}{t} + c; \quad z''_t = -\frac{b}{t^2}$$

Based on data `mm3` we can fit the log–Woods curve to each cow-lactation. As a help for you, we will show an easy way of doing this using the `lmBy()` function in `doBy`:

```
lmb1 <- lmBy(log(my)~log(dfc)+dfc|cowlact, data=mm3)
lmb1
```

```
lmBy(formula = log(my) ~ log(dfc) + dfc | cowlact, data = mm3)
```

The `lmBy()` function simply partitions data into different strata, here according to `cowlact`, and fits the specified model on the left hand side of the bar (the "|") to each stratum. Regression coefficients are obtained with

```
head(coef(lmb1))
```

```
       (Intercept)  log(dfc)          dfc
0263.3    2.421351 0.4276059 -0.009034712
0266.3    1.306783 0.6229534 -0.007213810
0287.3    2.820300 0.1799657 -0.004106817
0297.2    2.944411 0.1896532 -0.004086665
0301.2    2.583349 0.3370936 -0.007853868
0302.2    2.516251 0.2340775 -0.003249289
```

```
colnames(coef(lmb1))
```

```
[1] "(Intercept)" "log(dfc)"    "dfc"
```

To each stratum there is a race and lactation number, and we will for subesequent analyses need this information. To make this information available we can do as follows:

```
lmb1 <- lmBy(log(my)~log(dfc)+dfc|cowlact, data=mm3, id=~race+lactno)
lmb1
```

```
lmBy(formula = log(my) ~ log(dfc) + dfc | cowlact, data = mm3,
    id = ~race + lactno)
```

and we can the retrieve this information with

```
head(coef(lmb1, augment=TRUE))
```

```
       (Intercept)  log(dfc)          dfc     race lactno cowlact
0263.3    2.421351 0.4276059 -0.009034712      RDM      3  0263.3
0266.3    1.306783 0.6229534 -0.007213810 Holstein      3  0266.3
0287.3    2.820300 0.1799657 -0.004106817      RDM      3  0287.3
0297.2    2.944411 0.1896532 -0.004086665      RDM      2  0297.2
0301.2    2.583349 0.3370936 -0.007853868      RDM      2  0301.2
0302.2    2.516251 0.2340775 -0.003249289      RDM      2  0302.2
```

```
colnames(coef(lmb1, augment=TRUE))
```

```
[1] "(Intercept)" "log(dfc)"    "dfc"         "race"        "lactno"
[6] "cowlact"
```

1. Based on the mathematical expression for the log–Woods curve, find the time t_{max} for the maximum milk yield expressed in terms of the parameters a, b and c. (Hint: You will have to find the derivative; set the derivative equal to zero and solve for t).

Solution: A direct calculation shows that $z_t' = \frac{b}{t} + c$ and solving $z_t' = 0$ for t gives the time for maximum yield as $t_{max} = -b/c$.

2. Calculate for each cow-lactation t_{max} from the estimated regression coefficients.

Solution:

```
bb1 <- coef(lmb1, augment=TRUE)
bb1$t.max <- -bb1[,2]/bb1[,3]
```

3. Investigate - by a statistical analysis - whether there is evidence that the time for the maximum milk yield depends on race and lactation number. You do so by going through the steps above that you made when analyzing the 305–day milk yield. Simply writing "Yes, there is an effect" or "no, there is no effect" is a completely inadequate answer to this question. (Hint, there may be some outliers in data that you may want to exclude from your analysis).

4. Recall that we can think of $z_t' \approx z_{t+1} - z_t$ as the rate of change in milk yield when going from day t to day $t + 1$. There is another name for such a quantity: The velocity (so at t_{max}, the velocity is zero). In a similar way we can think of the second derivative z_t'' as $z_t'' \approx z_{t+1}' - z_t'$, i.e. the rate of change of velocity. There is another name for such a quantity: The acceleration. Since $z_t'' = -b/t^2$, the parameter b has a natural interpretation: it is the (negative) acceleration of the milk yield curve at day 1 after calving; we could call it the "day1-acceleration". Investigate - by a statistical analysis - whether there is evidence that day1-acceleration depends on race and lactation number.

1.5 Modelling with Woods curves – by random effects models

We continue working on log–Woods curves but now from a random effects model perspective.

1. Following up on the lectures, consider a random regression coefficients model made so that there are interactions between `dfc` (and `log(dfc)` and the race and lactation number for the systematic effects. Fit such a model, and produce diagnostic plots. Comment on the results. (Hint: Remember to set `REML=FALSE` when calling `lmer()`).

2. Simplify the model by removing non–significant terms. In this process you should continuously convince yourself that the model you consider is adequate. (Hint: When

comparing models, you should use the `anova()` function. (The `KRmodcomp()` function is likeliy to cause you trouble because the current implementation is not very efficient)).

3. Once you have reached a final model you are faced with the challenge of interpreting the parameter estimates of the model. Please do so!

Solution:

```
lme1 <- lmer(log(my)~race+log(dfc)+dfc+race:log(dfc)+race:dfc+(1|cowlact), data=mm3,
               REML=FALSE)
```

```
lme2 <- update(lme1, .~.-race)
anova(lme1, lme2)
```

```
Data: mm3
Models:
lme2: log(my) ~ log(dfc) + dfc + (1 | cowlact) + race:log(dfc) + race:dfc
lme1: log(my) ~ race + log(dfc) + dfc + race:log(dfc) + race:dfc +
lme1:     (1 | cowlact)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lme2  9 9902.5 9982.1 -4942.3
lme1 11 9884.0 9981.4 -4931.0 22.454      2  1.331e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 Sitka Spruce trees

TOPICS: Generalized linear models

In a study on the effect of ozone on the growth of trees, 54 sitka spruce trees were grown in an atmosphere enriched with ozone, and 25 trees were grown under normal conditions. The 79 trees of the same age were randomly assigned to the two groups at the beginning of the growth experiment.

In 1989, the second year of growth, the size of the trees was measured at roughly monthly intervals.

The data set `sitka89` in our data library `LiSciData` contains the size of the trees measured as $\log(h * d^2)$ where h is the height of a tree and d its diameter, the time of measurement in days after January, 1st, 1988, an identification number for each tree and the indicator of the treatment.

1. Plot the growth curves for all trees and the mean of the growth curves for the ozone and the control group

2. Propose and fit a model assuming that all observations are independent. Use the time-variable as a factor.

3. Fit a model taking into account the repeated measurements on the same tree.

4. Compare the parameter estimates and their standard errors of the two fits.

5. Test the null-hypothesis that there is no effects of ozone.

6. Provide the effect-estimates for each tree.

7. Make some residual plots.

3 Esophageal cancer

1. The data set `esoph` contains records from a case-control study of esophageal cancer. For 88 combinations of age , alcohol consumption and tobacco consumption the number of cases and controls are given.

   ```
   data(esoph)
   ```

2. Make some descriptive plots of data.

3. Formulate a model for your data.

4. Fit the model with interactions between all three predictors.

5. Eliminate factors to simplify the model as far as is possible.

6. All three factors are ordered. Convert the factors to numerical variables, for example by using the `recodeVar()` function of the package `doBy`.

7. Can the model be simplified using the numerical representation of the predictors?

8. Analyse the residuals of your final model.