Topics in Statistical Sciences 2 – Exam exercise 3

Søren Højsgaard and Torben Tvedebrink

Version: 23/11/2017 13:37

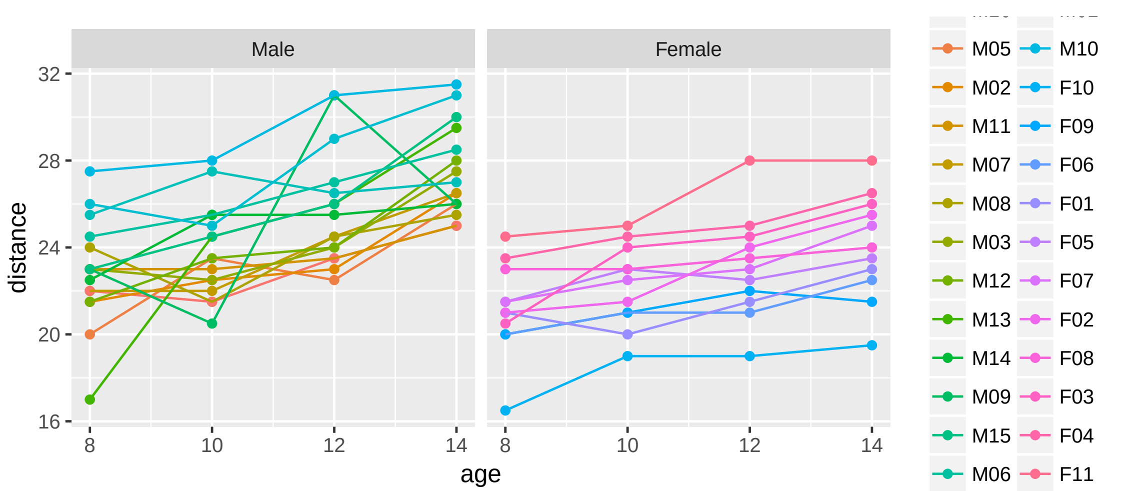
This exercise is about the generalized estimating equations as discussed in lectures 7–9 of Topics in Statistical Sciences 2. During the oral exam you will have 20 min to present the exercise. You decide what topics to cover and how to present them, however, we will ask questions to any part of the covered curricula, exercise and presentation.

1 The orthodont data

```
data(Orthodont, package="nlme")
ort <- Orthodont
head(ort)
```

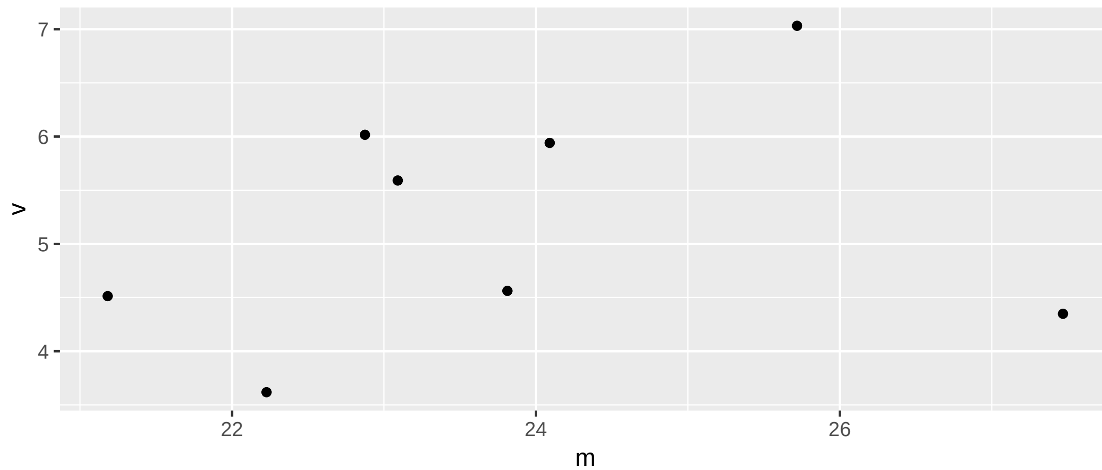
```
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

```
library(ggplot2)
library(tidyverse)
qplot(age, distance, group=Subject, data=ort, color=Subject) + geom_path() + facet_grid(~Sex)
```



There is clear indication of correlation within each subject, but no clear indication of variance heterogeneity.

```r
osum <- ort %>% group_by(Sex, age) %>% summarize(m=mean(distance), v=var(distance))
qplot(m, v, data=osum)
```



2 Fitting a GEE model

This exam exercise consists of the following components

1. Write up (or down) the estimating equations for estimating β in a setting there $g(\mu) = \eta = X\beta$ in a GEE "model" where 1) the variance function is $V(\mu) = 1$, 2) the link function is $g(\mu) = \mu$, 3) the working correlation $R(\alpha)$ is the identity matrix, and 4) all weights (w_i) are 1.

2. Implement the iterative scheme for estimating β in a sligtly extended setting of that above in a function called `fit_normal_gee`: You must allow for two type of working correlation matrices: `"independence"` and `"unstructured"`. Notice: We reserve ourselves the right to examine your code at the exam, and if your code simply consist of calling `geeglm` or similar functions, then this will be regarded as cheating.

   ```r
   fit_normal_gee <- function(formula, id, corstr="independence", phi=NULL){
     # Your code goes here
   }
   ```

   Above, `id` is the vector identifying clusters (as in `geeglm`)

3. The code should run at the time of exam in case we provide you with an extra dataset.

4. Apply the function to the Orthodont data in two settings: Two parallel regression lines and two non–parallel regression lines.

5. Do the models fit well to data? You may want to consider, among other things, the residuals and the fitted values.

6. Test the smaller of the two models against the larger.

7. Feel free to benchmark your code against `geeglm`

2.1 Requirements to your implementation

Regarding `fit_nromal_gee`, the minimal requirements are

Input:

- formula: a model formula
- corstr: Either `"independence"` or `"unstructured"`.
- phi: the dispersion parameter. If NULL, phi is estimated from data; if phi is a single positive number then this value is used; otherwise and error is signaled.

Output: A list with the following components:

- coef: regression coefficients
- vcov: the variance-covariance matrix of the regression coefficients
- p: the number of regression coefficients
- resid: the pearson residuals
- fit: the fitted values

2.2 Hints to your implementation

The hints to exam exercise 1 should help you a lot.