Complex Data - lab5

Stanisław Wilczyński

Muscatine data set

```
musc.dat <- read.table("../data/muscatine.txt",na.strings=".", as.is=T)
names(musc.dat) <- c("id", "gender", "baseage", "age", "occasion", "y")
musc.dat$cage <- musc.dat$age - 12
```

Looking at the summary on the p.1 we can clearly see some patterns:

- for both males and females percentage of obese increases in time for children aged 5-9
- for both males and females percentage of obese is stable in time for children aged 9-11
- for both males and females percentage of obese decreases in time for children aged 11-15
- the percentage of obese female is greater then percentage of obese male for almost all time points and ages

```
## Model fit
musc.gee <- gee(y~gender*cage + gender*I(cage^2),
  id=id,
  family="binomial",
  data=musc.dat,
  corstr="unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)           gender             cage        I(cage^2) 
##     -1.213042282      0.096202285      0.032413849     -0.018328452 
##      gender:cage gender:I(cage^2) 
##     -0.004269978      0.003724784
```

```
## Summary of the output
summary(musc.gee)
```

```
## 
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998) 
## 
## Model:
##  Link:                      Logit 
##  Variance to Mean Relation: Binomial 
##  Correlation Structure:     Unstructured 
## 
## Call:
## gee(formula = y ~ gender * cage + gender * I(cage^2), id = id, 
##     data = musc.dat, family = "binomial", corstr = "unstructured")
## 
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max 
## -0.2546318 -0.2306407 -0.2078458 -0.1602929  0.8899486 
## 
## 
```

```
## Coefficients:
##                      Estimate  Naive S.E.     Naive z Robust S.E.
## (Intercept)      -1.213290962 0.050018892 -24.2566543 0.051238583
## gender            0.107476876 0.070256905   1.5297696 0.072088777
## cage              0.039761374 0.013631563   2.9168609 0.013692772
## I(cage^2)        -0.017732142 0.003563414  -4.9761663 0.003513991
## gender:cage       0.004886449 0.018957440   0.2577589 0.018942421
## gender:I(cage^2)  0.003346875 0.004952592   0.6757826 0.004835746
##                     Robust z
## (Intercept)      -23.6792451
## gender             1.4908961
## cage               2.9038221
## I(cage^2)         -5.0461548
## gender:cage        0.2579633
## gender:I(cage^2)   0.6921114
##
## Estimated Scale Parameter:  0.9967266
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5715344 0.2497781
## [2,] 0.5715344 1.0000000 0.3020448
## [3,] 0.2497781 0.3020448 1.0000000
```

For this model we can see that included interactions yielded very small coefficients compared to main effects. Therefore, it is reasonable to fit a model without interactions.

```
## Model fit - no interactions
musc.gee.noInt <- gee(y~gender + cage + I(cage^2),
id=id,
family="binomial",
data=musc.dat,
corstr="unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)      gender        cage   I(cage^2)
## -1.22751283  0.12462968  0.03027391 -0.01643142
```

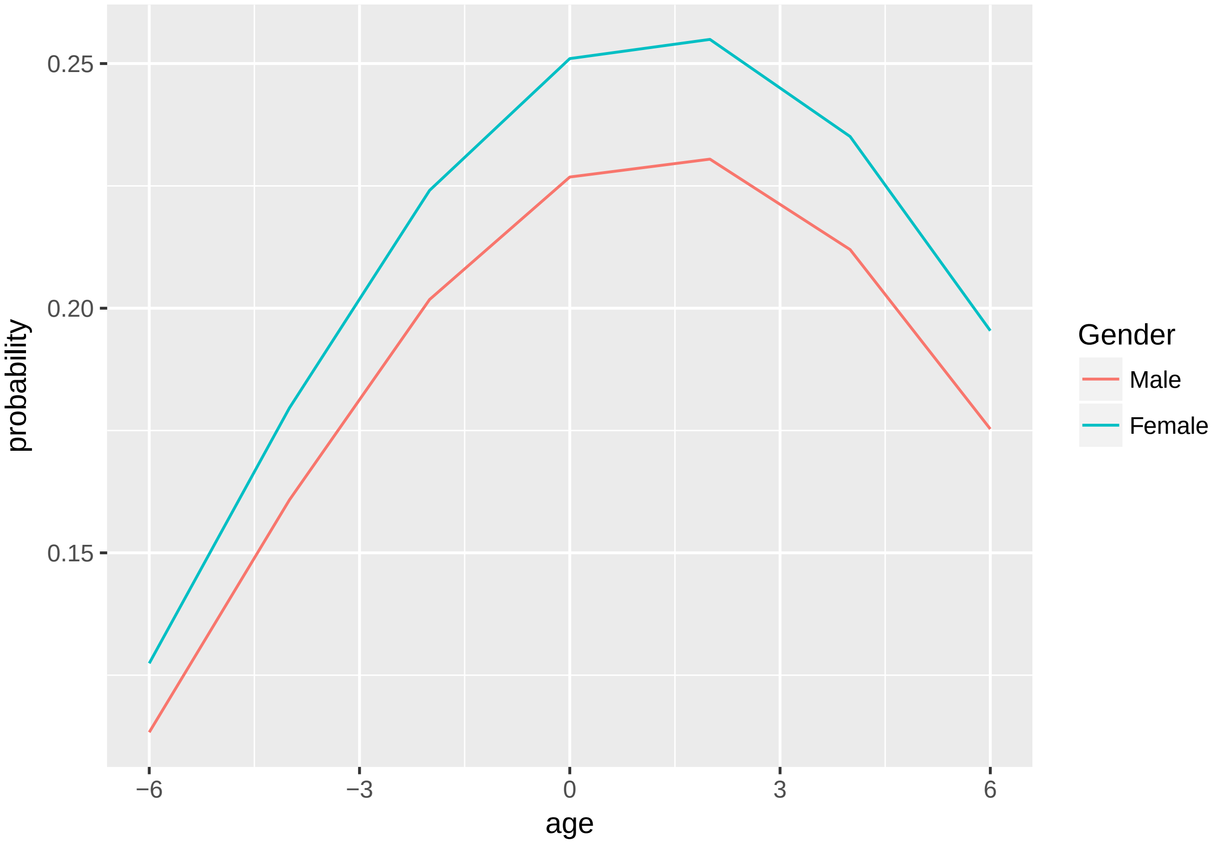
```
## Summary of the output
summary(musc.gee.noInt)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = y ~ gender + cage + I(cage^2), id = id, data = musc.dat,
##     family = "binomial", corstr = "unstructured")
```

```
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.2549353 -0.2304687 -0.2119811 -0.1608183  0.8866698
##
##
## Coefficients:
##                Estimate  Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept) -1.22640178 0.046523246 -26.361054 0.048163087 -25.463521
## gender       0.13320549 0.059977579   2.220921 0.063012084   2.113967
## cage         0.04239088 0.009467761   4.477392 0.009450856   4.485401
## I(cage^2)   -0.01601166 0.002472635  -6.475546 0.002411279  -6.640320
##
## Estimated Scale Parameter:  0.9965039
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5718216 0.2496181
## [2,] 0.5718216 1.0000000 0.3015565
## [3,] 0.2496181 0.3015565 1.0000000
```

As in the task description we can conclude for our model that young females have higher probability of being obese (1.142 time higher odds ratio) and that quadratic curve fits our data set well. The consequence of neglecting interactions is that the patterns of change in rates of obesity (profiles) do not depend on gender (check plot below).

```
## predictions
musc.sel.dat <- matrix(unlist(expand.grid(unique(musc.dat$gender),
unique(musc.dat$cage))), ncol=2)
musc.sel.dat <- cbind(rep(1,dim(musc.sel.dat)[1]),
musc.sel.dat, musc.sel.dat[,2]^2)
colnames(musc.sel.dat) <- c("Int", "gender","cage", "cage2")
musc.lin.pred <- musc.sel.dat %*% matrix(coef(musc.gee.noInt), ncol=1)
musc.exp.pred <- exp(musc.lin.pred)/(1+exp(musc.lin.pred))
musc.all.pred <- cbind(musc.sel.dat, musc.lin.pred, musc.exp.pred)
musc.all.pred.plot <- as.data.frame(musc.all.pred[,c(2,3,6)])
colnames(musc.all.pred.plot) <- c("gender","age", "probability")
musc.all.pred.plot$gender <- as.factor(musc.all.pred.plot$gender)
ggplot(musc.all.pred.plot, aes(x=age, y=probability, colour = gender)) + geom_line() + scale_colour_dis
                         breaks=c(0,1),
                         labels=c("Male", "Female"))
```



Depress data set

```
depress.dat <- read.table("../data/depress.txt",na.strings=".")
names(depress.dat) <- c("id", "y", "severe", "drug", "time")
depress.gee <- gee(y~severe + drug*time,
  id=id,
  family="binomial",
  data=depress.dat,
  corstr="unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

## (Intercept)      severe        drug        time   drug:time
## -0.02798843 -1.31391092 -0.05960381  0.48241209  1.01744498
```

```
sum.gee <- summary(depress.gee)
sum.gee
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = y ~ severe + drug * time, id = id, data = depress.dat,
##     family = "binomial", corstr = "unstructured")
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.94773674 -0.40645713  0.05226326  0.38927858  0.79975454
##
##
## Coefficients:
##                Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept) -0.02552611  0.1664840 -0.1533247   0.1726392 -0.1478581
## severe      -1.30484850  0.1448724 -9.0068787   0.1450136 -8.9981088
## drug        -0.05438636  0.2261876 -0.2404480   0.2271321 -0.2394481
## time         0.47587182  0.1150534  4.1360955   0.1190418  3.9975178
## drug:time    1.01297603  0.1870636  5.4151419   0.1865407  5.4303205
##
## Estimated Scale Parameter:  0.9823364
## Number of Iterations:  3
##
## Working Correlation
##             [,1]        [,2]        [,3]
## [1,]  1.00000000  0.07393977 -0.02741128
## [2,]  0.07393977  1.00000000 -0.05669559
## [3,] -0.02741128 -0.05669559  1.00000000
```

```
## GLMM
depress.glmer <- glmer(y ~ severe + drug*time + (1|id),
family = binomial,
data=depress.dat)
sum.glmm <- summary(depress.glmer)
sum.glmm
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: y ~ severe + drug * time + (1 | id)
##    Data: depress.dat
##
##      AIC      BIC   logLik deviance df.resid
##   1173.9   1203.5   -581.0   1161.9     1014
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2849 -0.8268  0.2326  0.7964  2.0181
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 0.003231 0.05684
## Number of obs: 1020, groups:  id, 340
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02797    0.16407  -0.170    0.865
## severe      -1.31488    0.15263  -8.615  < 2e-16 ***
## drug        -0.05967    0.22239  -0.268    0.788
## time         0.48274    0.11566   4.174 3.00e-05 ***
## drug:time    1.01817    0.19150   5.317 1.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) severe drug   time
## severe    -0.389
## drug      -0.614 -0.005
## time      -0.673 -0.124  0.524
## drug:time  0.462 -0.121 -0.742 -0.562
```

```
coefs <- cbind(sum.gee$coefficients[,c(1,2)], sum.glmm$coefficients[,c(1,2)])
colnames(coefs) <- c("Coeffcients (gee)", "Std. Err. (gee)", "Coeffcients (glmm)", "Std. Err. (glmm)")
kable(coefs, format = "latex", booktabs=TRUE,
          caption = "Coefficients") %>% kable_styling(latex_options="HOLD_position")
```

Table 1: Coefficients

	Coeffcients (gee)	Std. Err. (gee)	Coeffcients (glmm)	Std. Err. (glmm)
(Intercept)	-0.0255261	0.1664840	-0.0279652	0.1640663
severe	-1.3048485	0.1448724	-1.3148827	0.1526252
drug	-0.0543864	0.2261876	-0.0596721	0.2223938
time	0.4758718	0.1150534	0.4827369	0.1156629
drug:time	1.0129760	0.1870636	1.0181673	0.1915039

As we can see from the table above the differences in coefficients and their standard errors are quite small for these two models.

Task 1

We just have to extract time trends for these two therapies.

- standard treatment: $logit\{P(Y_{ij} = 1|b_{i1})\} = \beta_1 + \beta_2 severe_i + \beta_4 time_{ij} + b_{i1}$
- new treatment: $logit\{P(Y_{ij} = 1|b_{i1})\} = \beta_1 + \beta_2 severe_i + \beta_3 + (\beta_4 + \beta_5) time_{ij} + b_{i1}$

Therefore the difference between old and new treatment is $\beta_3 + \beta_5 time_{ij}$. It means that for $time_{ij} = 0$ the difference is β_3, for $time_{ij} = 1$ the difference is $\beta_3 + \beta_5$ and for $time_{ij} = 2$ the difference is $\beta_3 + 2\beta_5$.

As stated in the task, our main goal was to discover if there is a difference in probability of remission between these tow treatments. We can see that p-value for β_5 is almost zero and we reject the null hypothesis that there is no difference.

Task 2

Due to the problem with finding out what confidence interval should be calculated, we stick with analysis of how the odds ratio differ for used treatments:

- the odds of remission increase by $e^{\beta_4} = 1.62$ with each time period for patients on standard treatment
- the odds of remission increase by $e^{\beta_4 + \beta_5} = 4.49$ with each time period for patients on new treatment

Effect of initial diagnosis:

- for patients diagnosed with severe depression the odds of remission are $e^{\beta_1} = 0.27$ times the odds for the subject with mild depression

Random effect:

- As expected by looking at Table 1., the random effect is almost neglectable: $Var(b) = 0.0032$ and this means that values of b for each observation are very small compared to other variables multiplied by proper betas