Complex Data - lab1

Stanisław Wilczyński

15 May 2018

Task 1 and Task 2

In the first two tasks we compare two different estimators of error variance covariance matrices. We consider unstructured matrices. The first one is REML estimator:

```
lead <- read.table(file = "../data/lead.txt", header = FALSE)
names(lead) <- c("id", "baseline", paste("week", c(1,4,6), sep=""))
lead.uni <- melt(lead, id.vars = c("id"), value.name = "y")
lead.uni <- lead.uni[c(1,3)]
lead.uni <- lead.uni[order(lead.uni$id),]
lead.uni$time <- rep(c(0,1,4,6))
lead.uni$time.cat <- rep(1:4)
lead.cat <- gls(y~factor(time.cat),
                correlation = corSymm(form=~1 | id),
                weights = varIdent(form= ~1 | factor(time.cat)),
                data=lead.uni)
lead.cat.summary <- summary(lead.cat)
lead.cat.sigma <- lead.cat.summary$sigma
covariance.matrix <- getVarCov(lead.cat)
print(covariance.matrix)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 25.210 15.466 15.138 22.986
## [2,] 15.466 58.867 44.029 35.965
## [3,] 15.138 44.029 61.657 33.022
## [4,] 22.986 35.965 33.022 85.494
##   Standard Deviations: 5.021 7.6725 7.8522 9.2463
```

The second one is ML estimator:

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

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 24.706 15.156 14.835 22.525
## [2,] 15.156 57.690 43.148 35.246
## [3,] 14.835 43.148 60.424 32.360
## [4,] 22.525 35.246 32.360 83.782
##   Standard Deviations: 4.9705 7.5954 7.7733 9.1533
```

We can see that they don't differ much. Actually the only distinction when considering summaries for both

models (first using REML estimators, second one using ML estimators) is the residual standard error (or using notation form exercises - σ_{11}). They are 5.0209696 and 4.9704665 respectively. Therefore these variance covariance matrices differ just by multiplication constant.

Task 3 and Task 4

In task 3 and 4 we should choose some assumptions about the structure of the variance covariance matrix from the first task. Unfortunately, at the first glance the matrix does not have any structure at all. The only clearly visible thing is that the variances are getting bigger with time. First we try to test if assumptions of equal variance is plausible. From the unstructured variance covariance matrix it does not look so. To compare the models we use *anova* function, which performs likelihood ratio test. We received following output:

```
lead.cat.new <- gls(y~factor(time.cat),
correlation=corSymm(form= ~1 | id),
weights=varIdent(),
data=lead.uni)
print(getVarCov(lead.cat.new))
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 57.154 24.226 22.812 24.093
## [2,] 24.226 57.154 41.108 26.098
## [3,] 22.812 41.108 57.154 23.499
## [4,] 24.093 26.098 23.499 57.154
##   Standard Deviations: 7.56 7.56 7.56 7.56
```

```
print(anova(lead.cat.new, lead.cat))
```

```
##              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lead.cat.new     1 11 1324.325 1360.384 -651.1625
## lead.cat         2 14 1308.337 1354.231 -640.1687 1 vs 2 21.98756   1e-04
```

We can clearly see that p-value was very small - we reject the null hypothesis that the unstructured model explains as much variability as the simpler one.

Next we choose some better assumption: variances are powers of *time* variable.

```
lead.cat.new <- gls(y~factor(time.cat),
correlation=corSymm(form= ~1 | id),
weights=varPower(form = ~time+1),
data=lead.uni)
print(getVarCov(lead.cat.new))
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 32.233 14.660 18.917 25.833
## [2,] 14.660 44.602 38.422 28.417
## [3,] 18.917 38.422 68.520 32.385
## [4,] 25.833 28.417 32.385 80.222
##   Standard Deviations: 5.6774 6.6785 8.2777 8.9567
```

```
print(anova(lead.cat.new, lead.cat))
```

```
##              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lead.cat.new     1 12 1310.289 1349.627 -643.1447
## lead.cat         2 14 1308.337 1354.231 -640.1687 1 vs 2 5.951938   0.051
```

Here the p-value is just above significance level - we do not reject the null hypothesis and such assumption about the such structure of variance covariance matrix is believable.

Due to the problem with seeing a structure in correlations, we checked many different possible correlations structure. However none of them seemed good enough. For example for parameters *correlation=corCompSymm(form= ∼1 | id), weights=varIdent(form = ∼1 | factor(time.cat))* we got:

```
lead.cat.new <- gls(y~factor(time.cat),
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form = ~1 | factor(time.cat)),
data=lead.uni)
print(getVarCov(lead.cat.new))
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 26.804 19.321 19.961 23.915
## [2,] 19.321 56.362 28.945 34.678
## [3,] 19.961 28.945 60.157 35.827
## [4,] 23.915 34.678 35.827 86.350
##   Standard Deviations: 5.1773 7.5074 7.7561 9.2925
```

```
print(anova(lead.cat.new, lead.cat))
```

```
##              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lead.cat.new     1  9 1311.470 1340.973 -646.7351
## lead.cat         2 14 1308.337 1354.231 -640.1687 1 vs 2 13.13289  0.0222
```

Therefore we reject the null hypothesis - for this case unstructured model is significantly different than structured one.

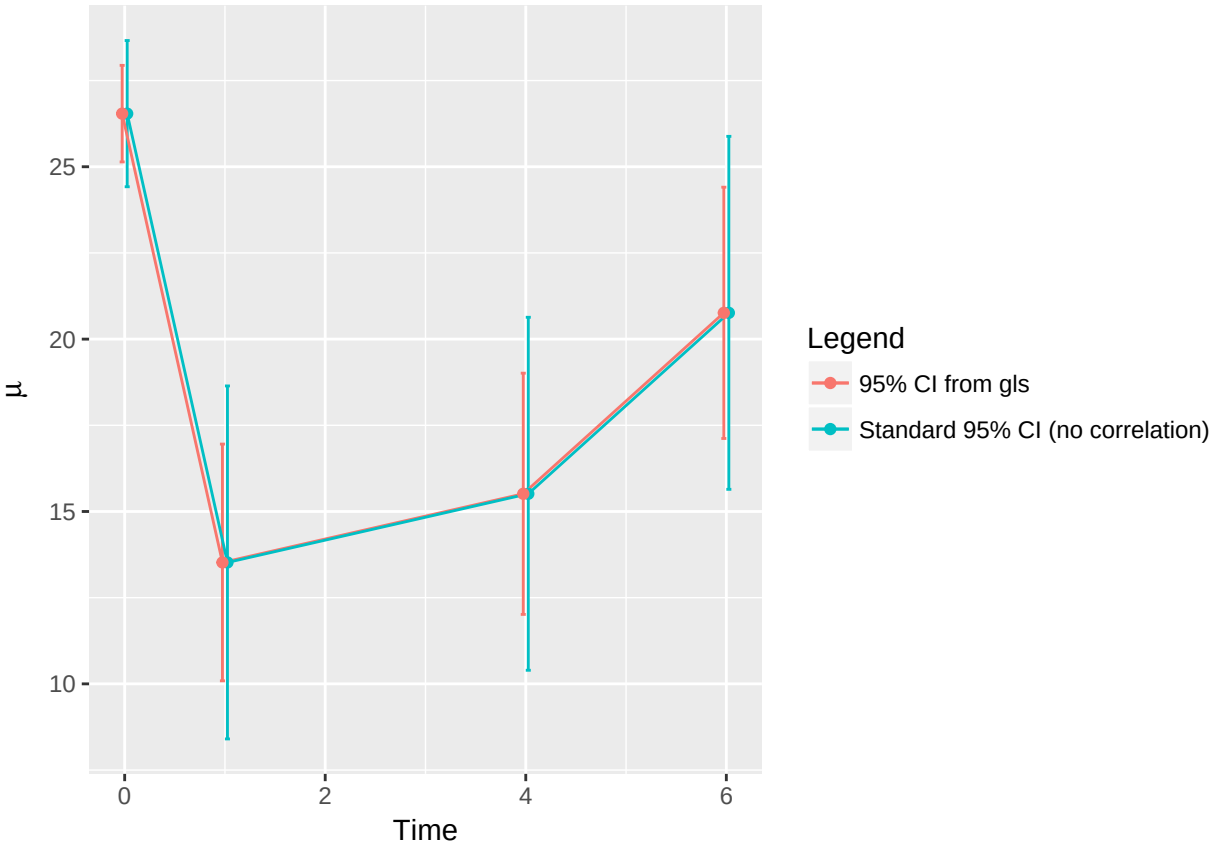
Task 5

In this task we compare confidence intervals for means at different time points in two settings: taking into account correlations and the opposite. In this case these confidence intervals are the same as confidence intervals for coefficients β, because our explanatory variables are just levels of a factor *time.cat*. We expect to get narrower CIs for model which better explains our data, so the model which takes into account the correlation.

```
lmod <- lm(y~factor(time.cat), data = lead.uni)
intervs_lm <- confint(lmod)
intervs <- intervals(lead.cat)$coef
for(i in 2:4){
  intervs[i,] <- intervs[i,] + intervs[1,]
  intervs_lm[i,] <- intervs_lm[i,] + intervs_lm[1,]
}
intervs <- as.data.frame(intervs)

intervs$type <- "95% CI from gls"
colnames(intervs) <- c("lower", "mean", "upper", "type")
intervs_lm <- as.data.frame(intervs_lm)
intervs_lm[,3] <- intervs[,2]
intervs_lm[,4] <- "Standard 95% CI (no correlation)"
intervs_lm <- intervs_lm[,c(1,3,2,4)]
colnames(intervs_lm) <- c("lower", "mean", "upper", "type")
intervs_plot <- rbind(intervs, intervs_lm)
intervs_plot$time <- c(0,1,4,6)
```

```r
pd <- position_dodge(0.1)
ggplot(intervs_plot, aes(x=intervs_plot$time, y=intervs_plot$mean, color = intervs_plot$type)) +
    geom_errorbar(aes(ymin=intervs_plot$lower, ymax=intervs_plot$upper), width=.1, position = pd) +
    geom_line(position = pd) +
    geom_point(position = pd) +
    labs(x="Time", y=TeX('$\\mu$'), colour="Legend")
```



Of course in both settings means are the same. However,as expected the CIs for means from *gls* are narrower because model taking into account correlation better explains the variability in our data.

Task 6

In this task we just have to test a contrast $L = (0, 1, -1, 0)$. This means that $H_0 : L^T\beta = 0$. Again we can use *anova* function which takes the model and contrast as parameters:

```r
anova(lead.cat, L=c(0,1,-1,0))
```

```
## Denom. DF: 196
##  F-test for linear combination(s)
## factor(time.cat)2 factor(time.cat)3
##                 1                -1
##   numDF  F-value p-value
## 1     1 6.111072  0.0143
```

Here we can see that p-value is lower that 0.05. Therefore we reject the null hypothesis at significance level 0.05 and infer that these means are significantly different.