# Complex Data - lab3

*Stanisław Wilczyński*

## Data input and Calculating Means

```r
wtloss <- read.table("../data/weightloss.dat",
header=F)
## Give names to variables
names(wtloss) <- c("id", paste("y", 1:4, sep=""), "program")
## Univariate format
wtloss.uni <- data.frame(id=rep(wtloss$id, each=4),
wgtloss=as.numeric(t(as.matrix(wtloss[,2:5]))),
program=rep(wtloss$program, each=4),
month=seq(0,9,3),
time.cat=rep(1:4))
wtloss.uni$prog.fac <- factor(wtloss.uni$program, labels=c("1:encourage", "2:none"))
attach(wtloss.uni)
wgt.mean <- tapply(wgtloss, list(month, program), mean)
wgt.sd <- tapply(wgtloss, list(month, program), sd)
detach(wtloss.uni)
```

## Plotting Response Profiles

```r
xyplot(wgtloss~month|prog.fac, type='l',groups=id,data=wtloss.uni)
```
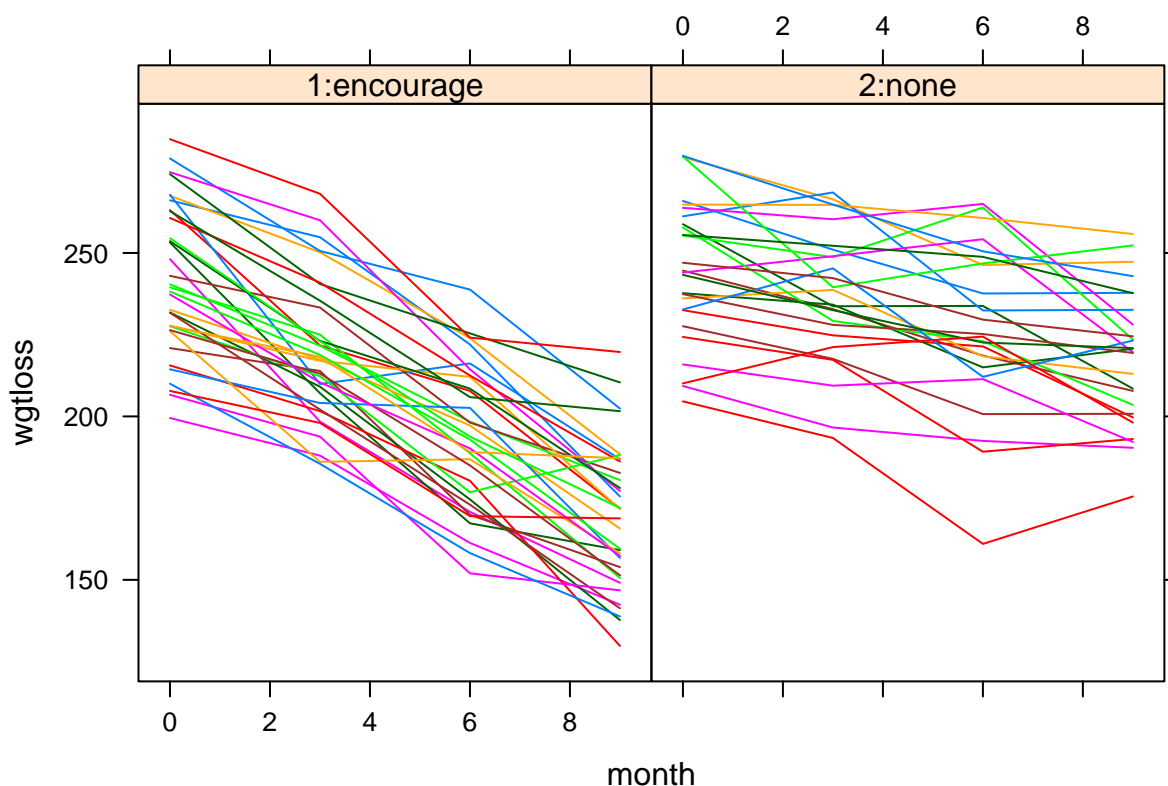
# Choice of covariance structure

We will consider three covariance structures:
1. Unstructured (UN)
2. Compound symmetry (CS)
3. Autoregressive (AR-1)

```
## CATEGORICAL time, UN, REML
wtloss.un.cat <- gls(wgtloss~factor(month)*prog.fac,
correlation=corSymm(form= ~1 | id),
weights=varIdent(form= ~1 | month),
data=wtloss.uni)
#summary(wtloss.un.cat)

## CATEGORICAL time, CS, REML
wtloss.cs.cat <- gls(wgtloss~factor(month)*prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni)
#summary(wtloss.cs.cat)

## CATEGORICAL time, AR1, REML
wtloss.ar1.cat <- gls(wgtloss~factor(month)*prog.fac,
correlation=corAR1(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni)
#summary(wtloss.ar1.cat)
```

## Compound symmetry vs unstructured

$H_0$ : compound symmetry is adequate for the data
$H_1$ : unstructured is required

```
un_vs_cs <- anova(wtloss.un.cat, wtloss.cs.cat)
```

|  | CS vs. UN | |
|---|---|---|
| Structure | -2 REML Log-Likelihood | Number of Cov. Parameters |
| Compound Symmetry | 1905.8232573 | 10 |
| Unstructured | 1900.8766893 | 18 |
| Difference | 4.9465681 | 8 |

LRT yields $G^2 = 4.9465681$ with 8 df (p-value = 0.7632715) so we do not reject the null hypothesis at $\alpha = 0.05$ and conclude that the assumption of compound symmetry covariance structure is adequate for the data.

## Autoregressive vs unstructered

$H_0$ : autoregressive is adequate for the data
$H_1$: unstructured is required

```
un_vs_ar <- anova(wtloss.un.cat, wtloss.ar1.cat)
```

| Structure | -2 REML Log-Likelihood | Number of Cov. Parameters |
|---|---|---|
| Autoregressive | 1931.246837 | 10 |
| Unstructured | 1900.8766893 | 18 |
| Difference | 30.3701477 | 8 |

LRT yields $G^2 = 30.3701477$ with 8 df (p-value $= 1.8178034 \times 10^{-4}$), so we reject the null hypothesis at $\alpha = 0.05$ and conclude that the assumption of autoregressive covariance structure is not adequate when compared to unstructured.

**Compound symmetry vs autoregressive**

```
cs_vs_ar1 <- anova(wtloss.cs.cat, wtloss.ar1.cat)
```

Since CS and AR-1 have the same number of parameters $= 10$, no LRT is necessary. We can directly compare their likelihoods, or -2*log(likelihood):

- $-2 * log(likelihood)$ for CS $= 1905.8232573$

- $-2 * log(likelihood)$ for AR-1 $= 1931.246837$

Since $-2 * log(likelihood)$ for CS is smaller than for AR-1, CS has a higher likelihood and we conclude that CS is an adequate model for the covariance structure when compared to AR-1.

**The most adequate covariance structure is compound symmetry model.**

Are all these tests correct?

No, the last test is theoretically not correct, because for non nested models we shouldn't draw conclusions just basing on comparison of log-likelihoods or performing LRT. We should compare AIC or BIC instead. However, in this case the number of parameters is the same, so comparing likelihoods is equivalent to comparing AICs. Moreover, when comparing many models instead of pairwise tests we can use a better strategy and choose the one with the lowest value of AIC.

| Structure | -2 REML Log-Likelihood | Number of Parameters | AIC |
|---|---|---|---|
| Compound Symmetry | 1905.8232573 | 10 | 1925.8232573 |
| Autoregressive | 1931.246837 | 10 | 1951.246837 |
| Unstructured | 1900.8766893 | 18 | 1936.8766893 |

Thus, we will use a compound symmetry covariance structure for the remainder of the lab.

## Single Degree of Freedom Contrasts

**AUC - test for equality of the area under the curve in two groups.**

```
t <- wtloss.uni$month
L2 <- 0.5*c(t[1] + t[2] - 2*t[4], t[3] - t[1], t[4]-t[2], t[4]-t[3])
coefs <- summary(wtloss.cs.cat)$coefficients
#Estimated mean AUC in encouragement program
AUC_enc <- sum(L2 * wgt.mean[,1])
#Estimated mean AUC in no encouragement program
AUC_none <- sum(L2 * wgt.mean[,2])
```

The obtained values of AUCs are:

Estimated mean AUC in encouragement program is -319.5705882.
Estimated mean AUC in no encouragement program is -117.7038462.

As written in the textbook to test for the equality of AUCs, we employ the contrast $(-L_2, L_2)$.

```
L <- c(-L2, L2)
anova(wtloss.cs.cat, L=L)
```

```
## Denom. DF: 232
##  F-test for linear combination(s)
##                 (Intercept)               factor(month)3
##                         7.5                         -3.0
##               factor(month)6               factor(month)9
##                        -3.0                         -1.5
##               prog.fac2:none factor(month)3:prog.fac2:none
##                        -7.5                          3.0
## factor(month)6:prog.fac2:none factor(month)9:prog.fac2:none
##                         3.0                          1.5
##    numDF  F-value p-value
## 1      1 759.3707  <.0001
```

The p-value is almost 0, so we reject the null hypothesis that the response profile is the same for two treatments.

## Parametric curves

**Quadratic time trend**

```
## QUADRATIC time, CS, ML,
wtloss.cs.quad <- gls(wgtloss~month*prog.fac + I(month^2)*prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni,
method="ML")

## CATEGORICAL time, CS, ML
wtloss.cs.cat.ml <- gls(wgtloss~factor(month)*prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni,
method="ML")
```

$H_0$ : quadratic model
$H_1$ : saturated mode

```
qm_vs_sm <- anova(wtloss.cs.cat.ml, wtloss.cs.quad)
```

Testing the Quadratic trend

| Structure | -2 Log Likelihood | Number of Parameters |
|---|---|---|
| Quadratic model | 1935.7000029 | 8 |
| Saturated model | 1935.3313128 | 10 |
| Difference | 0.3686901 | 2 |

LRT yields $G^2 = 0.3686901$ with 2 df (p-value $= 0.8316488$), so we fail to reject the null hypothesis at $\alpha = 0.05$ and conclude that the model with Month as a quadratic effect seems to fit the data adequately. (Note: $\chi^2_{2,0.95} = 5.99$)

**Linear time trend**

```
wtloss.cs.lin <- gls(wgtloss~month*prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni,
method="ML")
```

$H_0$ : linear model
$H_1$ : quadratic model

```
qm_vs_lm <- anova(wtloss.cs.quad, wtloss.cs.lin)
```

| Linear vs. Quadratic | | |
|---|---|---|
| Structure | -2 Log Likelihood | Number of Parameters |
| Linear model | 1936.295409 | 6 |
| Quadratic model | 1935.7000029 | 8 |
| Difference | 0.5954061 | 2 |

LRT yields $G^2 = 0.5954061$ with 2 df (p-value $= 0.7425218$), so we fail to reject the null hypothesis at $\alpha = 0.05$ and conclude that the model with Month as a linear effect fits the data better than the model with Month as a quadratic effect. (Note: $chi^2_{2,0.95} = 5.99$)

**Testing for intersections in the linear model**

```
## Linear model, NO interacations
wtloss.cs.lin.noint <- gls(wgtloss~month + prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni,
method="ML")
```

$H_0 : \beta_4 = 0$
$H_1 : \beta_4 \neq 0$

```
anova(wtloss.cs.lin)
```

```
## Denom. DF: 236
##               numDF  F-value p-value
## (Intercept)       1 6970.808  <.0001
## month             1 1011.995  <.0001
## prog.fac          1   24.301  <.0001
## month:prog.fac    1  184.368  <.0001
#we employ proper contrast
anova(wtloss.cs.lin, L=c(0,0,0,1))
```

```
## Denom. DF: 236
##  F-test for linear combination(s)
## [1] 1
```

```
##   numDF  F-value p-value
## 1     1 184.3676  <.0001
```

Since the p-value for month*program is almost 0, we reject the null hypothesis and conclude that there is a significant interaction between month and program.

Thus, our final model is the linear model with interaction.

```
final.wtloss.cs.lin <- gls(wgtloss~month*prog.fac,
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni)
summary(final.wtloss.cs.lin)
```

```
## Generalized least squares fit by REML
##   Model: wgtloss ~ month * prog.fac
##   Data: wtloss.uni
##        AIC      BIC    logLik
##   1941.074 1961.857 -964.5368
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | id
##  Parameter estimate(s):
##       Rho
## 0.8083168
##
## Coefficients:
##                       Value Std.Error    t-value p-value
## (Intercept)        242.37676  3.655475   66.30513  0.0000
## month               -8.10431  0.245748 -32.97814  0.0000
## prog.fac2:none       2.99477  5.553065    0.53930  0.5902
## month:prog.fac2:none  5.08329  0.373318   13.61650  0.0000
##
##  Correlation:
##                      (Intr) month  prg.2:
## month                -0.303
## prog.fac2:none       -0.658  0.199
## month:prog.fac2:none  0.199 -0.658 -0.303
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -3.0172539 -0.6942911 -0.0369479  0.7699925  2.2892673
##
## Residual standard error: 21.95552
## Degrees of freedom: 240 total; 236 residual
```

In terms of effectiveness we can conclude that the first program is much more effective than the second one: in the first group the estimated rate of change equals $\hat{\beta}_2 = $ -8.1043137 whereas in the second group estimated rate of change is $\hat{\beta}_2 + \hat{\beta}_4 = $ -3.0210256.

**Interpreting quadratic trends:**

```
wtloss.cs.quad.noint <- gls(wgtloss~month + prog.fac + I(month^2),
correlation=corCompSymm(form= ~1 | id),
weights=varIdent(form= ~1),
data=wtloss.uni)
ncoefs <- wtloss.cs.quad.noint$coefficients
coefs <- c(ncoefs[1], ncoefs[2], ncoefs[4], ncoefs[3])
change_rates <- coefs[2] + 2*coefs[3]*t[1:4]
expected1 <- coefs[1] + coefs[2]*t[1:4] + coefs[3]*t[1:4]*t[1:4]
```

$\hat{\beta}_1 = 232.0118529$ expected mean response at baseline of subjects in program 1.
$\hat{\beta}_4 = 25.8695701$ change in expected response at baseline of subjects in program 2 vs. program 1.
Rate of change in program 1 is $\beta_2 + 2\beta_3 time_{ij}$ .
Plugging in the above estimates,

| | Program 1 | |
|---|---|---|
| time | Rate of Change | Expected Response |
| 0 | -5.4490556 | 232.0118529 |
| 1 | -5.7507222 | 215.2121863 |
| 2 | -6.0523889 | 197.5075196 |
| 3 | -6.3540556 | 178.8978529 |

Thus, the mean response for Program 1 decreases over time. For the second program the effect is exactly the same - in this case for every time point 25.8695701 is added to the expected response, because in our model we don't include any interactions and the group effect is incorporated just by adding a constant.