Complex Data lab 3

Anna Zaleska

```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)
#wgt.mean
#wgt.sd
detach(wtloss.uni)
```

```r
 # Use lattice package
#xyplot(wgtloss~month|prog.fac, type='l',groups=id,data=wtloss.uni)
```

The most general mean models with different covariance classes:

```r
## 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)
```

Unstructured vs compound symmetry

H_0 : compound symmetry is adequate for the data
H_1 : unstructred is required

```r
anova(wtloss.un.cat, wtloss.cs.cat)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## wtloss.un.cat     1 18 1936.877 1998.918 -950.4383
## wtloss.cs.cat     2 10 1925.823 1960.291 -952.9116 1 vs 2 4.946568  0.7633
```

CS vs. UN

Structure	-2 REML Log-Likelihood	Number of Cov. Parameters
Compound Symmetry	1905.8232	10
Unstructured	1900.8766	18
Difference	4.9466	8

LRT yields $G^2 = 4.9466$ with 8 df (p =0.7633), 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.

Unstructured vs autoregressive

H_0 : autoregressive is adequate for the data
H_1 : unstructred is required

```r
anova(wtloss.un.cat, wtloss.ar1.cat)
```

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio
## wtloss.un.cat      1 18 1936.877 1998.918 -950.4383
## wtloss.ar1.cat     2 10 1951.247 1985.714 -965.6234 1 vs 2 30.37015
##                p-value
## wtloss.un.cat
## wtloss.ar1.cat   2e-04
```

AR-1 vs. UN

Structure	-2 REML Log-Likelihood	Number of Cov. Parameters
Autoregressive	1931.2468	10
Unstructured	1900.8766	18
Difference	30.3702	8

LRT yields $G^2 = 30.3702$ with 8 df (p =0.0002), so we reject the null hypothesis at $\alpha = 0.05$ and conclude that the assumption of autoregressive covariance structure is inappropriate, worse? when compared to unstructured

Autoregressive vs compound symmetry

```r
anova(wtloss.cs.cat, wtloss.ar1.cat)
```

```
##                Model df      AIC      BIC    logLik
## wtloss.cs.cat      1 10 1925.823 1960.291 -952.9116
## wtloss.ar1.cat     2 10 1951.247 1985.714 -965.6234
```

```r
#same number of parameters so we compare -2*log(likelihood)
```

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 = 1905.8232

- $-2 * log(likelihood)$ for AR-1 = 1931.2468

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?
anova(wtloss.un.cat, wtloss.ar1.cat)
anova(wtloss.un.cat, wtloss.cs.cat)
anova(wtloss.ar1.cat, wtloss.cs.cat)

Structure	-2 REML Log-Likelihood	Number of Cov. Parameters	AIC
Compoud Symmetry	1905.8232	10	1925.823
Autoregressive	1931.2468	10	1951.247
Unstructured	1900.8766	18	1936.877

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

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

Parametric curves

Quadratic time trend

```
## QUADRATIC time, CS, REML,
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")
#summary(wtloss.cs.quad)

## 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")
#summary(wtloss.cs.cat.ml)
```

H_0 : quadratic model
H_1 : saturated mode

```
anova(wtloss.cs.cat.ml, wtloss.cs.quad)
```

```
##                    Model df      AIC      BIC    logLik   Test   L.Ratio
## wtloss.cs.cat.ml       1 10 1955.331 1990.138 -967.6657
## wtloss.cs.quad         2  8 1951.700 1979.545 -967.8500 1 vs 2 0.3686901
##                    p-value
## wtloss.cs.cat.ml
## wtloss.cs.quad      0.8316
```

Testing the Quadratic trend

Structure	-2 Log Likelihood	Number of Cov. Parameters
Quadratic model	1935.7	8
Saturated model	1935.3314	10
Difference	0.3686	2

LRT yields G2 =0.369 (p = 0.8316), 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")
#summary(wtloss.cs.lin)
```

H_0 : linear model
H_1 : quadratic model

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

```
##                  Model df      AIC      BIC    logLik   Test   L.Ratio
```

```
## wtloss.cs.lin      1  6 1948.295 1969.179 -968.1477
## wtloss.cs.quad     2  8 1951.700 1979.545 -967.8500 1 vs 2 0.5954061
##                p-value
## wtloss.cs.lin
## wtloss.cs.quad  0.7425
```

Linear vs. Quadratic

Structure	-2 Log Likelihood	Number of Cov. Parameters
Linear model	1936.2954	6
Quadratic model	1935.7	8
Difference	0.5954	2

LRT yields G2 =0.596 (p = 0.7425), 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

```
## 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")
#summary(wtloss.cs.lin.noint)
```

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

```
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
```

Since the p-value for month*program is < 0.0001, we reject the null hypothesis and conclude that there is an interaction between month and program.

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

What can you conclude about the two weight programs in terms of their effectiveness?

Interpreting quadratic trends:

$\hat{\beta}_1$ =?? expected mean response at baseline of subjects in program 1.
$\hat{\beta}_1$=?? 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		
1		
2		
3		

Thus, the mean response for Program 1 ?? (increases/decreases) over time. What about Program 2?