GLM Case Study: TLC Trial

Treatment of Lead-Exposed Children (TLC) Trial

Exposure to lead, often due to deteriorating lead-based paint in older homes, can damage cognitive function, especially in children. The CDC has decided that children with blood lead level over 10 $\mu g/dL$ are at risk.

Chelating agents can be used to treat lead poisoning, which were usually introduced by injection and required hospitalization. A new agent, succimer, can be given orally. In 1990, the Treatment of Lead-Exposed Children (TLC) Trial Group conducted a placebo-controlled, randomized trial of succimer in children with blood lead levels of 20-44 $\mu g/dL$. The children in the study were aged 12-33 months at enrollment. They received up to three 26-day courses of succimer or placebo and were followed for 3 years.

The data set we will look at were a random sample of 100 children, with blood levels measured at baseline, week 1, 4 and 6.

Question of Interest: whether succimer reduces blood lead levels over time relative to placebo.
## Data

Table 1: Blood lead levels (µg/dL) at baseline, week 1, 4 and 6 for 10 children in the TLC trial

<table>
<thead>
<tr>
<th>ID</th>
<th>Group</th>
<th>Baseline</th>
<th>Week 1</th>
<th>Week 4</th>
<th>Week 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P</td>
<td>30.8</td>
<td>26.9</td>
<td>25.8</td>
<td>23.8</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>26.5</td>
<td>14.8</td>
<td>19.5</td>
<td>21.0</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>25.8</td>
<td>23.0</td>
<td>19.1</td>
<td>23.2</td>
</tr>
<tr>
<td>4</td>
<td>P</td>
<td>24.7</td>
<td>24.5</td>
<td>22.0</td>
<td>22.5</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>20.4</td>
<td>2.8</td>
<td>3.2</td>
<td>9.4</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>20.4</td>
<td>5.4</td>
<td>4.5</td>
<td>11.9</td>
</tr>
<tr>
<td>7</td>
<td>P</td>
<td>28.6</td>
<td>20.8</td>
<td>19.2</td>
<td>18.4</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>33.7</td>
<td>31.6</td>
<td>28.5</td>
<td>25.1</td>
</tr>
<tr>
<td>9</td>
<td>P</td>
<td>19.7</td>
<td>14.9</td>
<td>15.3</td>
<td>14.7</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>31.1</td>
<td>31.2</td>
<td>29.2</td>
<td>30.1</td>
</tr>
</tbody>
</table>
Summary Statistics

Read in the Data and Compute Some Summary Statistics

```r
> tlc <- read.table("data/tlc.txt",
+    col.names = c("ID", "Group", "week.0",
+        "week.1", "week.4", "week.6"))
> tlc[1:4,]
   ID Group week.0 week.1 week.4 week.6
1  1  P    30.8   26.9  25.8  23.8
2  2  A    26.5   14.8  19.5  21.0
3  3  A    25.8   23.0  19.1  23.2
4  4  P    24.7   24.5  22.0  22.5
>  
> do.call("rbind", tapply(tlc$week.0, tlc$Group, summary))

   Min. 1st Qu. Median  Mean 3rd Qu. Max.
A 19.7  22.13   26.20  26.54  29.55  41.1
P 19.7  21.88   25.25  26.27  29.73  38.1
>
> by(tlc[,-(1:2)], tlc$Group,
+   function (x) cbind(mean = mean(x), sd = sd(x))

tlc$Group: A

    mean    sd
week.0 26.540 5.020936
week.1 13.522 7.672487
week.4 15.514 7.852207
week.6 20.762 9.246332

------------------------------------------------------------

tlc$Group: P

    mean    sd
week.0 26.272 5.024107
week.1 24.660 5.461180
week.4 24.070 5.753127
week.6 23.646 5.639808
```
Explore the Data

First we need convert it to long format:

```r
> tlcL <- reshape (tlc, direction = "long", idvar = "ID",
+ varying = 3:6)
> names (tlcL)[3:4] <- c("Week", "Lead")
> tlcL[95:105,]

<table>
<thead>
<tr>
<th>ID</th>
<th>Group</th>
<th>Week</th>
<th>Lead</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.0</td>
<td>A</td>
<td>0</td>
<td>31.2</td>
</tr>
<tr>
<td>96.0</td>
<td>A</td>
<td>0</td>
<td>31.4</td>
</tr>
<tr>
<td>97.0</td>
<td>A</td>
<td>0</td>
<td>41.1</td>
</tr>
<tr>
<td>98.0</td>
<td>A</td>
<td>0</td>
<td>29.4</td>
</tr>
<tr>
<td>99.0</td>
<td>A</td>
<td>0</td>
<td>21.9</td>
</tr>
<tr>
<td>100.0</td>
<td>A</td>
<td>0</td>
<td>20.7</td>
</tr>
<tr>
<td>1.1</td>
<td>P</td>
<td>1</td>
<td>26.9</td>
</tr>
<tr>
<td>2.1</td>
<td>A</td>
<td>1</td>
<td>14.8</td>
</tr>
<tr>
<td>3.1</td>
<td>A</td>
<td>1</td>
<td>23.0</td>
</tr>
<tr>
<td>4.1</td>
<td>P</td>
<td>1</td>
<td>24.5</td>
</tr>
<tr>
<td>5.1</td>
<td>A</td>
<td>1</td>
<td>2.8</td>
</tr>
</tbody>
</table>
```

Scatterplot, by treatment group, with LOESS smoothing curve.

```r
library (lattice)
xyplot (Lead ~ Week | Group, data = tlcL,
  groups = tlcL$ID, type = "l",
  panel = function (x, y, subscripts, groups, ...) {
    panel.superpose (x, y,
      panel.groups = "panel.xyplot",
      subscripts,
      groups, col = "gray40", ...)
    panel.loess (x, y, col = "red", lwd = 2, ...)
  })
```
Figure 1: Plot of blood lead levels, by treatment group.

Notes

- Complete and balanced data.
- Interested in marginal inference: i.e., compare the mean profiles of the two groups over time.
- Randomized trial.
- The mean profile does not appear to be linear, especially for the treatment group.
Correlation Structure

```r
panel.hist <- function (x, ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5))
    h <- hist(x, plot = FALSE, probability = TRUE)
    breaks <- h$breaks
    nB <- length(breaks)
    y <- h$counts
    y <- y / max(y)
    rect(breaks[-nB], 0, breaks[-1], y,
         col = "cyan", ...)
    xd <- density(x)
    xd$y <- xd$y / max(xd$y)
    lines(xd, col = "brown", lwd = 1.5)
}
panel.cor <- function(x, y, digits = 2, prefix = ",", cex.cor) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y, use = "pairwise.complete.obs"))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if (missing(cex.cor))
        cex <- 0.8 / strwidth(txt)
    text(0.5, 0.5, txt, cex = cex * r)
}
pairs(tlc[,3:6], diag.panel = panel.hist,
      upper.panel = panel.cor,
      lower.panel = panel.smooth)
```
Figure 2: Pairwise scatter-plot of blood lead levels at baseline, week 1, 4 and 6 for children in TLC trial.
Figure 3: Pairwise scatter-plot of blood lead levels at baseline, week 1, 4 and 6 for children in TLC trial, by treatment group.
Objectives of Analysis

The null hypothesis of no treatment effect can be expressed in different ways:

• $H_0 : \mu_j(A) = \mu_j(P)$ for all $j = 1, 2, 3, 4$.
  - Time is treated as a factor.
  - This null can be expressed in terms of both the regression coefficients for the treatment and time $\times$ treatment interactions.

• $H_0 : \mu_j(A) - \mu_1(A) = \mu_j(P) - \mu_1(P)$ for all $j = 2, 3, 4$.
  - Emphasis on the treatment effect on the changes, i.e., time $\times$ treatment interaction.
  - Less restrictive, allows the baseline lead levels to differ between groups.

• Model the response profile via a parametric (or non-parametric) model, i.e., a linear or quadratic model, and test the time $\times$ treatment interaction effect.
  - Linear model is not appropriate.
Simple Linear Model

> temp <- lm (Lead ~ factor (Week) * Group, data = tlcL)
> summary (temp)

Call:
lm(formula = Lead ~ factor(Week) * Group, data = tlcL)

Residuals:
            Min       1Q   Median       3Q      Max
-16.662 -4.621  -0.993  3.672  43.138

Coefficients:  
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                     26.540    0.937  28.324  < 2e-16 ***
factor(Week)1                   -13.018    1.325  -9.824  < 2e-16 ***
factor(Week)4                   -11.026    1.325  -8.321  1.47e-15 ***
factor(Week)6                   -5.778    1.325  -4.360  1.66e-05 ***
GroupP                           -0.268    1.325   -0.202    0.8398
factor(Week)1:GroupP            11.406    1.874   6.086  2.75e-09 ***
factor(Week)4:GroupP            8.824    1.874   4.709  3.47e-06 ***
factor(Week)6:GroupP            3.152    1.874   1.682    0.0934 .

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.626 on 392 degrees of freedom
Multiple R-Squared: 0.3284, Adjusted R-squared: 0.3164
F-statistic: 27.38 on 7 and 392 DF,  p-value: < 2.2e-16

> anova (temp)

Analysis of Variance Table

Response: Lead  
                         Df   Sum Sq  Mean Sq   F value     Pr(>F)
factor(Week)            3  3272.8  1090.90   24.850  9.701e-15 ***
Group                   1  3110.9  3110.90   70.862  7.281e-16 ***
factor(Week):Group      3  2030.4   676.80   15.417  1.685e-09 ***
Residuals              392 17208.8   43.90


Model Diagnosis

> par (mfrow = c (2, 2))
> plot (temp)

Figure 4: Simple Linear Model
GEE

In R, GEE (for linear model, it just means robust variance estimation) is implemented by libraries `gee` and a newer `geepack` (the function name is `geese`).

```r
> library(gee)
> tlcL <- tlcL[order(tlcL$Group, tlcL$ID, tlcL$Week),]

Note that it is necessary to sort the data by ID first.
By default, `gee` uses “working independence” correlation matrix.

```r
> temp <- gee(Lead ~ factor(Week) * Group, id = ID, data = tlcL)
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gge S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link: Identity
Variance to Mean Relation: Gaussian
Correlation Structure: Independent

Call:
gee(formula = Lead ~ factor(Week) * Group, id = ID, data = tlcL)

Summary of Residuals:
Min 1Q Median 3Q Max
-16.6620 -4.6205 -0.9930 3.6725 43.1380

Coefficients:

 extinct Naive S.E. Naive z Robust S.E. Robust z
(Intercept) 26.540 0.937 28.324 0.703 37.756
factor(Week)1 -13.018 1.325 -9.824 1.021 -12.755
factor(Week)4 -11.026 1.325 -8.321 1.053 -10.469
factor(Week)6 -5.778 1.325 -4.360 1.126 -5.130
GroupP -0.268 1.325 -0.202 0.994 -0.270
factor(Week)1:GroupP 11.406 1.874 6.086 1.109 10.288
```
factor(Week)4:GroupP  8.824  1.874  4.709  1.141  7.734
factor(Week)6:GroupP  3.152  1.874  1.682  1.244  2.534

Estimated Scale Parameter:  43.9
Number of Iterations:  1

Working Correlation
[1,]   1  0  0  0
[2,]   0  1  0  0
[3,]   0  0  1  0
[4,]   0  0  0  1

• The “naive” SEs are based on the specified correlation matrix (what we called “model-based” SEs). Note that here they are the same as in the simple linear model.

• The coefficients are the same as in OLS.

• The robust estimates of SE are smaller (more efficient).

• There appears to be an outlier but we will ignore it.

• Since GEE is not based on likelihood, we can’t use likelihood ratio or score tests. We can use Wald test to test the null hypothesis of no Week:Group interaction effect but some programming seems necessary.

• `temp$robust.variance` gives the full covariance matrix for $\beta$. 

Exchangeable correlation

```r
> temp <- gee(Lead ~ factor(Week) * Group, id = ID,
+            corstr = "exchangeable", data = tlcL)
[1] "Beginning Ggee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
> summary(temp)
```

```
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
ggee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link: Identity
Variance to Mean Relation: Gaussian
Correlation Structure: Exchangeable

Call:
gee(formula = Lead ~ factor(Week) * Group, id = ID, data = tlcL,
    corstr = "exchangeable")

Summary of Residuals:
  Min   1Q Median   3Q  Max
-16.662 -4.621  -0.993  3.673  43.138

Coefficients:

(Intercept) 26.540 0.937 28.324 0.703 37.756
factor(Week)1 -13.018 0.847 -15.369 1.021 -12.755
factor(Week)4 -11.026 0.847 -13.017 1.053 -10.469
factor(Week)6 -5.778 0.847 -6.821 1.126 -5.130
GroupP -0.268 1.325 -0.202 0.994 -0.270
factor(Week)1:GroupP 11.406 1.198 9.522 1.109 10.288
factor(Week)4:GroupP 8.824 1.198 7.366 1.141 7.734
factor(Week)6:GroupP 3.152 1.198 2.631 1.244 2.534

Estimated Scale Parameter: 43.9
Number of Iterations: 1

Working Correlation

[1,] 1.000 0.591 0.591 0.591
[2,] 0.591 1.000 0.591 0.591
[3,] 0.591 0.591 1.000 0.591
[4,] 0.591 0.591 0.591 1.000
```
Unstructured correlation

> temp <- gee (Lead ~ factor (Week) * Group, id = ID, 
+ corstr = "unstructured", data = tlcL)
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
> summary (temp)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gem S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link: Identity
Variance to Mean Relation: Gaussian
Correlation Structure: Unstructured

Call:
gee(formula = Lead ~ factor(Week) * Group, id = ID, data = tlcL, 
corstr = "unstructured")

Summary of Residuals:
    Min  1Q Median  3Q Max
-16.662 -4.620 -0.993  3.672 43.138

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Naive S.E.</th>
<th>Naive z</th>
<th>Robust S.E.</th>
<th>Robust z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>26.540</td>
<td>0.937</td>
<td>28.324</td>
<td>0.703</td>
</tr>
<tr>
<td>factor(Week)1</td>
<td>-13.018</td>
<td>0.996</td>
<td>-13.072</td>
<td>1.021</td>
</tr>
<tr>
<td>factor(Week)4</td>
<td>-11.026</td>
<td>0.984</td>
<td>-11.207</td>
<td>1.053</td>
</tr>
<tr>
<td>factor(Week)6</td>
<td>-5.778</td>
<td>0.932</td>
<td>-6.202</td>
<td>1.126</td>
</tr>
<tr>
<td>GroupP</td>
<td>-0.268</td>
<td>1.325</td>
<td>-0.202</td>
<td>0.994</td>
</tr>
<tr>
<td>factor(Week)1:GroupP</td>
<td>11.406</td>
<td>1.408</td>
<td>8.099</td>
<td>1.109</td>
</tr>
<tr>
<td>factor(Week)4:GroupP</td>
<td>8.824</td>
<td>1.391</td>
<td>6.342</td>
<td>1.141</td>
</tr>
<tr>
<td>factor(Week)6:GroupP</td>
<td>3.152</td>
<td>1.318</td>
<td>2.392</td>
<td>1.244</td>
</tr>
</tbody>
</table>

Estimated Scale Parameter:  43.9
Number of Iterations: 1

Working Correlation

```
[1,]  1.000  0.435  0.449  0.506
[2,]  0.435  1.000  0.809  0.676
[3,]  0.449  0.809  1.000  0.698
[4,]  0.506  0.676  0.698  1.000
```

- The robust standard error estimates are same for different correlation models.
Generalized Least Squares

- R library *nlme* provides a function *gls* that does generalized least squares estimation.

- The difference with *gee* is that it does not compute sandwich standard error estimates.

```r
> temp <- gls (Lead ~ factor (Week) * Group,
+     data = tlcL, method = "ML",
+     correlation = corCompSymm (form = ~ 1 | ID))
```

Generalized least squares fit by maximum likelihood
Model: Lead ~ factor(Week) * Group
Data: tlcL
```

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>26.54</td>
<td>0.937</td>
<td>28.32</td>
<td>0.0000</td>
</tr>
<tr>
<td>factor(Week)1</td>
<td>-13.02</td>
<td>0.843</td>
<td>-15.45</td>
<td>0.0000</td>
</tr>
<tr>
<td>factor(Week)4</td>
<td>-11.03</td>
<td>0.843</td>
<td>-13.08</td>
<td>0.0000</td>
</tr>
<tr>
<td>factor(Week)6</td>
<td>-5.78</td>
<td>0.843</td>
<td>-6.86</td>
<td>0.0000</td>
</tr>
<tr>
<td>GroupP</td>
<td>-0.27</td>
<td>1.325</td>
<td>-0.20</td>
<td>0.8399</td>
</tr>
<tr>
<td>factor(Week)1:GroupP</td>
<td>11.41</td>
<td>1.192</td>
<td>9.57</td>
<td>0.0000</td>
</tr>
<tr>
<td>factor(Week)4:GroupP</td>
<td>8.82</td>
<td>1.192</td>
<td>7.40</td>
<td>0.0000</td>
</tr>
<tr>
<td>factor(Week)6:GroupP</td>
<td>3.15</td>
<td>1.192</td>
<td>2.64</td>
<td>0.0085</td>
</tr>
</tbody>
</table>

Correlation:

<table>
<thead>
<tr>
<th></th>
<th>(Intr)</th>
<th>fc(W)1</th>
<th>fc(W)4</th>
<th>fc(W)6</th>
<th>GroupP</th>
<th>f(W)1: f(W)4:</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(Week)1</td>
<td>-0.450</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(Week)4</td>
<td>-0.450</td>
<td>0.500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(Week)6</td>
<td>-0.450</td>
<td>0.500</td>
<td>0.500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GroupP</td>
<td>-0.707</td>
<td>0.318</td>
<td>0.318</td>
<td>0.318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>factor(Week)1:GroupP</td>
<td>0.318</td>
<td>-0.707</td>
<td>-0.354</td>
<td>-0.354</td>
<td>-0.450</td>
<td></td>
</tr>
<tr>
<td>factor(Week)4:GroupP</td>
<td>0.318</td>
<td>-0.354</td>
<td>-0.707</td>
<td>-0.354</td>
<td>-0.450</td>
<td>0.500</td>
</tr>
<tr>
<td>factor(Week)6:GroupP</td>
<td>0.318</td>
<td>-0.354</td>
<td>-0.354</td>
<td>-0.707</td>
<td>-0.450</td>
<td>0.500 0.500</td>
</tr>
</tbody>
</table>

Standardized residuals:
```
<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.540</td>
<td>-0.704</td>
<td>-0.151</td>
<td>0.560</td>
<td>6.575</td>
</tr>
</tbody>
</table>
```

Residual standard error: 6.56
Degrees of freedom: 400 total; 392 residual
• By default, REML is used. We requested maximum likelihood by specifying the `method` argument. In this case, there is very little difference.

Generalized least squares fit by REML

Model: Lead ~ factor(Week) * Group

Data: tlcL

AIC  BIC  logLik
2481 2520  -1230

Correlation Structure: Compound symmetry

Formula: ~1 | ID

Parameter estimate(s):

Rho
0.596

... 

Standardized residuals:

Min  Q1  Med  Q3  Max
-2.514 -0.697 -0.150 0.554 6.510

Residual standard error: 6.63

Degrees of freedom: 400 total; 392 residual

• Since REML is “conditional” on the fixed effects, when comparing models with different fixed effects (regression coefficients), maximum likelihood should be used.

• `gls` does anova (F-test).

```r
> anova (temp)
Denom. DF: 392

                      numDF F-value  p-value
(Intercept)           1   1532 <.0001
factor(Week)          3     60 <.0001
Group                 1     25 <.0001
factor(Week):Group    3     37 <.0001
```

> intervals (temp)
Approximate 95% confidence intervals

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>24.697</td>
<td>26.540</td>
<td>28.38</td>
</tr>
<tr>
<td>factor(Week)1</td>
<td>-14.675</td>
<td>-13.018</td>
<td>-11.36</td>
</tr>
<tr>
<td>factor(Week)4</td>
<td>-12.683</td>
<td>-11.026</td>
<td>-9.37</td>
</tr>
<tr>
<td>factor(Week)6</td>
<td>-7.435</td>
<td>-5.778</td>
<td>-4.12</td>
</tr>
<tr>
<td>GroupP</td>
<td>-2.874</td>
<td>-0.268</td>
<td>2.34</td>
</tr>
<tr>
<td>factor(Week)1:GroupP</td>
<td>9.063</td>
<td>11.406</td>
<td>13.75</td>
</tr>
<tr>
<td>factor(Week)4:GroupP</td>
<td>6.481</td>
<td>8.824</td>
<td>11.17</td>
</tr>
<tr>
<td>factor(Week)6:GroupP</td>
<td>0.809</td>
<td>3.152</td>
<td>5.50</td>
</tr>
</tbody>
</table>

Correlation structure:

<table>
<thead>
<tr>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rho</td>
<td>0.5</td>
<td>0.596</td>
</tr>
</tbody>
</table>

Residual standard error:

<table>
<thead>
<tr>
<th>lower</th>
<th>est.</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.94</td>
<td>6.56</td>
<td>7.25</td>
</tr>
</tbody>
</table>
Bootstrap standard error estimates

The `glsD` function in library Design is an enhanced version of gls that can estimate standard error via bootstrap.

Note: since the data is “cluster”, bootstrap is done at the cluster level.

```R
> library (Design)
Loading required package: Hmisc
Hmisc library by Frank E Harrell Jr

Type library(help='Hmisc'), ?Overview, or ?Hmisc.Overview')
to see overall documentation.

> temp <- glsD (Lead ~ factor (Week) * Group,
+       data = tlcL,
+       correlation = corCompSymm (form = ~ 1 | ID),
+       \textcolor{red}{B = 1000})
> temp
Generalized least squares fit by REML
  Model: Lead ~ factor(Week) * Group
  Data: tlcL
  Log-restricted-likelihood: -1230.311

Using bootstrap variance estimates

|                | Coef  | S.E   | Z    | Pr(>|Z|)   |
|----------------|-------|-------|------|------------|
| Intercept      | 26.540| 0.6946| 38.21| < 2.2e-16  |
| Week=1         | -13.018| 0.9982| -13.04| < 2.2e-16  |
| Week=4         | -11.026| 1.0338| -10.67| < 2.2e-16  |
| Week=6         | -5.778 | 1.1726| -4.93 | 8.331e-07  |
| Group=P        | -0.268 | 0.9341| -0.29 | 0.77418    |
| Week=1 * Group=P| 11.406| 1.0925| 10.44 | < 2.2e-16  |
| Week=4 * Group=P| 8.824 | 1.1221| 7.86  | 3.719e-15  |
| Week=6 * Group=P| 3.152 | 1.2788| 2.46  | 0.01371    |

Correlation Structure: Compound symmetry
  Formula: ~1 | ID
  Parameter estimate(s):
    Rho
    0.5954417
  Degrees of freedom: 400 total; 392 residual
Residual standard error: 6.625722
Clusters: 100
Bootstrap repetitions: 1000
Bootstraps were all balanced with respect to clusters
Ratio of Original Variances to Bootstrap Variances

<table>
<thead>
<tr>
<th></th>
<th>Week=1</th>
<th>Week=4</th>
<th>Week=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.82</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>Group=P</td>
<td>2.01</td>
<td>1.19</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Bootstrap Nonparametric 0.95 Confidence Limits for Correlation Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.450</td>
<td>0.718</td>
</tr>
</tbody>
</table>

> anova (temp)

Wald Statistics Response: Lead

<table>
<thead>
<tr>
<th>Factor</th>
<th>Chi-Square</th>
<th>d.f.</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week (Factor+Higher Order Factors)</td>
<td>203.53</td>
<td>6</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>111.87</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Group (Factor+Higher Order Factors)</td>
<td>115.70</td>
<td>4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>111.87</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Week * Group (Factor+Higher Order Factors)</td>
<td>111.87</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>206.31</td>
<td>7</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Estimating the contrasts:

> tlcL$wc <- factor (tlcL$Week)
> tempB <- glsD (Lead ~ wc * Group,
+     data = tlcL,
+     correlation = corCompSymm (form = ~ 1 | ID))
>
> wcl <- levels (tlcL$wc)
> contrast (tempB,
+     list (Group = "A", wc = wcl),
+     list (Group = "P", wc = wcl))

| wc | Contrast S.E. | Lower   | Upper   | Z      | Pr(>|z|)  |
|----|---------------|---------|---------|--------|----------|
| 0  | 0.268         | 1.325144| -2.329235| 2.865235| 0.20     | 0.8397   |
| 1  | -11.138       | 1.325144| -13.735235| -8.5407645| -8.41   | 0.0000   |
| 4  | -8.556        | 1.325144| -11.153235| -5.9587645| -6.46   | 0.0000   |
| 6  | -2.884        | 1.325144| -5.481235| -0.2867645| -2.18   | 0.0295   |
Estimating the mean responses:

```r
> newdata <- data.frame (expand.grid (wcl, c("A", "P")))
> names (newdata) <- c("wc", "Group")
> cbind (newdata, predict (tempB, newdata = newdata,
+     conf.int = 0.95))

<table>
<thead>
<tr>
<th>wc</th>
<th>Group</th>
<th>linear.predictors</th>
<th>se.fit</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td>26.540</td>
<td>0.9370</td>
<td>24.7034</td>
<td>28.3765</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>13.522</td>
<td>0.9370</td>
<td>11.6854</td>
<td>15.3585</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>15.514</td>
<td>0.9370</td>
<td>13.6774</td>
<td>17.3505</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>20.762</td>
<td>0.9370</td>
<td>18.9254</td>
<td>22.5985</td>
</tr>
<tr>
<td>0</td>
<td>P</td>
<td>26.272</td>
<td>0.9370</td>
<td>24.4354</td>
<td>28.1085</td>
</tr>
<tr>
<td>1</td>
<td>P</td>
<td>24.660</td>
<td>0.9370</td>
<td>22.8234</td>
<td>26.4965</td>
</tr>
<tr>
<td>4</td>
<td>P</td>
<td>24.070</td>
<td>0.9370</td>
<td>22.2334</td>
<td>25.9065</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>23.646</td>
<td>0.9370</td>
<td>21.8094</td>
<td>25.4825</td>
</tr>
</tbody>
</table>

tlc.means <- data.frame (newdata,
                         predict (tempB, newdata = newdata,
                         conf.int = 0.95))

names (tlc.means)[3] <- "Lead"

xYplot (Cbind (Lead, lower, upper) ~ as.numeric (as.character (wc)),
       group = Group,
       ylim = c(10, 30), xlab = "Weeks",
       ylab = "Mean Blood Lead Level",
       data = tlc.means)
```
Figure 5: Mean Blood Lead Levels with 95% CI.
More about Weighted Least Square/GEE

- For a fixed $W$, under the only assumption $E(Y) = X\beta$, the WLS estimator $\hat{\beta}(W)$ is
  1. Unbiased.
  2. Consistent and asymptotically normal.
  3. Efficient if a consistent estimator of $\text{Var}(Y_i)$ is available.

- If the $W(\alpha)$ depends on the data, i.e., $\alpha$ has to be estimated from the data. The WLE estimator $\hat{\beta}(\hat{W})$ solves:

$$X^T\hat{W}(Y - X\beta) = 0,$$

and is not necessarily unbiased.

- $\alpha$ can be estimated using simple methods of moments or ML or REML (even without normality assumption).

- However, under mild assumptions, $\hat{\beta}(\hat{W})$ is still consistent and asymptotically normal and has the same asymptotic variance as $\hat{\beta}(W)$ if $\hat{W}$ converges to $W$. So again, using a consistent estimator of $\text{Var}(Y_i)$ ensures asymptotic efficiency.

- $\alpha$ is not parameter of interest and nothing is said about them. There are extensions of GEE that models variance parameters by specifying higher moments. (Leave to later).

- GEE method trades off some efficiency with consistency, depending upon whether the correlation structure is correctly specified.

- Using a reasonable working correlation matrix can improve efficiency.

- When there is missing data, or highly unbalanced design, the robust approach becomes problematic.

- The GEE method is most appropriate when the number of subjects is much larger than the number of observations per subject, and complete balanced design.
Dealing with Baseline Outcome

An example of pre-post data

When only two measurements are taken for each subject, say pre- and post-treatments \((Y_i^0\) and \(Y_i^1\)) (i.e. \(n = 2\)). Let \(X\) be treatment indicator. Consider the three possible models:

\[
Y_i^1 = \mu + \beta_1 X_i + \epsilon_i \quad (1)
\]
\[
(Y_i^1 - Y_i^0) = \mu^* + \beta_1^* X_i + \epsilon_i \quad (2)
\]
\[
Y_i^1 = \mu^{**} + \beta_1^{**} X_i + \beta_2 Y_i^0 + \epsilon_i \quad (3)
\]

- For randomized trials, it can be shown that \(\beta_1 = \beta_1^* = \beta_1^{**}\). The last two models may be more precise.

- For observational studies, the “post-only” model (1) is generally not satisfactory. The “change” model (2) and the “adjust” model (3) have different interpretations and often quite different values for \(\beta_1\).

\[
> \text{summary (lm (week.1 ~ Group, data = tlc))}
\]

Call: lm(formula = week.1 ~ Group, data = tlc)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 13.5220 | 0.9418 | 14.358 | < 2e-16 *** |
| GroupP | 11.1380 | 1.3319 | 8.363 | 4.24e-13 *** |

\[
> \text{summary (lm (I(week.1 - week.0) ~ Group, data = tlc))}
\]

Call: lm(formula = I(week.1 - week.0) ~ Group, data = tlc)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -13.0180 | 0.7919 | -16.44 | <2e-16 *** |
| GroupP | 11.4060 | 1.1199 | 10.18 | <2e-16 *** |

\[
> \text{summary (lm (week.1 ~ Group + week.0, data = tlc))}
\]

Call: lm(formula = week.1 ~ Group + week.0, data = tlc)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -6.5810 | 3.0336 | -2.169 | 0.0325 * |
| GroupP | 11.3410 | 1.0991 | 10.318 | < 2e-16 *** |
| week.0 | 0.7575 | 0.1105 | 6.855 | 6.61e-10 *** |
Baseline Response for Longitudinal Data

In the case where more than two observations ("waves") are taken, consider the four ways of handling the baseline value:

1. Retain it as part of the outcome vector and make no assumptions about group differences in the mean response at baseline.

2. Retain it as part of the outcome and assume the group means are equal at baseline, such as in a randomized trial.

3. Subtract the baseline response from all remaining responses.

4. Use the baseline value as a covariate in the analysis.

- Method 1 vs. method 2
  - In randomized trial, both methods 1 and 2 yield valid estimates of group difference, but method 2 is in general more powerful.
  - In observational studies, method 2 is not appropriate generally and only method 1 should be used.
  - In methods 1 and 2, the null hypothesis is that the Group by Week interaction effects are zero.
  - There is no Group main effect in method 2.

- Method 3 vs. method 4
  - The interpretation of the regression coefficients is different, for all three factors in the model!
  - Method 4 is more powerful than method 3.
  - In methods 3 and 4, the null hypothesis is that both the Group main effect and Group by Week interaction effects are zero.

- Method 1 vs. method 3
  - Methods 1 and 3 produce identical tests and estimates of effects (check this yourself).
Recommend to use method 1 because (1) it’s easier to construct test of the null hypothesis for method 1 in softwares, and (2) when there are subjects with missing baseline response, all of their data are excluded from method 3.

• Method 2 vs. method 4

  – Methods 2 and 4 are similar.
  – Method 2 is preferred over method 4 for the same reasons in the comparison of methods 1 and 3.
  – An additional constraint of method 4:

\[
\text{Cov}(Y_{i1}, Y_{i2}) = \text{Cov}(Y_{i1}, Y_{i3}) = \cdots = \text{Cov}(Y_{i1}, Y_{in})
\]

  – Methods 2 and 4 are only appropriate when it is reasonable to assume the baseline means are equal between groups (for randomized trial) or can be (conceptually at least) “held” equal between groups (for observational studies).
Method 1

```r
> full.1 <- gls (Lead ~ factor (Week) * Group, method = "ML",
+     data = tlcL,
+     correlation = corCompSymm (form = ~ 1 | ID))

> full.1
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) * Group
  Data: tlcL
  Log-likelihood: -1235.411

Coefficients:
             (Intercept)  factor(Week)1  factor(Week)4  factor(Week)6
GroupP factor(Week)1:GroupP factor(Week)4:GroupP factor(Week)6:GroupP
      -0.268        11.406        8.824        3.152

> anova(full.1)
Denom. DF: 392

numDF        F-value       p-value
(Intercept)          1     1533.2616       <.0001
factor(Week)         3     60.1967        <.0001
Group               1    24.9235         <.0001
factor(Week):Group   3    37.3452         <.0001

> reduced.1 <- gls (Lead ~ factor (Week) + Group, method = "ML",
+     data = tlcL,
+     correlation = corCompSymm (form = ~ 1 | ID))
> anova (full.1, reduced.1)

Model df   AIC    BIC  logLik   Test  L.Ratio p-value
full.1     10 2490.822 2530.736 -1235.411 1 vs 2  98.54295  <.0001
reduced.1   7 2583.365 2611.305 -1284.682
```

27
Method 2
This model is unusual since it includes the interaction terms without the main effects. R seems to be reluctant to do that. Using formula \( \text{Lead} \sim \text{factor(Week)} \ast \text{Group} - \text{Group} \) does not work.

```r
> tlcL$W1P <- (tlcL$Week == 1) & (tlcL$Group == "P")
> tlcL$W4P <- (tlcL$Week == 4) & (tlcL$Group == "P")
> tlcL$W6P <- (tlcL$Week == 6) & (tlcL$Group == "P")
> full.2 <- gls (Lead ~ factor(Week) + W1P + W4P + W6P,
+                  data = tlcL, method = "ML",
+                  correlation = corCompSymm (form = ~ 1 | ID))

> full.2
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) + W1P + W4P + W6P
  Data: tlcL
  Log-likelihood: -1235.432

Coefficients:
  (Intercept) factor(Week)1 factor(Week)4 factor(Week)6 W1PTRUE W4PTRUE W6PTRUE
  26.406000  -12.963798  -10.971798  -5.723798  11.297597  8.715597  3.043597

> anova(full.2)
Denom. DF: 393
  numDF  F-value p-value
(Intercept)  1 1540.6464  <.0001
factor(Week) 3  60.5016  <.0001
  W1P  1   71.5526  <.0001
  W4P  1   58.0041  <.0001
  W6P  1    8.0498   0.0048

> reduced.2 <- gls (Lead ~ factor(Week),
+                  data = tlcL, method = "ML",
+                  correlation = corCompSymm (form = ~ 1 | ID))

> anova (full.2, reduced.2)
  Model df AIC BIC logLik Test L.Ratio p-value
full.2  1  9  2488.864  2524.787 -1235.432
reduced.2 2  6  2604.437  2628.386 -1296.219 1 vs 2  121.5736  <.0001
```
Method 3

```r
> tlcL2 <- reshape (tlc, direction = "long", idvar = "ID",
+    varying = 4:6)
> names (tlcL2)[3:5] <- c("BaseLead", "Week", "Lead")
> tlcL2$ChangeLead <- tlcL2$Lead - tlcL2$BaseLead
> tlcL2 <- tlcL2[order (tlcL2$Group, tlcL2$ID, tlcL2$Week),]
>
> full.3 <- gls (ChangeLead ~ factor (Week) * Group, method = "ML",
+    data = tlcL2,
+    correlation = corCompSymm (form = ~ 1 | ID))

> full.3
Generalized least squares fit by maximum likelihood
Model: ChangeLead ~ factor(Week) * Group
Data: tlcL2
Log-likelihood: -923.4243

Coefficients:
(Intercept) factor(Week)4 factor(Week)6 GroupP
factor(Week)4:GroupP factor(Week)6:GroupP
-2.582 -8.254

> anova(full.3)
Denom. DF: 294
numDF F-value p-value
(Intercept) 1 158.68628 <.0001
factor(Week) 2 14.37365 <.0001
Group 1 65.97800 <.0001
factor(Week):Group 2 24.02362 <.0001

> reduced.3 <- gls (ChangeLead ~ factor (Week), method = "ML",
+    data = tlcL2,
+    correlation = corCompSymm (form = ~ 1 | ID))

> anova (full.3, reduced.3)
Model df AIC BIC logLik Test L.Ratio p-value
full.3 1 8 1862.849 1892.479 -923.4243 1 vs 2 96.94567 <.0001
reduced.3 2 5 1953.794 1972.313 -971.8971
```

Method 4

```r
> full.4 <- gls (Lead ~ factor (Week) * Group + BaseLead, method = "ML",
+       data = tlcL2,
+       correlation = corCompSymm (form = ~ 1 | ID))

> full.4
Generalized least squares fit by maximum likelihood
Model: Lead ~ factor(Week) * Group + BaseLead
Data: tlcL2
Log-likelihood: -921.2781

Coefficients:
   (Intercept) factor(Week)4 factor(Week)6 GroupP
       -7.8737215      1.9920000      7.2400000     11.3540533
   BaseLead factor(Week)4:GroupP factor(Week)6:GroupP
       0.8061689     -2.5820000     -8.2540000

> anova(full.4)
Denom. DF: 293

                  numDF F-value p-value
(Intercept)       1    1867.5630 <.0001
factor(Week)      2     14.2760 <.0001
Group             1     63.7806 <.0001
BaseLead          1     72.3672 <.0001
factor(Week):Group 2     23.8605 <.0001

> reduced.4 <- gls (Lead ~ factor (Week) + BaseLead,
+       method = "ML", data = tlcL2,
+       correlation = corCompSymm (form = ~ 1 | ID))

> anova (full.4, reduced.4)

                  Model df   AIC   BIC  logLik  Test L.Ratio p-value
full.4            1   9 1860.556 1893.890 -921.2781
reduced.4         2   6 1952.686 1974.908 -970.3428 1 vs 2  98.12944  <.0001
```
Inference for Marginal Mean Effects

- Approximate Wald tests (and associated confidence intervals) can be used (with robust variance estimates if so desired).

- For small sample sizes, approximate $t$- or $F$-tests may be more accurate. However, the estimation of the proper number of degrees of freedom is non-trivial (SAS includes four different methods in PROC MIXED.)

- For nested models, likelihood ratio test can be used. However, it is not valid if the models are fitted using REML rather than ML.

- Other model selection criteria, such as AIC or BIC, can be used.
Model Diagnosis

- The model diagnosis for general linear model is similar to linear models.
- Library `nlme` provides several functions for examining `gls` objects.

Residual Plots

> plot (full.1, ID ~ resid (.), id = 0.01)

- The errors should center at about zero and the variances should be approximately equal.
• Variance and mean relationship: slight increase in variance with time.
• An outlier with ID 40.
> plot (full.1, resid (.), type = "p") ~ fitted (.)) | Group,  
+ id = ~ ID == 40)  
> plot (full.1, Lead ~ fitted (.))

- There are several types of residuals, raw, Pearson and normalized.
Checking normality assumption:

\[ > \text{qqnorm} (\text{full.1}, \sim \text{resid}(\cdot)) \]
SAS sample code

data lead;
  infile 'C:\tlc.dat';
  input id group $ y1 y2 y3 y4;
  y=y1; time=0; output;
  y=y2; time=1; output;
  y=y3; time=4; output;
  y=y4; time=6; output;
  drop y1-y4;
run;

* Method 1;
proc mixed METHOD=ML;
  class id group time;
  model y=group time group*time/S CHISQ;
  repeated time/type=CS subject=id R RCORR;
run;
Further Reading

- Chapter 5 of Fitzmaurice, Laird and Ware (2004).