General Linear Models - Case Study

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 μg/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 μg/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 we will look at were a random sample of 100 children, with blood levels measured at baseline, weeks 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

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

Summary Statistics

Read in the data and compute some summary statistics

```
> 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$Group, function(x) cbind(mean = mean(x), sd = sd(x),
+       min=apply(x,2,min),max=apply(x,2,max)))
tlc$Group: A
        mean    sd  min  max
week.0 26.54 5.021 19.7 41.1
week.1 13.52 7.672  2.8 39.0
week.4 15.51 7.852  3.0 40.4
week.6 20.76 9.246  4.1 63.9
-------------------------------------------------------------------------------------------------------
tlc$Group: P
        mean    sd  min  max
week.0 26.27 5.024 19.7 38.1
week.1 24.66 5.461 14.9 40.8
week.4 24.07 5.753 15.3 38.6
week.6 23.65 5.640 13.5 43.3
```

Explore the Data

First we need convert it to long format:

```
> tlcL <- reshape (tlc, direction = "long", idvar = "ID", varying = 3:6)

> names (tlcL)[3:4] <- c("Week", "Lead")

> tlcL[95:105,]
       ID Group Week Lead
95.0   95     A    0 31.2
96.0   96     A    0 31.4
97.0   97     A    0 41.1
98.0   98     A    0 29.4
99.0   99     A    0 21.9
100.0 100     A    0 20.7
1.1     1     P    1 26.9
2.1     2     A    1 14.8
3.1     3     A    1 23.0
4.1     4     P    1 24.5
5.1     5     A    1  2.8
```

Scatterplot, by treatment group, with LOESS smoothing curve.

```
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, ...)})
```

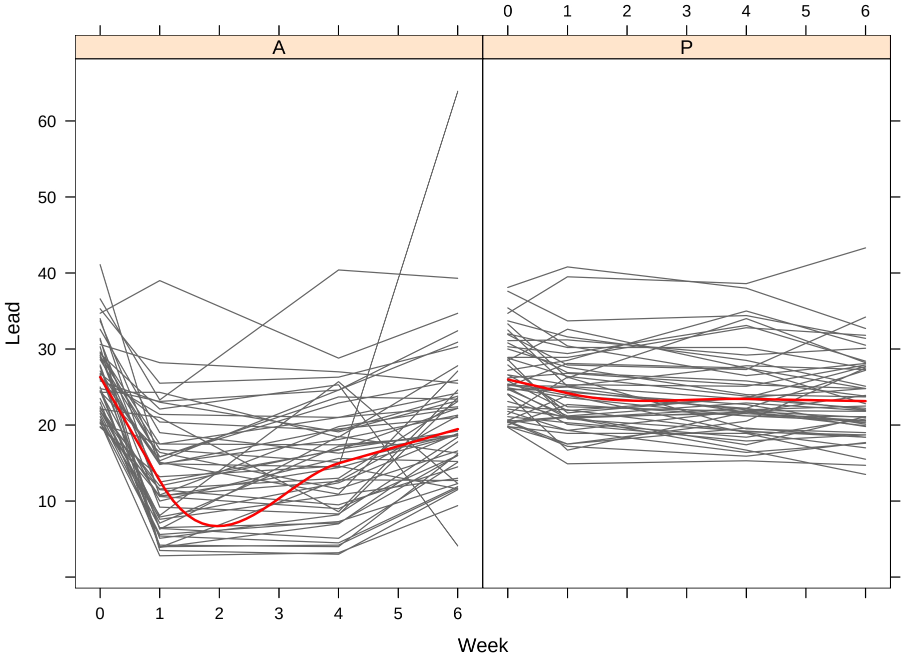


Figure 1: Plot of blood lead levels, by treatment group.

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

```
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)

pairs (subset (tlc, Group == "A", select = 3:6),diag.panel = panel.hist,
        upper.panel = panel.cor, lower.panel = panel.smooth)
pairs (subset (tlc, Group == "P", select = 3:6),diag.panel = panel.hist,
        upper.panel = panel.cor, lower.panel = panel.smooth)
```

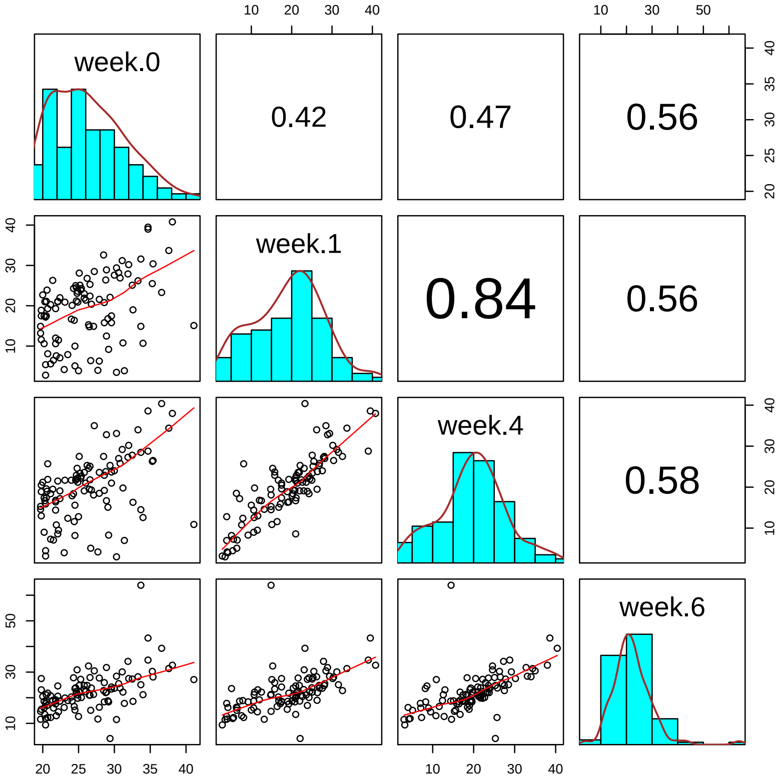
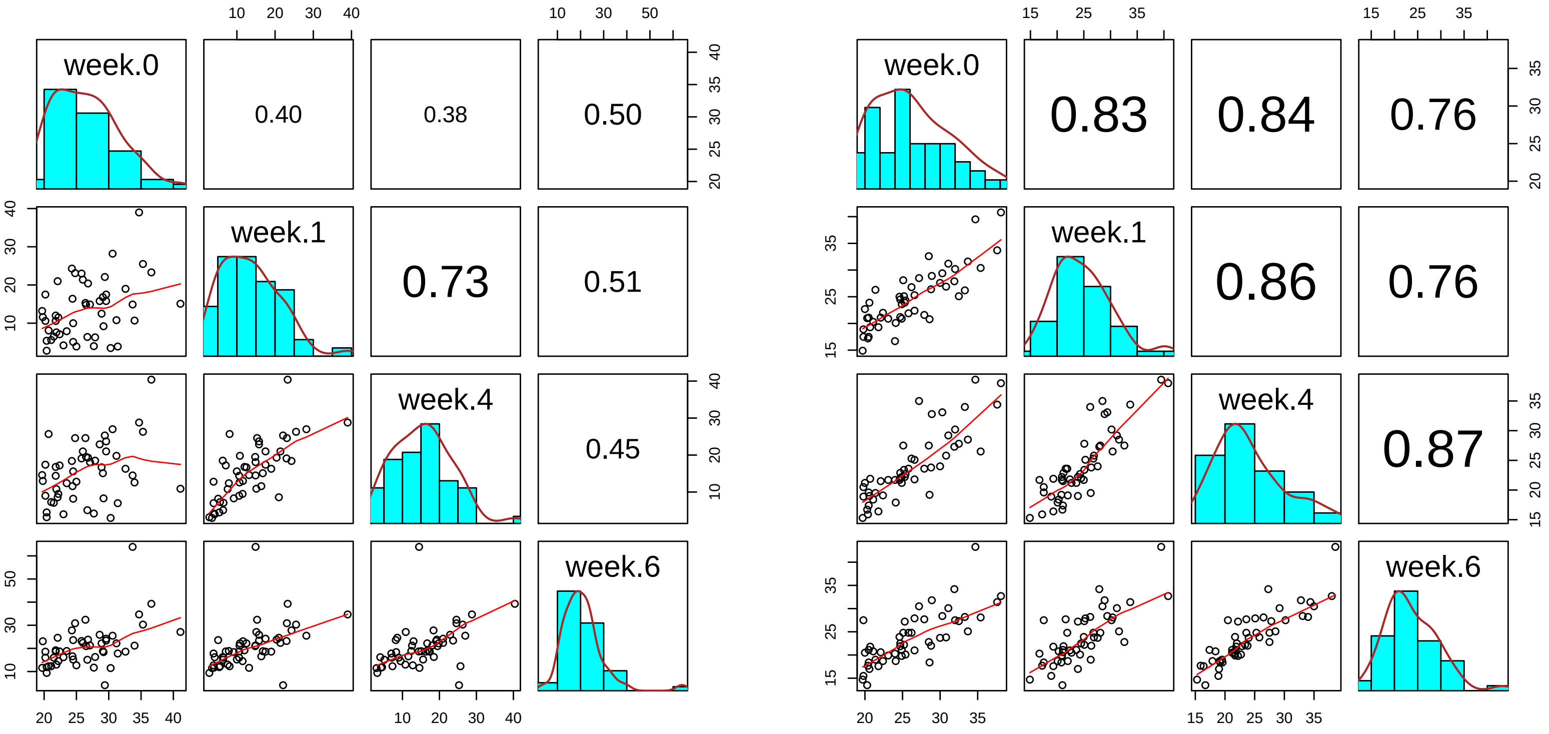


Figure 2: Pairwise scatter-plot of blood lead levels at baseline, week 1, 4 and 6 for children in TLC trial.



(a) Succimer

(b) Placebo

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 × 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 × 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 × treatment interaction effect.

Simple Linear Model

```
> trt <- factor(tlcL$Group,levels=sort(unique(tlcL$Group),T))
> temp <- lm (Lead ~ factor (Week) * trt, data = tlcL)
> summary (temp)

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

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

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          26.272      0.937  28.038  < 2e-16 ***
factor(Week)1        -1.612      1.325  -1.216   0.2245
factor(Week)4        -2.202      1.325  -1.662   0.0974 .
factor(Week)6        -2.626      1.325  -1.982   0.0482 *
trtA                  0.268      1.325   0.202   0.8398
factor(Week)1:trtA  -11.406      1.874  -6.086 2.75e-09 ***
factor(Week)4:trtA   -8.824      1.874  -4.709 3.47e-06 ***
factor(Week)6:trtA   -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  1090.9  24.850 9.701e-15 ***
trt                1  3110.9  3110.9  70.862 7.281e-16 ***
factor(Week):trt   3  2030.4   676.8  15.417 1.685e-09 ***
Residuals        392 17208.8    43.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Model Diagnosis

```
> par (mfrow = c (2, 2))
> plot (temp, which=1:4)
```

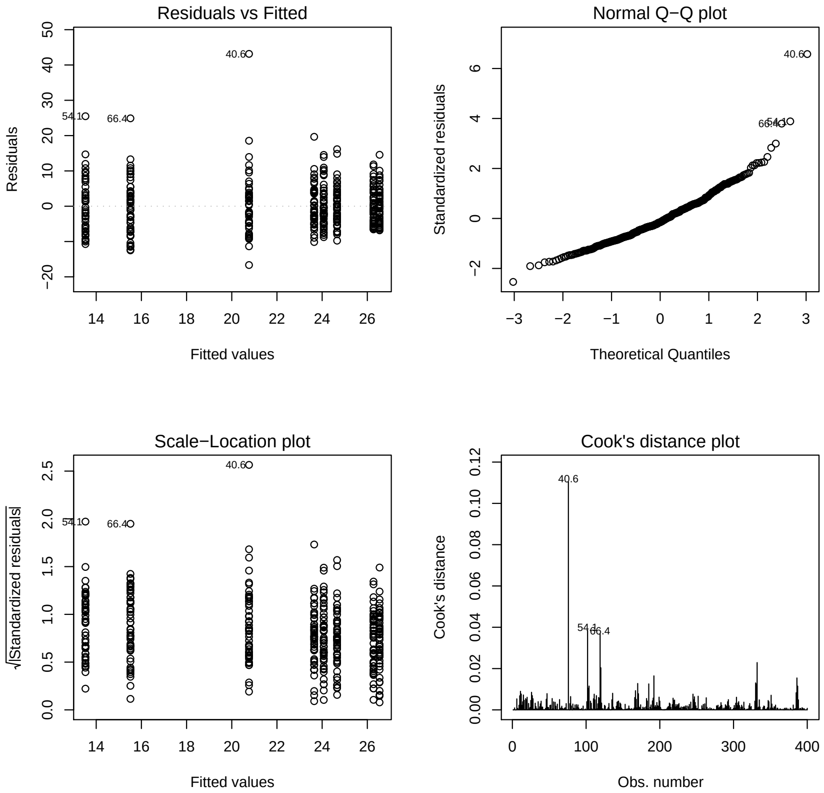


Figure 4: Simple Linear Model

GEE

In R, GEE (for linear model, it just means robust variance estimation) is implemented by library *gee*.

```
> 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.

```
> trt <- factor(tlcL$Group,levels=sort(unique(tlcL$Group),T))
> temp <- gee (Lead ˜ factor (Week) * trt, id = ID,
+              data = tlcL)
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1]  26.272  -1.612  -2.202  -2.626   0.268 -11.406  -8.824  -3.152
> summary (temp)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee 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) * trt, id = ID, data = tlcL)

Summary of Residuals:
     Min       1Q   Median       3Q      Max
-16.6620  -4.6205  -0.9930   3.6725  43.1380
Coefficients:
                   Estimate Naive S.E.    Naive z Robust S.E.    Robust z
(Intercept)          26.272  0.9370175 28.0378980   0.7033749  37.3513444
factor(Week)1        -1.612  1.3251428 -1.2164727   0.4330325  -3.7225846
factor(Week)4        -2.202  1.3251428 -1.6617077   0.4386752  -5.0196593
factor(Week)6        -2.626  1.3251428 -1.9816732   0.5278091  -4.9752834
trtA                  0.268  1.3251428  0.2022424   0.9944085   0.2695069
factor(Week)1:trtA  -11.406  1.8740349 -6.0863327   1.1086833 -10.2878794
factor(Week)4:trtA   -8.824  1.8740349 -4.7085569   1.1408849  -7.7343471
factor(Week)6:trtA   -3.152  1.8740349 -1.6819324   1.2439296  -2.5339055

Estimated Scale Parameter:  43.90009
Number of Iterations:  1

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

Notes

- 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 programing seems necessary.

- `temp$robust.variance` gives the full covariance matrix for β.

```
> temp$robust
                   (Intercept) factor(Week)1 factor(Week)4 factor(Week)6        trtA
(Intercept)         0.49473632   -0.04884672   -0.01922112   -0.07494656 -0.49473632
factor(Week)1      -0.04884672    0.18751712    0.10333952    0.08738576  0.04884672
factor(Week)4      -0.01922112    0.10333952    0.19243592    0.15252296  0.01922112
factor(Week)6      -0.07494656    0.08738576    0.15252296    0.27858248  0.07494656
trtA               -0.49473632    0.04884672    0.01922112    0.07494656  0.98884832
factor(Week)1:trtA  0.04884672   -0.18751712   -0.10333952   -0.08738576 -0.23983632
factor(Week)4:trtA  0.01922112   -0.10333952   -0.19243592   -0.15252296 -0.21662832
factor(Week)6:trtA  0.07494656   -0.08738576   -0.15252296   -0.27858248 -0.11854416
                   factor(Week)1:trtA factor(Week)4:trtA factor(Week)6:trtA
(Intercept)                0.04884672         0.01922112         0.07494656
factor(Week)1             -0.18751712        -0.10333952        -0.08738576
factor(Week)4             -0.10333952        -0.19243592        -0.15252296
factor(Week)6             -0.08738576        -0.15252296        -0.27858248
trtA                      -0.23983632        -0.21662832        -0.11854416
factor(Week)1:trtA         1.22917864         0.86059416         0.53279368
factor(Week)4:trtA         0.86059416         1.30161840         0.54664640
factor(Week)6:trtA         0.53279368         0.54664640         1.54736080
```

- Wald test for the Week:Group interaction.

```
> L <- rbind(c(0,0,0,0,0,1,0,0),c(0,0,0,0,0,0,1,0),c(0,0,0,0,0,0,0,1))
> lb <- L%*%(temp$coef)
> mid <- L%*%temp$robust%*%t(L)
> waldtemp <- t(lb)%*%solve(mid)%*%lb
> waldtemp #Wald statistics for Week:Group interaction
         [,1]
[1,] 109.9875
> 1-pchisq(waldtemp,3)
     [,1]
[1,]    0
```

GEE implemented in geeglm

In R, GEE is also implemented by a newer version `geepack` (the function name is `geese` corresponding to `gee`). The `geeglm` function in `geepack` follows the syntax of the `glm` function and has the `anova` method for comparing models by Wald tests.

```
> library(geepack)
> temp <- geeglm(Lead ~ factor (Week) * trt, id = ID,data = tlcL)
> summary (temp)

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

 Coefficients:
                   Estimate   Std.err         Wald       p(>W)
(Intercept)          26.272 0.7175089 1.340700e+03 0.000000e+00
factor(Week)1        -1.612 0.4417463 1.331632e+01 2.631060e-04
factor(Week)4        -2.202 0.4475124 2.421166e+01 8.630814e-07
factor(Week)6        -2.626 0.5358089 2.401981e+01 9.534954e-07
trtA                  0.268 1.0044556 7.118822e-02 7.896145e-01
factor(Week)1:trtA  -11.406 1.1121157 1.051881e+02 0.000000e+00
factor(Week)4:trtA   -8.824 1.1443119 5.946236e+01 1.243450e-14
factor(Week)6:trtA   -3.152 1.2473450 6.385564e+00 1.150522e-02
```

```
Estimated Scale Parameters:
            Estimate  Std.err
(Intercept) 43.02208 6.632351

Correlation: Structure = independenceNumber of clusters:   400   Maximum cluster size: 1
> anova(temp)
Analysis of 'Wald statistic' Table
Model: gaussian, link: identity
Response: Lead
Terms added sequentially (first to last)

                 Df      X2 P(>|Chi|)
factor(Week)      3  96.855     0.000
trt               1  25.527 4.363e-07
factor(Week):trt  3 109.376     0.000
```

Exchangeable correlation

```
> temp <- gee (Lead ~ factor (Week) * trt, id = ID,
+              corstr = "exchangeable", data = tlcL)
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1]  26.272  -1.612  -2.202  -2.626   0.268 -11.406  -8.824  -3.152
> summary (temp)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee 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) * trt, id = ID, data = tlcL,
    corstr = "exchangeable")

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

```
Coefficients:
                   Estimate Naive S.E.    Naive z Robust S.E.    Robust z
(Intercept)          26.272  0.9370175 28.0378980   0.7033749  37.3513444
factor(Week)1        -1.612  0.8470380 -1.9031023   0.4330325  -3.7225846
factor(Week)4        -2.202  0.8470380 -2.5996472   0.4386752  -5.0196593
factor(Week)6        -2.626  0.8470380 -3.1002150   0.5278091  -4.9752834
trtA                  0.268  1.3251428  0.2022424   0.9944085   0.2695069
factor(Week)1:trtA  -11.406  1.1978927 -9.5217212   1.1086833 -10.2878794
factor(Week)4:trtA   -8.824  1.1978927 -7.3662693   1.1408849  -7.7343471
factor(Week)6:trtA   -3.152  1.1978927 -2.6312875   1.2439296  -2.5339055

Estimated Scale Parameter:  43.90009
Number of Iterations:  1

Working Correlation
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.5914168 0.5914168 0.5914168
[2,] 0.5914168 1.0000000 0.5914168 0.5914168
[3,] 0.5914168 0.5914168 1.0000000 0.5914168
[4,] 0.5914168 0.5914168 0.5914168 1.0000000
```

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"
[1]  26.272  -1.612  -2.202  -2.626   0.268 -11.406  -8.824  -3.152
> summary (temp)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee 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) * trt, id = ID, data = tlcL,
    corstr = "unstructured")

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

```
Coefficients:
                  Estimate Naive S.E.    Naive z Robust S.E.    Robust z
(Intercept)         26.272  0.9370175 28.0378980   0.7033749  37.3513444
factor(Week)1       -1.612  0.9958441 -1.6187273   0.4330325  -3.7225846
factor(Week)4       -2.202  0.9838820 -2.2380732   0.4386752  -5.0196593
factor(Week)6       -2.626  0.9316319 -2.8187099   0.5278091  -4.9752834
trtA                 0.268  1.3251428  0.2022424   0.9944085   0.2695069
factor(Week)1:trtA -11.406  1.4083362 -8.0989182   1.1086833 -10.2878794
factor(Week)4:trtA  -8.824  1.3914193 -6.3417260   1.1408849  -7.7343471
factor(Week)6:trtA  -3.152  1.3175265 -2.3923618   1.2439296  -2.5339055

Estimated Scale Parameter:  43.90009
Number of Iterations:  1

Working Correlation
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.4352486 0.4487346 0.5057311
[2,] 0.4352486 1.0000000 0.8094551 0.6759677
[3,] 0.4487346 0.8094551 1.0000000 0.6975035
[4,] 0.5057311 0.6759677 0.6975035 1.0000000
```

- The GEE robust standard error estimates are robust to different working correlation structures.

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.

```
> temp <- gls (Lead ~ factor (Week) * trt,
+              data = tlcL, correlation = corCompSymm (form =  ~ 1 | ID))
> summary (temp)
Generalized least squares fit by REML
  Model: Lead ~ factor(Week) * trt
  Data: tlcL
       AIC      BIC    logLik
  2480.621 2520.334 -1230.311

Correlation Structure: Compound symmetry
 Formula: ~1 | ID
 Parameter estimate(s):
      Rho
0.5954401

Coefficients:
                     Value Std.Error   t-value p-value
(Intercept)         26.272 0.9370175 28.037898  0.0000
factor(Week)1       -1.612 0.8428574 -1.912542  0.0565
factor(Week)4       -2.202 0.8428574 -2.612542  0.0093
factor(Week)6       -2.626 0.8428574 -3.115592  0.0020
trtA                 0.268 1.3251428  0.202242  0.8398
factor(Week)1:trtA -11.406 1.1919804 -9.568950  0.0000
factor(Week)4:trtA  -8.824 1.1919804 -7.402807  0.0000
factor(Week)6:trtA  -3.152 1.1919804 -2.644339  0.0085
```

```
 Correlation:
                   (Intr) fc(W)1 fc(W)4 fc(W)6 trtA   f(W)1: f(W)4:
factor(Week)1      -0.450
factor(Week)4      -0.450  0.500
factor(Week)6      -0.450  0.500  0.500
trtA               -0.707  0.318  0.318  0.318
factor(Week)1:trtA  0.318 -0.707 -0.354 -0.354 -0.450
factor(Week)4:trtA  0.318 -0.354 -0.707 -0.354 -0.450  0.500
factor(Week)6:trtA  0.318 -0.354 -0.354 -0.707 -0.450  0.500  0.500

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.5147478 -0.6973588 -0.1498706  0.5542799  6.5106944

Residual standard error: 6.625714
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.

```
> temp <- gls (Lead ~ factor (Week) * trt, method = "ML",
+              data = tlcL, correlation = corCompSymm (form =  ~ 1 | ID))
> summary (temp)
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) * trt
  Data: tlcL
       AIC      BIC    logLik
  2490.822 2530.736 -1235.411

Correlation Structure: Compound symmetry
 Formula: ~1 | ID
 Parameter estimate(s):
  Rho
0.596
```

```
...

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.5402789 -0.7044388 -0.1513922  0.5599072  6.5767945

Residual standard error: 6.559122
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).

```
> anova(temp)
Denom. DF: 392
                 numDF   F-value p-value
(Intercept)          1 1533.2616  <.0001
factor(Week)         3   60.1967  <.0001
trt                  1   24.9235  <.0001
factor(Week):trt     3   37.3452  <.0001
```

```
> intervals (temp)
Approximate 95% confidence intervals

 Coefficients:
                        lower    est.       upper
(Intercept)         24.429792  26.272 28.11420832
factor(Week)1       -3.269086  -1.612  0.04508638
factor(Week)4       -3.859086  -2.202 -0.54491362
factor(Week)6       -4.283086  -2.626 -0.96891362
trtA                -2.337276   0.268  2.87327599
factor(Week)1:trtA -13.749474 -11.406 -9.06252597
factor(Week)4:trtA -11.167474  -8.824 -6.48052597
factor(Week)6:trtA  -5.495474  -3.152 -0.80852597
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
        lower      est.    upper
Rho 0.5000673 0.5954401 0.679613
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
   lower     est.    upper
5.937677 6.559122 7.245609
```

Bootstrap standard error estimates

The **Gls** function in library **rms** is an enhanced version of gls that can estimate standard error via bootstrap.
Note: since the data are "clustered", bootstrap is done at the cluster level.

```
> library(rms)
> temp <- Gls (Lead ~ factor (Week) * trt,
+               data = tlcL, correlation = corCompSymm (form =  ~ 1 | ID),
+               B = 1000)
Loading required package: nlme
> temp

Generalized Least Squares Fit by REML

Gls(model = Lead ~ factor(Week) * trt, data = tlcL, correlation = corCompSymm(form = ~1 |
    ID), B = 1000)

Obs      400    Log-restricted-likelihood -1230.31
Clusters 100    Model d.f.                       7
g      4.920    sigma                       6.6257
                d.f.                           392

Using bootstrap variance estimates

               Coef     S.E.   Wald Z Pr(>|Z|)
Intercept       26.2720 0.7167  36.66 <0.0001
```

```
Week=1           -1.6120 0.4400  -3.66 0.0002
Week=4           -2.2020 0.4437  -4.96 <0.0001
Week=6           -2.6260 0.5429  -4.84 <0.0001
trt=A             0.2680 0.9807   0.27 0.7846
Week=1 * trt=A -11.4060 1.0973 -10.39 <0.0001
Week=4 * trt=A  -8.8240 1.1273  -7.83 <0.0001
Week=6 * trt=A  -3.1520 1.2419  -2.54 0.0111

Correlation Structure: Compound symmetry
 Formula: ~1 | ID
 Parameter estimate(s):
     Rho
0.5954401

Bootstrap repetitions: 1000
Bootstraps were all balanced with respect to clusters
Ratio of Original Variances to Bootstrap Variances

     Intercept  Week=1  Week=4  Week=6  trt=A  Week=1 * trt=A  Week=4 * trt=A  Week=6 * trt=A
          1.71    3.67    3.61    2.41   1.83            1.18            1.12            0.92

Bootstrap Nonparametric 0.95 Confidence Limits for Correlation Parameters

    Lower Upper
Rho 0.454 0.725
```

```
> anova(temp)
               Wald Statistics          Response: Lead

 Factor                                    Chi-Square d.f. P
 Week  (Factor+Higher Order Factors)       200.05     6    <.0001
  All Interactions                         115.33     3    <.0001
 trt  (Factor+Higher Order Factors)        117.60     4    <.0001
  All Interactions                         115.33     3    <.0001
 Week * trt  (Factor+Higher Order Factors) 115.33     3    <.0001
 TOTAL                                     201.33     7    <.0001
```

Estimating the contrasts:

```
> tlcL$wc <- factor (tlcL$Week)
> tempB <- Gls (Lead ~ wc  * trt,
+                data = tlcL,
+                correlation = corCompSymm (form =  ~ 1 | ID))
> wcl <- levels (tlcL$wc)
> contrast (tempB,
+           list (trt = "A", wc = wcl),
+           list (trt = "P", wc = wcl))
 wc Contrast S.E.     Lower      Upper      Z     Pr(>|z|)
 0    0.268  1.325143  -2.329232  2.8652322  0.20 0.8397
 1  -11.138  1.325143 -13.735232 -8.5407678 -8.41 0.0000
 4   -8.556  1.325143 -11.153232 -5.9587678 -6.46 0.0000
 6   -2.884  1.325143  -5.481232 -0.2867678 -2.18 0.0295
```

Estimating the mean responses:

```
> newdata <- data.frame (expand.grid (wcl, c("A", "P")))
> names (newdata) <- c("wc", "trt")
> cbind(newdata,predict(tempB,newdata=newdata, conf.int=0.95))
  wc trt linear.predictors    lower    upper
1  0   A            26.540 24.70348 28.37652
2  1   A            13.522 11.68548 15.35852
3  4   A            15.514 13.67748 17.35052
4  6   A            20.762 18.92548 22.59852
5  0   P            26.272 24.43548 28.10852
6  1   P            24.660 22.82348 26.49652
7  4   P            24.070 22.23348 25.90652
8  6   P            23.646 21.80948 25.48252
```

```
tlc.means <- data.frame (newdata, predict (tempB, newdata = newdata,conf.int = 0.95))
names (tlc.means)[3] <- "Lead"
tlc.means[,2] <-c(rep("Succimer",4),rep("Placebo",4))
xYplot (Cbind (Lead, lower, upper) ~ as.numeric (as.character (wc)),
        group = trt,
        ylim = c(10, 30), xlab = "Weeks",ylab = "Mean Blood Lead Level",
        type='l',lwd=2,lty=c(2,1),col=c("red","blue"),label.curves=F,keys="lines",
        data = tlc.means)
Key(.8,.25,col=c("red","blue"),lwd=2,lty=c(2,1))
```

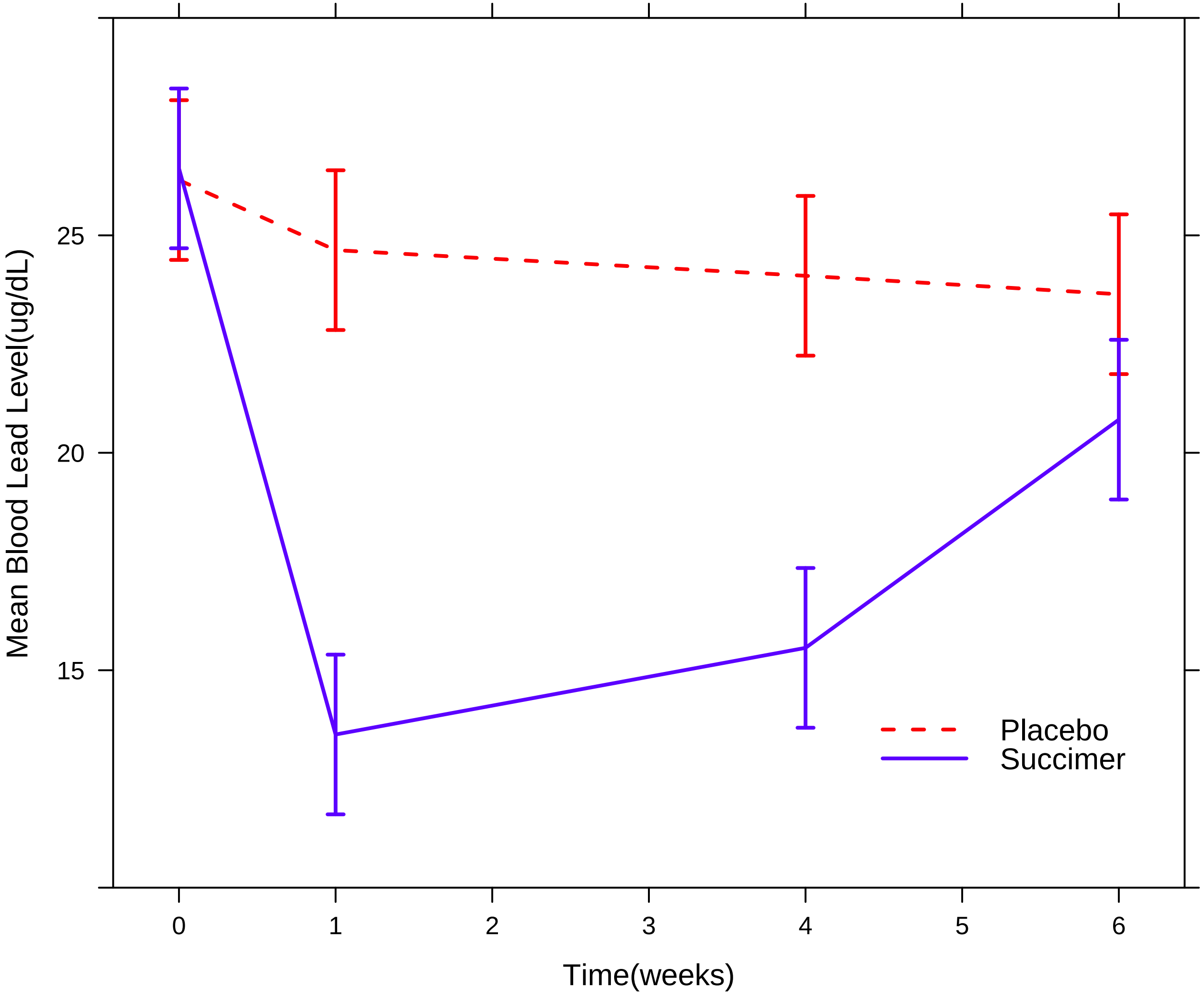


Figure 5: Mean Blood Lead Levels with 95% CI.

Dealing with Baseline Outcome

A simple example of pre-post data

When only two measurements are taken for each subject, say pre- and post-treatments (Y_{i0} and Y_{i1}) (i.e. $n = 2$). Let X be treatment indicator. Consider the three possible models:

$$\begin{aligned} Y_{i1} &= \mu + \beta_1 X_i + \epsilon_i & (1) \\ (Y_{i1} - Y_{i0}) &= \mu^* + \beta_1^* X_i + \epsilon_i & (2) \\ Y_{i1} &= \mu^{**} + \beta_1^{**} X_i + \beta_2 Y_{i0} + \epsilon_i & (3) \end{aligned}$$

- For randomized trials, it can be shown that $\beta_1 = \beta_1^* = \beta_1^{**}$.
- 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 β_1.

```
> trt <- factor(tlc$Group,levels=sort(unique(tlc$Group),T))
> summary (lm (week.1 ~ trt, data = tlc))
Call:
lm(formula = week.1 ~ trt, data = tlc)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.6600     0.9418  26.185  < 2e-16 ***
trtA        -11.1380     1.3319  -8.363 4.24e-13 ***

> y21 <- tlc$week.1-tlc$week.0
> summary (lm (y21 ~ trt))
Call:
lm(formula = y21 ~ trt)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6120     0.7919  -2.036   0.0445 *
trtA        -11.4060     1.1199 -10.184   <2e-16 ***

> summary (lm (week.1 ~ trt + week.0, data = tlc))
Call:
lm(formula = week.1 ~ trt + week.0, data = tlc)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7600     3.0050   1.584    0.116
trtA        -11.3410     1.0991 -10.318  < 2e-16 ***
week.0        0.7575     0.1105   6.855 6.61e-10 ***
```

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

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

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

Method 3: Subtract the baseline response from all remaining responses.

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

Method 1

```
> trt <- factor(tlcL$Group,levels=sort(unique(tlcL$Group),T))
> full.1 <- gls (Lead ~ factor (Week) * trt, method = "ML",
+                data = tlcL,
+                correlation = corCompSymm (form =  ~ 1 | ID))
> full.1
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) * trt
  Data: tlcL
  Log-likelihood: -1235.411

Coefficients:
       (Intercept)      factor(Week)1      factor(Week)4      factor(Week)6
            26.272             -1.612             -2.202             -2.626
              trtA factor(Week)1:trtA factor(Week)4:trtA factor(Week)6:trtA
             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
trt                  1   24.9235  <.0001
factor(Week):trt     3   37.3452  <.0001
> reduced.1 <- gls (Lead ~ factor (Week) + trt, 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        1 10 2490.822 2530.736 -1235.411
reduced.1     2  7 2583.365 2611.305 -1284.682 1 vs 2  98.543  <.0001
```

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 `Lead ~ factor(Week) * Group - Group` does not work.

```
> tlcL$W1A <- (tlcL$Week == 1) & (tlcL$Group == "A")
> tlcL$W4A <- (tlcL$Week == 4) & (tlcL$Group == "A")
> tlcL$W6A <- (tlcL$Week == 6) & (tlcL$Group == "A")
>
> full.2 <- gls (Lead ~ factor (Week) + W1A + W4A + W6A,
+                data = tlcL, method = "ML",
+                correlation = corCompSymm (form =  ~ 1 | ID))
> full.2
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) + W1A + W4A + W6A
  Data: tlcL
  Log-likelihood: -1235.432

Coefficients:
  (Intercept) factor(Week)1 factor(Week)4 factor(Week)6       W1ATRUE       W4ATRUE
    26.406000     -1.666202     -2.256202     -2.680202    -11.297597     -8.715597
      W6ATRUE
    -3.043597
```

```
> anova(full.2)
Denom. DF: 393
             numDF   F-value p-value
(Intercept)      1 1540.6464  <.0001
factor(Week)     3   60.5016  <.0001
W1A              1   71.5526  <.0001
W4A              1   58.0041  <.0001
W6A              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 2488.863 2524.787 -1235.432
reduced.2     2  6 2604.437 2628.386 -1296.219 1 vs 2 121.5737  <.0001
```

Method 3

```
> 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),]
> trt <- factor(tlcL2$Group,levels=sort(unique(tlcL2$Group),T))
> full.3 <- gls (ChangeLead ~ factor (Week) * trt, method = "ML",
+                 data = tlcL2,
+                 correlation = corCompSymm (form =  ~ 1 | ID))
> full.3
Generalized least squares fit by maximum likelihood
  Model: ChangeLead ~ factor(Week) * trt
  Data: tlcL2
  Log-likelihood: -923.4243

Coefficients:
       (Intercept)      factor(Week)4      factor(Week)6               trtA
            -1.612             -0.590             -1.014            -11.406
factor(Week)4:trtA factor(Week)6:trtA
             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
trt                  1  65.97800  <.0001
factor(Week):trt     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
reduced.3     2  5 1953.794 1972.313 -971.8971 1 vs 2 96.94567  <.0001
```

Method 4

```
> full.4 <- gls (Lead ~ factor (Week) * trt + BaseLead, method = "ML",
+                data = tlcL2,
+                correlation = corCompSymm (form =  ~ 1 | ID))
> full.4
Generalized least squares fit by maximum likelihood
  Model: Lead ~ factor(Week) * trt + BaseLead
  Data: tlcL2
  Log-likelihood: -921.2781

Coefficients:
       (Intercept)      factor(Week)4      factor(Week)6               trtA
         3.4803318         -0.5900000         -1.0140000        -11.3540533
          BaseLead factor(Week)4:trtA factor(Week)6:trtA
         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
trt                  1   63.7806  <.0001
BaseLead             1   72.3672  <.0001
factor(Week):trt     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
```

- Method 1 vs. method 2
  - In methods 1 and 2, the null hypothesis is that the Group by Week interaction effects are zero.
  - There is no treatment group main effect in 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.
- Method 3 vs. method 4
  - Methods 3 and 4 do not retain the baseline response as part of the outcome.
  - In methods 3 and 4, the null hypothesis is that both the Group main effect and Group by Week interaction effects are zero.
  - The interpretation of the regression coefficients is different, for all three factors in the model!
  - Method 4 is more powerful than method 3.

- 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).

Inference for Marginal Mean Effects

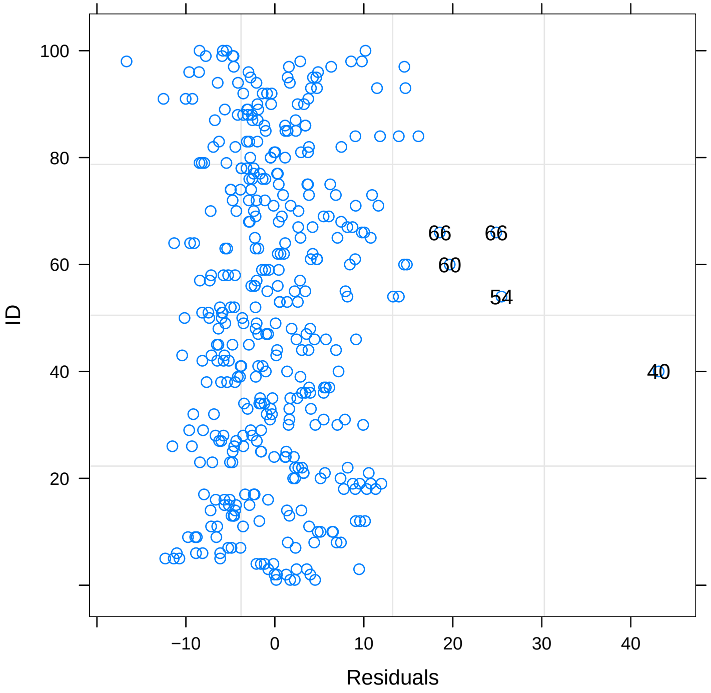
- Wald tests (and associated confidence intervals) can be used (with robust variance estimates if so desired).
- For nested models, likelihood ratio test can be used. However, it is not valid if the models are fitted using REML rather than ML when the constraints are on the mean.
- Likelihood ratio test can be used for hypotheses about the covariance parameters. Do not recommend testing covariance parameters using Wald tests because the distribution of the Wald test statistic for a variance parameter does not have an approximate normal distribution when sample size is small and the variance is close to zero.
- Other model selection criteria, such AIC or BIC, can be used for un-nested models.

Model Diagnosis

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

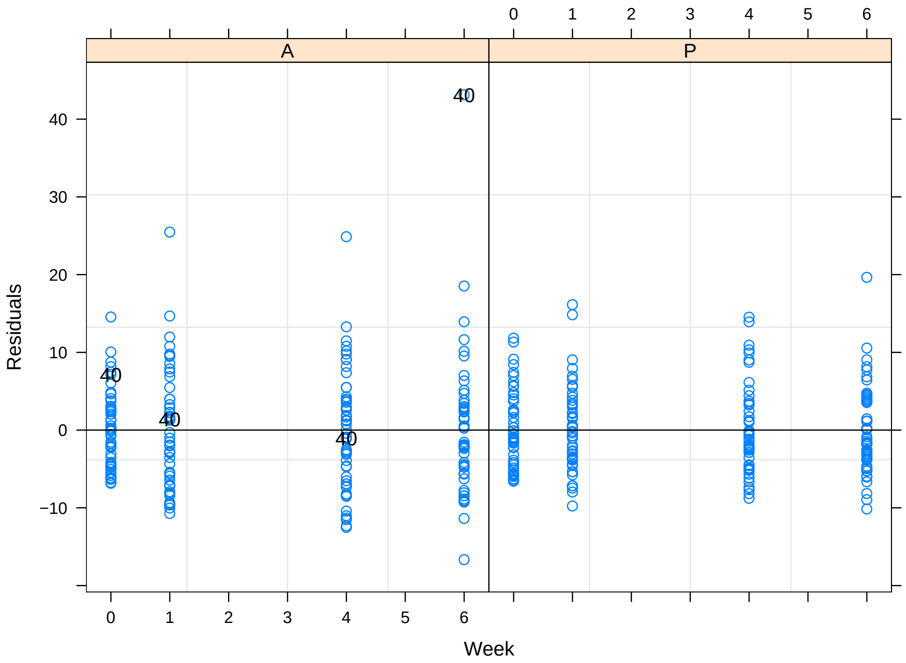
Residual Plots

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



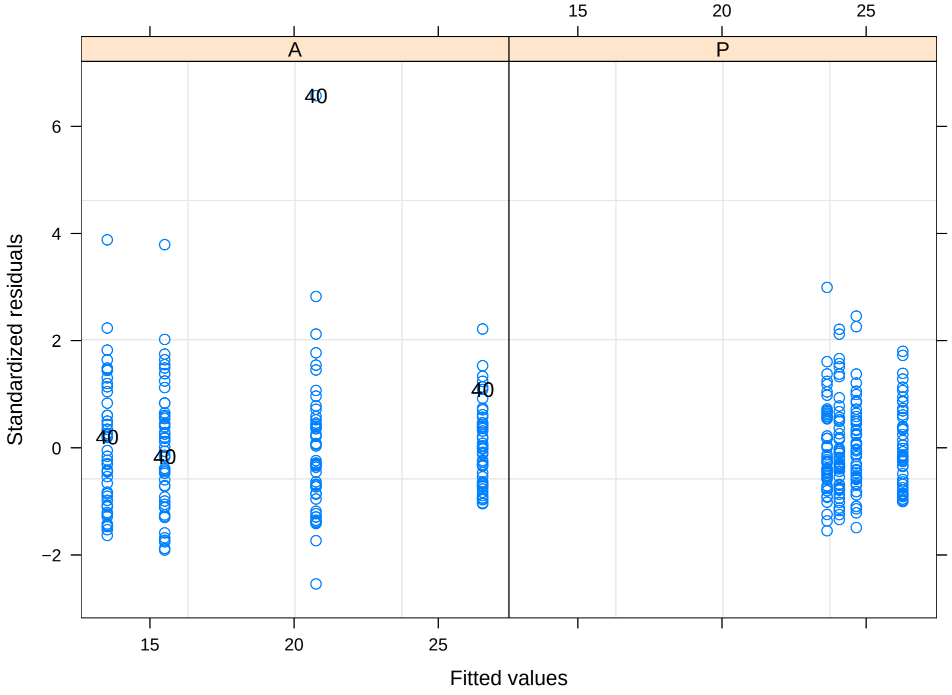
- The errors should center at about zero and the variances should be approximately equal.

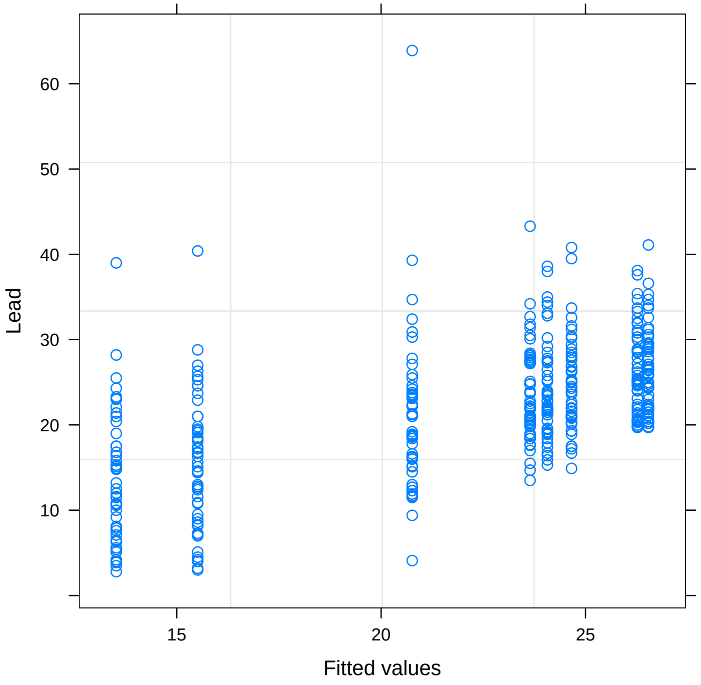
```
> plot (full.1, resid (.) ~ Week | Group, abline = 0,
+        id = ~ ID == 40)
```



- 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 (.))
```

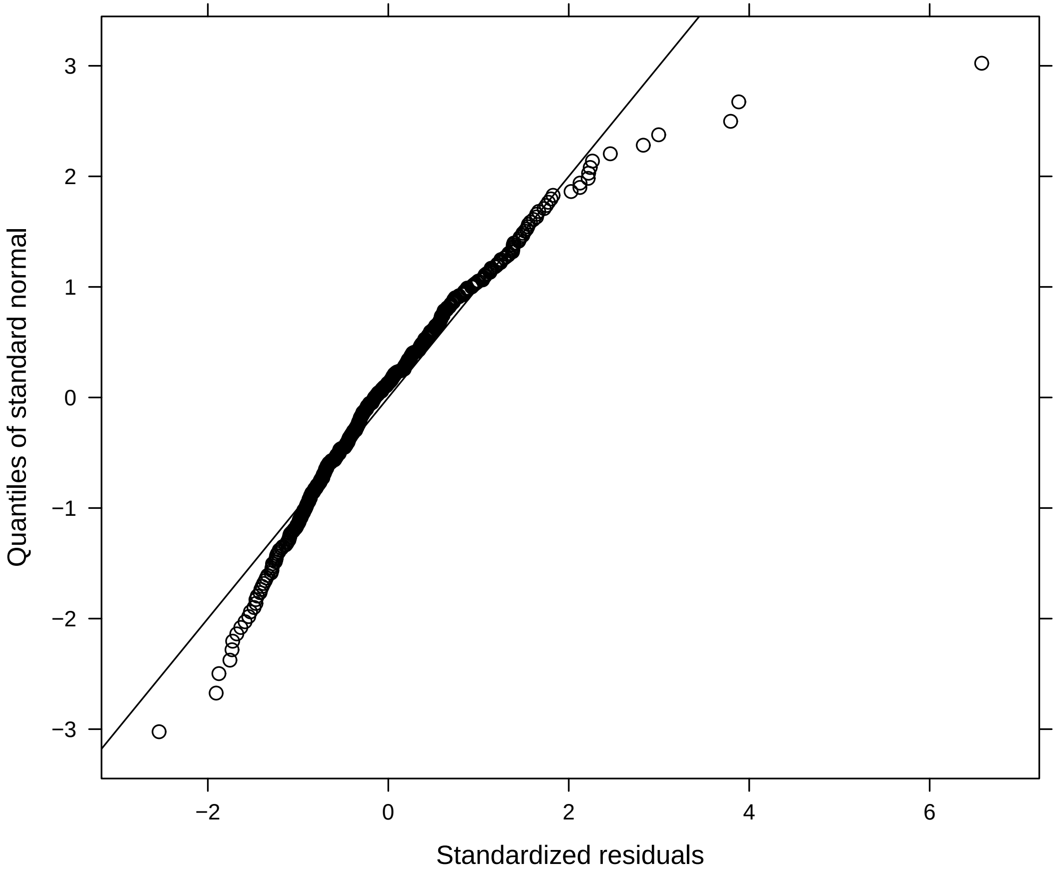




- There are several types of residuals, *raw*, *Pearson* and *normalized*.

Checking normality assumption:

```
> qqnorm (full.1,abline = c(0,1))
```



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: optional

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