A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 14

Analyzing Longitudinal Data II – Generalized Estimation Equations and Linear Mixed Effect Models: Treating Respiratory Illness and Epileptic Seizures

14.1 Introduction

14.2 Methods for Non-normal Distributions

14.3 Analysis Using R: GEE

14.3.1 Beat the Blues Revisited

To use the `gee` function, package **gee** (Carey et al., 2013) has to be installed and attached:

```
R> library("gee")
```

The `gee` function is used in a similar way to the `lme` function met in Chapter 12 with the addition of the features of the `glm` function that specify the appropriate error distribution for the response and the implied link function, and an argument to specify the structure of the working correlation matrix. Here we will fit an independence structure and then an exchangeable structure. The R code for fitting generalized estimation equations to the `BtheB_long` data (as constructed in Chapter 12) with identity working correlation matrix is as follows (note that the `gee` function assumes the rows of the *data.frame* `BtheB_long` to be ordered with respect to subjects):

```
R> osub <- order(as.integer(BtheB_long$subject))
R> BtheB_long <- BtheB_long[osub,]
R> btb_gee <- gee(bdi ~ bdi.pre + trt + length + drug,
+      data = BtheB_long, id = subject, family = gaussian,
+      corstr = "independence")
```

and with exchangeable correlation matrix:

```
R> btb_gee1 <- gee(bdi ~ bdi.pre + trt + length + drug,
+      data = BtheB_long, id = subject, family = gaussian,
+      corstr = "exchangeable")
```

The `summary` method can be used to inspect the fitted models; the results are shown in Figures 14.1 and 14.2.

```
R> summary(btb_gee)

 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 = bdi ~ bdi.pre + trt + length + drug, id = subject,
    data = BtheB_long, family = gaussian, corstr = "independence")

Summary of Residuals:
    Min      1Q  Median      3Q     Max
-21.650  -5.849   0.113   5.584  28.187


Coefficients:
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)    3.569     1.4833    2.41      2.2695    1.572
bdi.pre        0.582     0.0564   10.32      0.0916    6.355
trtBtheB      -3.237     1.1296   -2.87      1.7746   -1.824
length>6m      1.458     1.1380    1.28      1.4826    0.983
drugYes       -3.741     1.1766   -3.18      1.7827   -2.099

Estimated Scale Parameter:  79.3
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
```

Figure 14.1 R output of the `summary` method for the `btb_gee` model (slightly abbreviated).

14.3.2 Respiratory Illness

The baseline status, i.e., the status for `month == 0`, will enter the models as an explanatory variable and thus we have to rearrange the *data.frame* `respiratory` in order to create a new variable `baseline`:

```
R> data("respiratory", package = "HSAUR3")
R> resp <- subset(respiratory, month > "0")
R> resp$baseline <- rep(subset(respiratory, month == "0")$status,
+                       rep(4, 111))
R> resp$nstat <- as.numeric(resp$status == "good")
R> resp$month <- resp$month[, drop = TRUE]
```

The new variable `nstat` is simply a dummy coding for a poor respiratory status. Now we can use the data `resp` to fit a logistic regression model and GEE models with an independent and an exchangeable correlation structure as follows.

```
R> resp_glm <- glm(status ~ centre + trt + gender + baseline
```

```
R> summary(btb_gee1)
 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 = bdi ~ bdi.pre + trt + length + drug, id = subject,
    data = BtheB_long, family = gaussian, corstr = "exchangeable")

Summary of Residuals:
   Min     1Q Median     3Q    Max
-23.96  -6.64  -1.11   4.26  25.45


Coefficients:
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)    3.023     2.3039  1.3122      2.2320   1.3544
bdi.pre        0.648     0.0823  7.8741      0.0835   7.7583
trtBtheB      -2.169     1.7664 -1.2281      1.7361  -1.2495
length>6m     -0.111     1.7309 -0.0643      1.5509  -0.0718
drugYes       -3.000     1.8257 -1.6430      1.7316  -1.7323

Estimated Scale Parameter:  81.7
Number of Iterations:  5

Working Correlation
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.676 0.676 0.676
[2,] 0.676 1.000 0.676 0.676
[3,] 0.676 0.676 1.000 0.676
[4,] 0.676 0.676 0.676 1.000
```

Figure 14.2 R output of the `summary` method for the `btb_gee1` model (slightly abbreviated).

```
+      + age, data = resp, family = "binomial")
R> resp_gee1 <- gee(nstat ~ centre + trt + gender + baseline
+      + age, data = resp, family = "binomial", id = subject,
+      corstr = "independence", scale.fix = TRUE,
+      scale.value = 1)
R> resp_gee2 <- gee(nstat ~ centre + trt + gender + baseline
+      + age, data = resp, family = "binomial", id = subject,
+      corstr = "exchangeable", scale.fix = TRUE,
+      scale.value = 1)
```

The estimated treatment effect taken from the exchangeable structure GEE model is 1.299 which, using the robust standard errors, has an associated 95% confidence interval

```
R> se <- summary(resp_gee2)$coefficients["trttrt",
+                                        "Robust S.E."]
R> coef(resp_gee2)["trttrt"] +
+      c(-1, 1) * se * qnorm(0.975)
```

```
R> summary(resp_glm)

Call:
glm(formula = status ~ centre + trt + gender + baseline + age,
    family = "binomial", data = resp)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.90017    0.33765   -2.67   0.0077
centre2       0.67160    0.23957    2.80   0.0051
trttrt        1.29922    0.23684    5.49  4.1e-08
gendermale    0.11924    0.29467    0.40   0.6857
baselinegood  1.88203    0.24129    7.80  6.2e-15
age          -0.01817    0.00886   -2.05   0.0404

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 608.93  on 443  degrees of freedom
Residual deviance: 483.22  on 438  degrees of freedom
AIC: 495.2

Number of Fisher Scoring iterations: 4
```

Figure 14.3 R output of the `summary` method for the `resp_glm` model.

```
[1] 0.612 1.987
```

These values reflect effects on the log-odds scale. Interpretation becomes simpler if we exponentiate the values to get the effects in terms of odds. This gives a treatment effect of 3.666 and a 95% confidence interval of

```
R> exp(coef(resp_gee2)["trttrt"] +
+      c(-1, 1) * se * qnorm(0.975))
```

```
[1] 1.84 7.29
```

The odds of achieving a 'good' respiratory status with the active treatment is between about twice and seven times the corresponding odds for the placebo.

14.3.3 Epilepsy

Moving on to the count data in `epilepsy` from Table **??**, we begin by calculating the means and variances of the number of seizures for all interactions between treatment and period:

```
R> data("epilepsy", package = "HSAUR3")
R> itp <- interaction(epilepsy$treatment, epilepsy$period)
R> tapply(epilepsy$seizure.rate, itp, mean)
  placebo.1 Progabide.1   placebo.2 Progabide.2   placebo.3
       9.36        8.58        8.29        8.42        8.79
Progabide.3   placebo.4 Progabide.4
       8.13        7.96        6.71
R> tapply(epilepsy$seizure.rate, itp, var)
  placebo.1 Progabide.1   placebo.2 Progabide.2   placebo.3
      102.8       332.7        66.7       140.7       215.3
```

```
R> summary(resp_gee1)
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = nstat ~ centre + trt + gender + baseline + age,
    id = subject, data = resp, family = "binomial", corstr = "independence",
    scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
    Min      1Q  Median      3Q     Max
-0.9313 -0.3062  0.0897  0.3302  0.8431


Coefficients:
             Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.9002    0.33765  -2.666       0.460   -1.956
centre2        0.6716    0.23957   2.803       0.357    1.882
trttrt         1.2992    0.23684   5.486       0.351    3.704
gendermale     0.1192    0.29467   0.405       0.443    0.269
baselinegood   1.8820    0.24129   7.800       0.350    5.376
age           -0.0182    0.00886  -2.049       0.013   -1.397

Estimated Scale Parameter:  1
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
```

Figure 14.4 R output of the `summary` method for the `resp_gee1` model (slightly abbreviated).

```
Progabide.3   placebo.4 Progabide.4
      193.0        58.2       126.9
```

Some of the variances are considerably larger than the corresponding means, which for a Poisson variable may suggest that overdispersion may be a problem, see Chapter 7.

We can now fit a Poisson regression model to the data assuming independence using the `glm` function. We also use the GEE approach to fit an independence structure, followed by an exchangeable structure using the following R code:

```
R> per <- rep(log(2),nrow(epilepsy))
R> epilepsy$period <- as.numeric(epilepsy$period)
R> names(epilepsy)[names(epilepsy) == "treatment"] <- "trt"
R> fm <- seizure.rate ~ base + age + trt + offset(per)
R> epilepsy_glm <- glm(fm, data = epilepsy, family = "poisson")
```

```
R> summary(resp_gee2)
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Call:
gee(formula = nstat ~ centre + trt + gender + baseline + age,
    id = subject, data = resp, family = "binomial", corstr = "exchangeable",
    scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
    Min      1Q  Median      3Q     Max
-0.9313 -0.3062  0.0897  0.3302  0.8431


Coefficients:
             Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.9002     0.4785  -1.881       0.460   -1.956
centre2        0.6716     0.3395   1.978       0.357    1.882
trttrt         1.2992     0.3356   3.871       0.351    3.704
gendermale     0.1192     0.4176   0.286       0.443    0.269
baselinegood   1.8820     0.3419   5.504       0.350    5.376
age           -0.0182     0.0126  -1.446       0.013   -1.397

Estimated Scale Parameter:  1
Number of Iterations:  1

Working Correlation
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.336 0.336 0.336
[2,] 0.336 1.000 0.336 0.336
[3,] 0.336 0.336 1.000 0.336
[4,] 0.336 0.336 0.336 1.000
```

Figure 14.5 R output of the `summary` method for the `resp_gee2` model (slightly abbreviated).

```
R> epilepsy_gee1 <- gee(fm, data = epilepsy, family = "poisson",
+      id = subject, corstr = "independence", scale.fix = TRUE,
+      scale.value = 1)
R> epilepsy_gee2 <- gee(fm, data = epilepsy, family = "poisson",
+      id = subject, corstr = "exchangeable", scale.fix = TRUE,
+      scale.value = 1)
R> epilepsy_gee3 <- gee(fm, data = epilepsy, family = "poisson",
+      id = subject, corstr = "exchangeable", scale.fix = FALSE,
+      scale.value = 1)
```

As usual we inspect the fitted models using the `summary` method, the results are given in Figures 14.8, 14.9, 14.10, and 14.11.

```
R> layout(matrix(1:2, nrow = 1))
R> ylim <- range(epilepsy$seizure.rate)
R> placebo <- subset(epilepsy, treatment == "placebo")
R> progabide <- subset(epilepsy, treatment == "Progabide")
R> boxplot(seizure.rate ~ period, data = placebo,
+          ylab = "Number of seizures",
+          xlab = "Period", ylim = ylim, main = "Placebo")
R> boxplot(seizure.rate ~ period, data = progabide,
+          main  = "Progabide", ylab = "Number of seizures",
+          xlab = "Period", ylim = ylim)
```

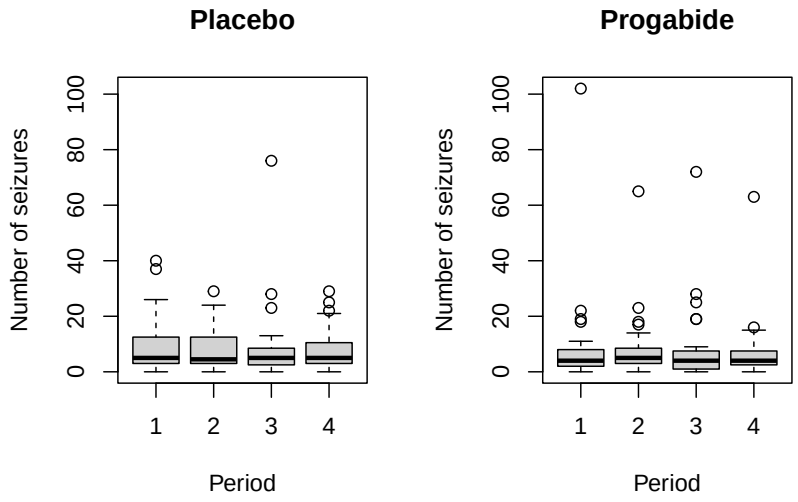

Figure 14.6 Boxplots of numbers of seizures in each two-week period post randomization for placebo and active treatments.

14.4 Analysis Using R: Random Effects

As an example of using generalized mixed models for the analysis of longitudinal data with a non-normal response, the following logistic model will be fitted to the respiratory illness data

$$\begin{aligned}\text{logit}(\text{P}(\text{status} = \text{good})) &= \beta_0 + \beta_1 \text{treatment} + \beta_2 \text{time} + \beta_3 \text{gender} \\ &\quad + \beta_4 \text{age} + \beta_5 \text{centre} + \beta_6 \text{baseline} + u\end{aligned}$$

where u is a subject-specific random effect.

The necessary R code for fitting the model using the `glmer` function from package **lme4** (Bates and Sarkar, 2014, Bates, 2005) is:

```
R> library("lme4")
```

```
R> layout(matrix(1:2, nrow = 1))
R> ylim <- range(log(epilepsy$seizure.rate + 1))
R> boxplot(log(seizure.rate + 1) ~ period, data = placebo,
+          main = "Placebo", ylab = "Log number of seizures",
+          xlab = "Period", ylim = ylim)
R> boxplot(log(seizure.rate + 1) ~ period, data = progabide,
+          main = "Progabide", ylab = "Log number of seizures",
+          xlab = "Period", ylim = ylim)
```

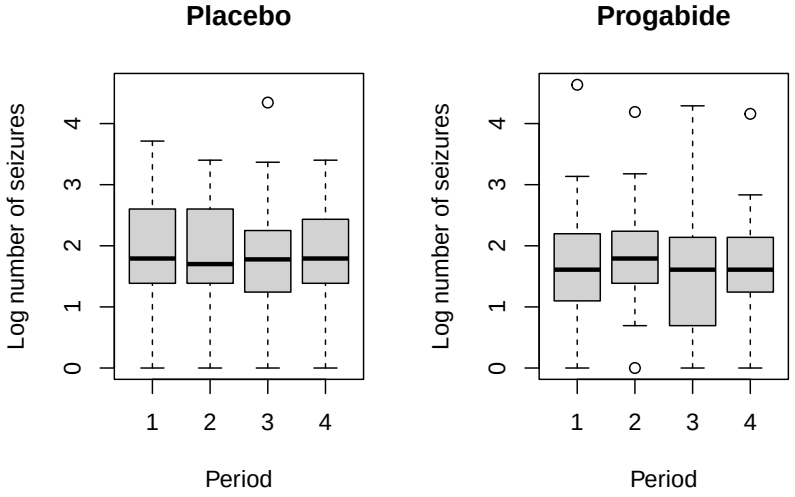

Figure 14.7 Boxplots of log of numbers of seizures in each two-week period post randomization for placebo and active treatments.

```
R> resp_lmer <- glmer(status ~ baseline + month +
+      trt + gender + age + centre + (1 | subject),
+      family = binomial(), data = resp)
R> exp(fixef(resp_lmer))
```

```
 (Intercept) baselinegood      month.L      month.Q
       0.191       21.954        0.816        0.972
     month.C       trttrt   gendermale          age
       0.701        8.725        1.269        0.975
     centre2
       2.825
```

The significance of the effects as estimated by this random effects model and by the GEE model described in Section 14.3.2 is generally similar. But as expected from our previous discussion the estimated coefficients are substantially larger. While the estimated effect of treatment on a randomly sampled

```
R> summary(epilepsy_glm)

Call:
glm(formula = fm, family = "poisson", data = epilepsy)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.130616   0.135619   -0.96   0.3355
base          0.022652   0.000509   44.48  < 2e-16
age           0.022740   0.004024    5.65  1.6e-08
trtProgabide -0.152701   0.047805   -3.19   0.0014

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2521.75  on 235  degrees of freedom
Residual deviance:  958.46  on 232  degrees of freedom
AIC: 1732

Number of Fisher Scoring iterations: 5
```

Figure 14.8 R output of the `summary` method for the `epilepsy_glm` model.

individual, given the set of observed covariates, is estimated by the marginal model using GEE to increase the log-odds of being disease free by 1.299, the corresponding estimate from the random effects model is 2.166. These are not inconsistent results but reflect the fact that the models are estimating different parameters. The random effects estimate is conditional upon the patient's random effect, a quantity that is rarely known in practice. Were we to examine the log-odds of the average predicted probabilities with and without treatment (averaged over the random effects) this would give an estimate comparable to that estimated within the marginal model.

```
R> summary(epilepsy_gee1)

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

Model:
 Link:                      Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Independent

Call:
gee(formula = fm, id = subject, data = epilepsy, family = "poisson",
    corstr = "independence", scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
   Min     1Q Median     3Q    Max
-4.920  0.181  1.707  4.885 69.966


Coefficients:
             Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1306   0.135619  -0.963     0.36515   -0.358
base           0.0227   0.000509  44.476     0.00124   18.332
age            0.0227   0.004024   5.651     0.01158    1.964
trtProgabide  -0.1527   0.047805  -3.194     0.17111   -0.892

Estimated Scale Parameter:  1
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
```

Figure 14.9 R output of the `summary` method for the `epilepsy_gee1` model (slightly abbreviated).

```
R> summary(epilepsy_gee2)
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Exchangeable

Call:
gee(formula = fm, id = subject, data = epilepsy, family = "poisson",
    corstr = "exchangeable", scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
   Min     1Q Median     3Q    Max
-4.920  0.181  1.707  4.885 69.966


Coefficients:
             Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1306   0.200442  -0.652     0.36515   -0.358
base           0.0227   0.000753  30.093     0.00124   18.332
age            0.0227   0.005947   3.824     0.01158    1.964
trtProgabide  -0.1527   0.070655  -2.161     0.17111   -0.892

Estimated Scale Parameter:  1
Number of Iterations:  1

Working Correlation
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.395 0.395 0.395
[2,] 0.395 1.000 0.395 0.395
[3,] 0.395 0.395 1.000 0.395
[4,] 0.395 0.395 0.395 1.000
```

Figure 14.10 R output of the `summary` method for the `epilepsy_gee2` model (slightly abbreviated).

```
R> summary(epilepsy_gee3)

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

Model:
 Link:                      Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Exchangeable

Call:
gee(formula = fm, id = subject, data = epilepsy, family = "poisson",
    corstr = "exchangeable", scale.fix = FALSE, scale.value = 1)

Summary of Residuals:
   Min     1Q Median     3Q    Max
-4.920  0.181  1.707  4.885 69.966


Coefficients:
             Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1306     0.4522  -0.289     0.36515   -0.358
base           0.0227     0.0017  13.339     0.00124   18.332
age            0.0227     0.0134   1.695     0.01158    1.964
trtProgabide  -0.1527     0.1594  -0.958     0.17111   -0.892

Estimated Scale Parameter:  5.09
Number of Iterations:  1

Working Correlation
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.395 0.395 0.395
[2,] 0.395 1.000 0.395 0.395
[3,] 0.395 0.395 1.000 0.395
[4,] 0.395 0.395 0.395 1.000
```

Figure 14.11 R output of the `summary` method for the `epilepsy_gee3` model (slightly abbreviated).

```
R> summary(resp_lmer)

...
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.6546     0.7762   -2.13    0.033
baselinegood   3.0890     0.5986    5.16 2.5e-07
month.L       -0.2035     0.2796   -0.73    0.467
month.Q       -0.0282     0.2791   -0.10    0.919
month.C       -0.3557     0.2808   -1.27    0.205
trttrt         2.1662     0.5516    3.93 8.6e-05
gendermale     0.2384     0.6661    0.36    0.720
age           -0.0256     0.0199   -1.28    0.200
centre2        1.0385     0.5418    1.92    0.055

...
```

Figure 14.12 R output of the `summary` method for the `resp_lmer` model (abbreviated).

Bibliography

Bates, D. (2005), "Fitting linear mixed models in R," *R News*, 5, 27–30, URL `http://CRAN.R-project.org/doc/Rnews/`.

Bates, D. and Sarkar, D. (2014), ***lme4****: Linear Mixed-Effects Models Using S4 Classes*, URL `http://CRAN.R-project.org/package=lme4`, R package version 1.1-5.

Carey, V. J., Lumley, T., and Ripley, B. D. (2013), ***gee****: Generalized Estimation Equation Solver*, URL `http://CRAN.R-project.org/package=gee`, R package version 4.13-18.