Chapter 7
Modeling Longitudinal Responses using Generalized Least Squares

In this chapter we consider models for a multivariate response variable represented by serial measurements over time within subject. This setup induces correlations between measurements on the same subject that must be taken into account to have optimal model fits and honest inference. Full likelihood model-based approaches have advantages including (1) optimal handling of imbalanced data and (2) robustness to missing data (dropouts) that occur not completely at random. The three most popular model-based full likelihood approaches are mixed effects models, generalized least squares, and Bayesian hierarchical models. For continuous Y, generalized least squares has a certain elegance, and a case study will demonstrate its use after surveying competing approaches. As OLS is a special case of generalized least squares, the case study is also helpful in developing and interpreting OLS models[a].

Some good references on longitudinal data analysis include[148, 159, 252, 414, 509, 635, 637].

7.1 Notation and Data Setup

Suppose there are N independent subjects, with subject i $(i = 1, 2, \ldots, N)$ having n_i responses measured at times $t_{i1}, t_{i2}, \ldots, t_{in_i}$. The response at time t for subject i is denoted by Y_{it}. Suppose that subject i has baseline covariates X_i. Generally the response measured at time $t_{i1} = 0$ is a covariate in X_i instead of being the first measured response Y_{i0}.

For flexible analysis, longitudinal data are usually arranged in a "tall and thin" layout. This allows measurement times to be irregular. In studies com-

---

[a] A case study in OLS—Chapter 7 from the first edition—may be found on the text's web site.

paring two or more treatments, a response is often measured at baseline (pre-randomization). The analyst has the option to use this measurement as Y_{i0} or as part of X_i. There are many reasons to put initial measurements of Y in X, i.e., to use baseline measurements as baseline .

[1]

7.2 Model Specification for Effects on $E(Y)$

Longitudinal data can be used to estimate overall means or the mean at the last scheduled follow-up, making maximum use of incomplete records. But the real value of longitudinal data comes from modeling the entire time course. Estimating the time course leads to understanding slopes, shapes, overall trajectories, and periods of treatment effectiveness. With continuous Y one typically specifies the time course by a mean time-response profile. Common representations for such profiles include

- k dummy variables for $k+1$ unique times (assumes no functional form for time but assumes discrete measurement times and may spend many d.f.)
- $k = 1$ for linear time trend, $g_1(t) = t$
- k–order polynomial in t
- $k + 1$–knot restricted cubic spline (one linear term, $k - 1$ nonlinear terms)

Suppose the time trend is modeled with k parameters so that the time effect has k d.f. Let the basis functions modeling the time effect be $g_1(t)$, $g_2(t), \ldots, g_k(t)$ to allow it to be nonlinear. A model for the time profile without interactions between time and any X is given by

$$E[Y_{it}|X_i] = X_i\beta + \gamma_1 g_1(t) + \gamma_2 g_2(t) + \ldots + \gamma_k g_k(t). \qquad (7.1)$$

To allow the slope or shape of the time-response profile to depend on some of the Xs we add product terms for desired interaction effects. For example, to allow the mean time trend for subjects in group 1 (reference group) to be arbitrarily different from the time trend for subjects in group 2, have a dummy variable for group 2, a time "main effect" curve with k d.f. and all k products of these time components with the dummy variable for group 2.

Once the right hand side of the model is formulated, predicted values, contrasts, and ANOVAs are obtained just as with a univariate model. For these purposes time is no different than any other covariate except for what is described in the next section.

7.3 Modeling Within-Subject Dependence

Sometimes understanding within-subject correlation patterns is of interest in itself. More commonly, accounting for intra-subject correlation is crucial for inferences to be valid. Some methods of analysis cover up the correlation

pattern while others assume a restrictive form for the pattern. The following table is an attempt to briefly survey available longitudinal analysis methods. LOCF and the summary statistic method are not modeling methods. LOCF is an ad hoc attempt to account for longitudinal dropouts, and summary statistics can convert multivariate responses to univariate ones with few assumptions (other than minimal dropouts), with some information loss.

What Methods To Use for Repeated Measurements / Serial Data? [a][b]

	Repeated Measures ANOVA	GEE	Mixed Effects Model	GLS	LOCF	Summary Statistic[c]
Assumes normality	×		×	×		
Assumes independence of measurements within subject	×[d]	×[e]				
Assumes a correlation structure[f]	×	×[g]	×	×		
Requires same measurement times for all subjects	×				?	
Does not allow smooth modeling of time to save d.f.	×					
Does not allow adjustment for baseline covariates	×					
Does not easily extend to non-continuous Y	×			×		
Loses information by not using intermediate measurements					×[h]	×
Does not allow widely varying # of observations per subject	×	×[i]			×	×[j]
Does not allow for subjects to have distinct trajectories[k]	×	×		×	×	
Assumes subject-specific effects are Gaussian			×			
Badly biased if non-random dropouts	?	×			×	
Biased in general					×	
Harder to get tests & CLs			×[l]		×[m]	
Requires large # subjects/clusters		×				
SEs are wrong	×[n]				×	
Assumptions are not verifiable in small samples	×	N/A	×	×	×	
Does not extend to complex settings such as time-dependent covariates and dynamic[o] models	×		×	×	×	?

[a] Thanks to Charles Berry, Brian Cade, Peter Flom, Bert Gunter, and Leena Choi for valuable input.

[b] GEE: generalized estimating equations; GLS: generalized least squares; LOCF: last observation carried forward.

[c] E.g., compute within-subject slope, mean, or area under the curve over time. Assumes that the summary measure is an adequate summary of the time profile and assesses the relevant treatment effect.

2

The most prevalent full modeling approach is mixed effects models in which baseline predictors are fixed effects, and random effects are used to describe subject differences and to induce within-subject correlation. Some disadvantages of mixed effects models are

- The induced correlation structure for Y may be unrealistic if care is not taken in specifying the model.
- Random effects require complex approximations for distributions of test statistics.
- The most commonly used models assume that random effects follow a normal distribution. This assumption may not hold.

It could be argued that an extended linear model (with no random effects) is a logical extension of the univariate OLS model [b]. This model, called the generalized least squares or growth curve model[221, 509, 510], was developed long before mixed effect models became popular.

We will assume that $Y_{it}|X_i$ has a multivariate normal distribution with mean given above and with variance-covariance matrix V_i, an $n_i \times n_i$ matrix that is a function of $t_{i1}, \ldots, t_{in_i}$. We further assume that the diagonals of V_i are all equal[b]. This *extended linear model* has the following assumptions:

- all the assumptions of OLS at a single time point including correct modeling of predictor effects and univariate normality of responses conditional on X

---

[d] Unless one uses the Huynh-Feldt or Greenhouse-Geisser correction

[e] For full efficiency, if using the working independence model

[f] Or requires the user to specify one

[g] For full efficiency of regression coefficient estimates

[h] Unless the last observation is missing

[i] The cluster sandwich variance estimator used to estimate SEs in GEE does not perform well in this situation, and neither does the working independence model because it does not weight subjects properly.

[j] Unless one knows how to properly do a weighted analysis

[k] Or uses population averages

[l] Unlike GLS, does not use standard maximum likelihood methods yielding simple likelihood ratio χ^2 statistics. Requires high-dimensional integration to marginalize random effects, using complex approximations, and if using SAS, unintuitive d.f. for the various tests.

[m] Because there is no correct formula for SE of effects; ordinary SEs are not penalized for imputation and are too small

[n] If correction not applied

[o] E.g., a model with a predictor that is a lagged value of the response variable

[b] E.g., few statisticians use subject random effects for univariate Y. Pinheiro and Bates [509, Section 5.1.2] state that "in some applications, one may wish to avoid incorporating random effects in the model to account for dependence among observations, choosing to use the within-group component Λ_i to directly model variance-covariance structure of the response."

[b] This procedure can be generalized to allow for heteroscedasticity over time or with respect to X, e.g., males may be allowed to have a different variance than females.

- the distribution of two responses at two different times for the same subject, conditional on X, is bivariate normal with a specified correlation coefficient
- the joint distribution of all n_i responses for the ith subject is multivariate normal with the given correlation pattern (which implies the previous two distributional assumptions)
- responses from two different subjects are uncorrelated.

7.4 Parameter Estimation Procedure

Generalized least squares is like weighted least squares but uses a covariance matrix that is not diagonal. Each subject can have her own shape of V_i due to each subject being measured at a different set of times. This is a maximum likelihood procedure. Newton-Raphson or other trial-and-error methods are used for estimating parameters. For a small number of subjects, there are advantages in using REML (restricted maximum likelihood) instead of ordinary MLE [159, Section 5.3] [509, Chapter 5][221] (especially to get a more unbiased estimate of the covariance matrix).

When imbalances of measurement times are not severe, OLS fitted ignoring subject identifiers may be efficient for estimating β. But OLS standard errors will be too small as they don't take intra-cluster correlation into account. This may be rectified by substituting a covariance matrix estimated using the Huber-White cluster sandwich estimator or from the cluster bootstrap. When imbalances are severe and intra-subject correlations are strong, OLS (or GEE using a working independence model) is not expected to be efficient because it gives equal weight to each observation; a subject contributing two distant observations receives $\frac{1}{5}$ the weight of a subject having 10 tightly-spaced observations.

7.5 Common Correlation Structures

We usually restrict ourselves to *isotropic* correlation structures which assume the correlation between responses within subject at two times depends only on a measure of the distance between the two times, not the individual times. We simplify further and assume it depends on $|t_1 - t_2|^c$. Assume that the correlation coefficient for Y_{it_1} vs. Y_{it_2} conditional on baseline covariates X_i for subject i is $h(|t_1 - t_2|, \rho)$, where ρ is a vector (usually a scalar) set of fundamental correlation parameters. Some commonly used structures when

---

c We can speak interchangeably of correlations of residuals within subjects or correlations between responses measured at different times on the same subject, conditional on covariates X.

times are continuous and are not equally spaced [509, Section 5.3.3] are shown below, along with the correlation function names from the R `nlme` package.

Compound symmetry: $h = \rho$ if $t_1 \neq t_2$, 1 if $t_1 = t_2$ `nlme corCompSymm`
  (Essentially what two-way ANOVA assumes)
Autoregressive-moving average lag 1: $h = \rho^{|t_1 - t_2|} = \rho^s$ `corCAR1`
  where $s = |t_1 - t_2|$
Exponential: $h = \exp(-s/\rho)$ `corExp`
Gaussian: $h = \exp[-(s/\rho)^2]$ `corGaus`
Linear: $h = (1 - s/\rho)[s < \rho]$ `corLin`
Rational quadratic: $h = 1 - (s/\rho)^2/[1 + (s/\rho)^2]$ `corRatio`
Spherical: $h = [1 - 1.5(s/\rho) + 0.5(s/\rho)^3][s < \rho]$ `corSpher`
Linear exponent AR(1): $h = \rho^{d_{min} + \delta \frac{s - d_{min}}{d_{max} - d_{min}}}$, 1 if $t_1 = t_2$[572]

The structures 3–7 use ρ as a scaling parameter, not as something restricted to be in $[0, 1]$

7.6 Checking Model Fit

The constant variance assumption may be checked using typical residual plots. The univariate normality assumption (but not multivariate normality) may be checked using typical Q-Q plots on residuals. For checking the correlation pattern, a *variogram* is a very helpful device based on estimating correlations of all possible pairs of residuals at different time points[d]. Pairs of estimates obtained at the same absolute time difference s are pooled. The variogram is a plot with $y = 1 - \hat{h}(s, \rho)$ vs. s on the x-axis, and the theoretical variogram of the correlation model currently being assumed is superimposed.

7.7 Sample Size Considerations

Section 4.4 provided some guidance about sample sizes needed for OLS. A good way to think about sample size adequacy for generalized least squares is to determine the effective number of independent observations that a given configuration of repeated measurements has. For example, if the standard error of an estimate from three measurements on each of 20 subjects is the same as the standard error from 27 subjects measured once, we say that the 20×3 study has an effective sample size of 27, and we equate power from the univariate analysis on n subjects measured once to $\frac{20n}{27}$ subjects measured three times. Faes et al.[181] have a nice approach to effective sample sizes with a variety of correlation patterns in longitudinal data. For an AR(1) correlation structure with n equally spaced measurement times on each of N subjects,

---

[d] Variograms can be unstable.

with the correlation between two consecutive times being ρ, the effective sample size is $\frac{n-(n-2)\rho}{1+\rho}N$. Under compound symmetry, the effective size is $\frac{nN}{1+\rho(n-1)}$.

7.8 R Software

The nonlinear mixed effects model package `nlme` of Pinheiro & Bates in Rprovides many useful functions. For fitting linear models, fitting functions are `lme` for mixed effects models and `gls` for generalized least squares without random effects. The `rms` package has a front-end function `Gls` so that many features of `rms` can be used:

`anova`: all partial Wald tests, test of linearity, pooled tests
`summary`: effect estimates (differences in $\hat{Y}$) and confidence limits
`Predict` and `plot`: partial effect plots
`nomogram`: nomogram
`Function`: generate R function code for the fitted model
`latex`: LaTeX representation of the fitted model.

In addition, `Gls` has a cluster bootstrap option (hence you do not use `rms`'s `bootcov` for `Gls` fits). When `B` is provided to `Gls( )`, bootstrapped regression coefficients and correlation estimates are saved, the former setting up for bootstrap percentile confidence limits[e] The `nlme` package has many graphics and fit-checking functions. Several functions will be demonstrated in the case study.

7.9 Case Study

Consider the dataset in Table 6.9 of Davis [148, pp. 161–163] from a multi-center, randomized controlled trial of botulinum toxin type B (BotB) in patients with cervical dystonia from nine U.S. sites. Patients were randomized to placebo ($N = 36$), 5000 units of BotB ($N = 36$), or 10,000 units of BotB ($N = 37$). The response variable is the total score on the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS), measuring severity, pain, and disability of cervical dystonia (high scores mean more impairment). TWSTRS is measured at baseline (week 0) and weeks 2, 4, 8, 12, 16 after treatment began. The dataset name on the dataset wiki page is `cdystonia`.

---

[e] To access regular `gls` functions named `anova` (for likelihood ratio tests, AIC, etc.) or `summary` use `anova.gls` or `summary.gls`.

7.9.1 Graphical Exploration of Data

Graphics which follow display raw data as well as quartiles of TWSTRS by
time, site, and treatment. A table shows the realized measurement schedule.

```
require(rms)
```

```
getHdata(cdystonia)
attach(cdystonia)

# Construct unique subject ID
uid ← with(cdystonia, factor(paste(site, id)))

# Tabulate patterns of subjects' time points
table(tapply(week, uid,
            function(w) paste(sort(unique(w)), collapse=' ')))
```

0	0 2 4	0 2 4 12 16	0 2 4 8	0 2 4 8 12
1	1	3	1	1
0 2 4 8 12 16	0 2 4 8 16	0 2 8 12 16	0 4 8 12 16	0 4 8 16
94	1	2	4	1

```
# Plot raw data, superposing subjects
xl ← xlab('Week'); yl ← ylab('TWSTRS-total score')
ggplot(cdystonia, aes(x=week, y=twstrs, color=factor(id))) +
        geom_line() + xl + yl + facet_grid(treat ∼ site) +
        guides(color=FALSE) # Fig. 7.1
```

```
# Show quartiles
ggplot(cdystonia, aes(x=week, y=twstrs)) + xl + yl +
  ylim(0, 70) + stat_summary(fun.data="median_hilow",
                              conf.int=0.5, geom='smooth') +
  facet_wrap(∼ treat, nrow=2)   # Fig. 7.2
```

Next the data are rearranged so that Y_{i0} is a baseline covariate.

```
baseline ← subset(data.frame(cdystonia,uid), week == 0,
                   -week)
baseline ← upData(baseline, rename=c(twstrs='twstrs0'),
                  print=FALSE)
followup ← subset(data.frame(cdystonia,uid), week > 0,
                  c(uid,week,twstrs))
rm(uid)
both     ← merge(baseline, followup, by='uid')

dd       ← datadist(both)
options(datadist='dd')
```
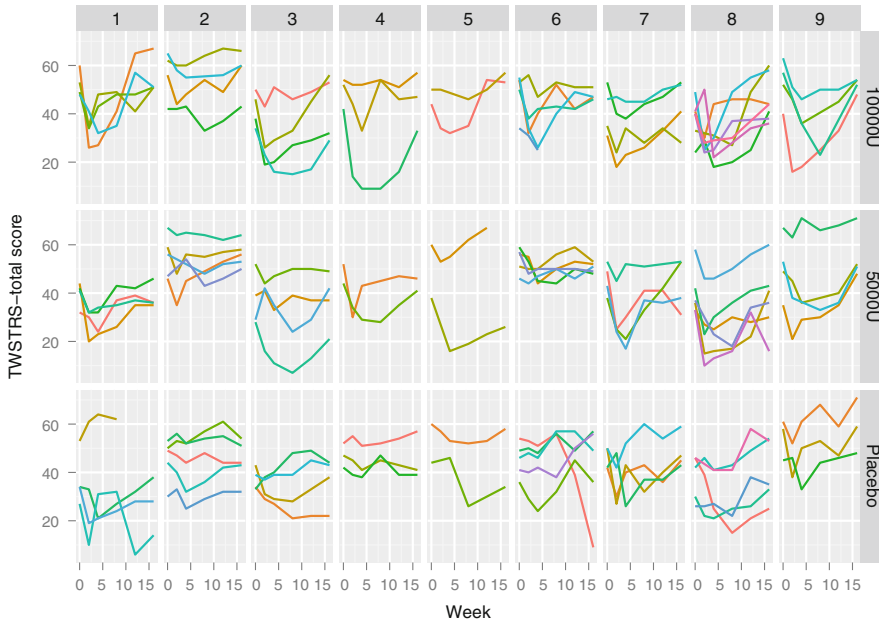
Fig. 7.1 Time profiles for individual subjects, stratified by study site and dose

7.9.2 Using Generalized Least Squares

We stay with baseline adjustment and use a variety of correlation structures, with constant variance. Time is modeled as a restricted cubic spline with 3 knots, because there are only 3 unique interior values of `week`. Below, six correlation patterns are attempted. In general it is better to use scientific knowledge to guide the choice of the correlation structure.

```
require(nlme)
```

```
cp← list(corCAR1,corExp,corCompSymm,corLin,corGaus,corSpher)
z  ← vector('list',length(cp))
for(k in 1:length(cp)) {
  z[[k]] ← gls(twstrs ∼ treat * rcs(week, 3) +
              rcs(twstrs0, 3) + rcs(age, 4) * sex, data=both,
              correlation=cp[[k]](form = ∼week | uid))
}
```

```
anova(z[[1]],z[[2]],z[[3]],z[[4]],z[[5]],z[[6]])
```

```
        Model df      AIC      BIC    logLik
z[[1]]      1 20 3553.906 3638.357 -1756.953
z[[2]]      2 20 3553.906 3638.357 -1756.953
z[[3]]      3 20 3587.974 3672.426 -1773.987
z[[4]]      4 20 3575.079 3659.531 -1767.540
z[[5]]      5 20 3621.081 3705.532 -1790.540
z[[6]]      6 20 3570.958 3655.409 -1765.479
```
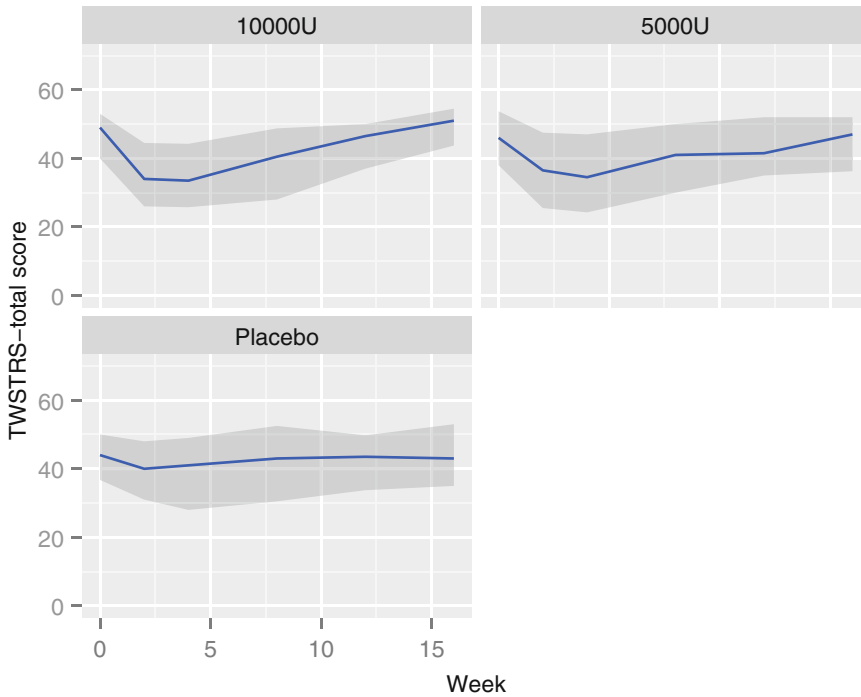
Fig. 7.2 Quartiles of `TWSTRS` stratified by dose

AIC computed above is set up so that smaller values are best. From this the continuous-time AR1 and exponential structures are tied for the best. For the remainder of the analysis we use `corCAR1`, using `Gls`.

```
a ← Gls(twstrs ∼ treat * rcs(week, 3) + rcs(twstrs0, 3) +
        rcs(age, 4) * sex, data=both,
        correlation=corCAR1(form=∼week | uid))
```

```
print(a, latex=TRUE)
```

Generalized Least Squares Fit by REML

```
Gls(model = twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) +
    rcs(age, 4) * sex, data = both, correlation = corCAR1
    (form = ~week | uid))
```

Obs	522	Log-restricted-likelihood	-1756.95
Clusters	108	Model d.f.	17
g	11.334	σ	8.5917
		d.f.	504

	Coef	S.E.	t	$\Pr(>|t|)$
Intercept	-0.3093	11.8804	-0.03	0.9792
treat=5000U	0.4344	2.5962	0.17	0.8672
treat=Placebo	7.1433	2.6133	2.73	0.0065
week	0.2879	0.2973	0.97	0.3334
week'	0.7313	0.3078	2.38	0.0179
twstrs0	0.8071	0.1449	5.57	< 0.0001
twstrs0'	0.2129	0.1795	1.19	0.2360
age	-0.1178	0.2346	-0.50	0.6158
age'	0.6968	0.6484	1.07	0.2830
age"	-3.4018	2.5599	-1.33	0.1845
sex=M	24.2802	18.6208	1.30	0.1929
treat=5000U * week	0.0745	0.4221	0.18	0.8599
treat=Placebo * week	-0.1256	0.4243	-0.30	0.7674
treat=5000U * week'	-0.4389	0.4363	-1.01	0.3149
treat=Placebo * week'	-0.6459	0.4381	-1.47	0.1411
age * sex=M	-0.5846	0.4447	-1.31	0.1892
age' * sex=M	1.4652	1.2388	1.18	0.2375
age" * sex=M	-4.0338	4.8123	-0.84	0.4023

```
Correlation Structure: Continuous AR(1)
 Formula: ~week | uid
 Parameter estimate(s):
      Phi
0.8666689
```

$\hat{\rho} = 0.867$, the estimate of the correlation between two measurements taken one week apart on the same subject. The estimated correlation for measurements 10 weeks apart is $0.867^{10} = 0.24$.

```
v ← Variogram(a, form=~ week | uid)
plot(v)   # Figure 7.3
```

The empirical variogram is largely in agreement with the pattern dictated by AR(1).

Next check constant variance and normality assumptions.

```
both$resid ← r ← resid(a); both$fitted ← fitted(a)
yl ← ylab('Residuals')
p1 ← ggplot(both, aes(x=fitted, y=resid)) + geom_point() +
     facet_grid(~ treat) + yl
p2 ← ggplot(both, aes(x=twstrs0, y=resid)) + geom_point()+yl
p3 ← ggplot(both, aes(x=week, y=resid)) + yl + ylim(-20,20) +
     stat_summary(fun.data="mean_sdl", geom='smooth')
p4 ← ggplot(both, aes(sample=resid)) + stat_qq() +
     geom_abline(intercept=mean(r), slope=sd(r)) + yl
gridExtra::grid.arrange(p1, p2, p3, p4, ncol=2)   # Figure 7.4
```
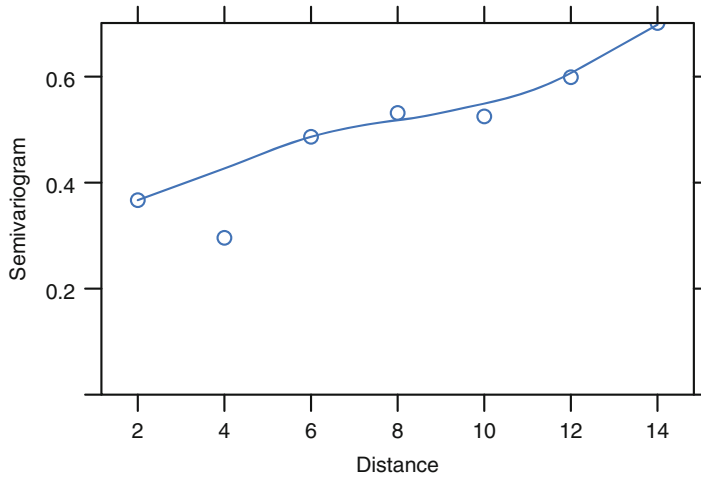
Fig. 7.3 Variogram, with assumed correlation pattern superimposed

These model assumptions appear to be well satisfied, so inferences are likely to be trustworthy if the more subtle multivariate assumptions hold.

Now get hypothesis tests, estimates, and graphically interpret the model.

```
plot(anova(a))       # Figure 7.5
```

```
ylm ← ylim(25, 60)
p1 ← ggplot(Predict(a, week, treat, conf.int=FALSE),
            adj.subtitle=FALSE, legend.position='top') + ylm
p2 ← ggplot(Predict(a, twstrs0), adj.subtitle=FALSE) + ylm
p3 ← ggplot(Predict(a, age, sex), adj.subtitle=FALSE,
            legend.position='top') + ylm
gridExtra::grid.arrange(p1, p2, p3, ncol=2)   # Figure 7.6
```

```
latex(summary(a),file='', table.env=FALSE) # Shows for week 8
```

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
week	4	12	8	6.69100	1.10570	4.5238	8.8582
twstrs0	39	53	14	13.55100	0.88618	11.8140	15.2880
age	46	65	19	2.50270	2.05140	-1.5179	6.5234
treat — 5000U:10000U	1	2		0.59167	1.99830	-3.3249	4.5083
treat — Placebo:10000U	1	3		5.49300	2.00430	1.5647	9.4212
sex — M:F	1	2		-1.08500	1.77860	-4.5711	2.4011

```
# To get results for week 8 for a different reference group
# for treatment, use e.g. summary(a, week=4, treat='Placebo')

# Compare low dose with placebo, separately at each time
```
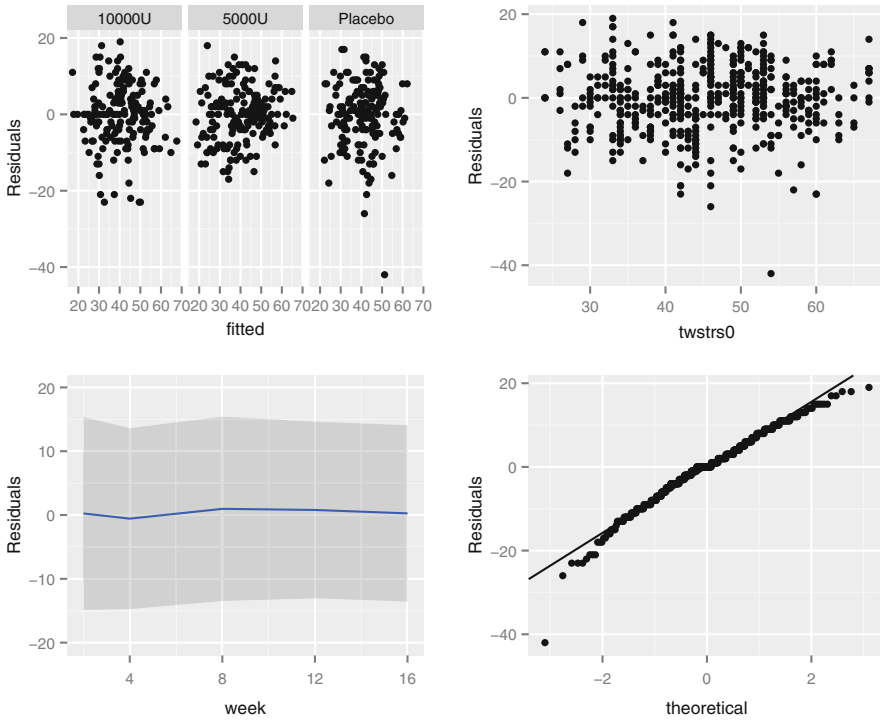
Fig. 7.4 Three residual plots to check for absence of trends in central tendency and in variability. Upper right panel shows the baseline score on the x-axis. Bottom left panel shows the mean $\pm 2 \times$SD. Bottom right panel is the QQ plot for checking normality of residuals from the GLS fit.
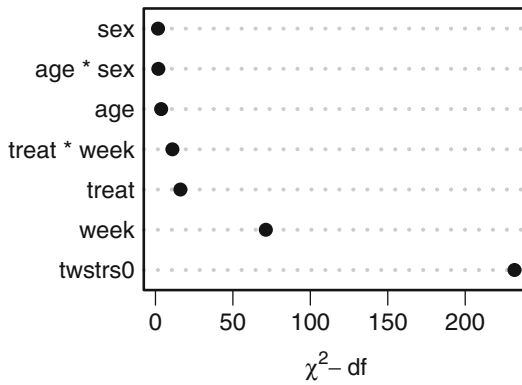

Fig. 7.5 Results of `anova` from generalized least squares fit with continuous time AR1 correlation structure. As expected, the baseline version of Y dominates.
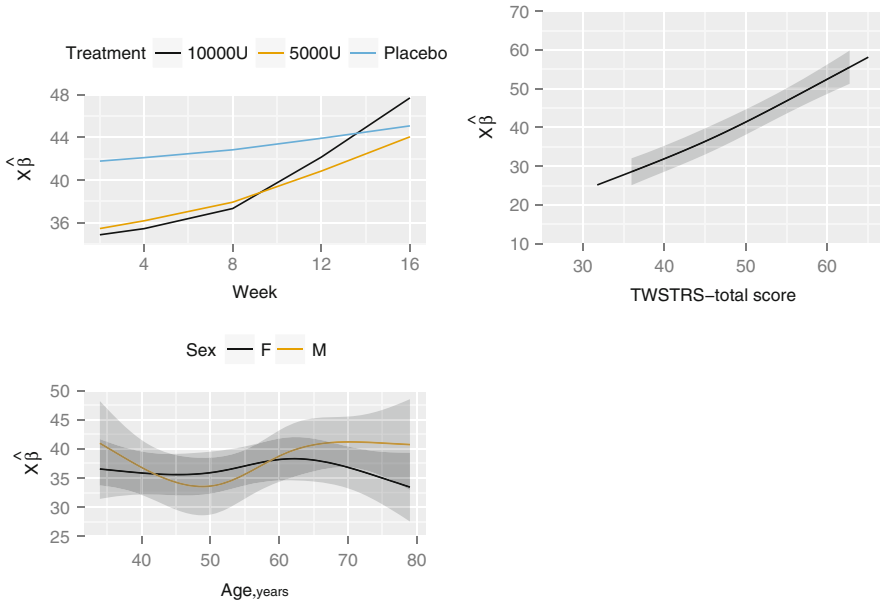
Fig. 7.6 Estimated effects of time, baseline TWSTRS, age, and sex

```
k1 ← contrast(a, list(week=c(2,4,8,12,16), treat='5000U'),
                 list(week=c(2,4,8,12,16), treat='Placebo'))
options(width=80)
print(k1, digits=3)
```

```
   week twstrs0 age sex Contrast S.E.  Lower  Upper      Z Pr(>|z|)
1     2      46  56   F    -6.31 2.10 -10.43 -2.186 -3.00   0.0027
2     4      46  56   F    -5.91 1.82  -9.47 -2.349 -3.25   0.0011
3     8      46  56   F    -4.90 2.01  -8.85 -0.953 -2.43   0.0150
4*   12      46  56   F    -3.07 1.75  -6.49  0.361 -1.75   0.0795
5*   16      46  56   F    -1.02 2.10  -5.14  3.092 -0.49   0.6260

Redundant contrasts are denoted by *

Confidence intervals are 0.95 individual intervals
```

```
# Compare high dose with placebo
k2 ← contrast(a, list(week=c(2,4,8,12,16), treat='10000U'),
                 list(week=c(2,4,8,12,16), treat='Placebo'))
print(k2, digits=3)
```

```
   week twstrs0 age sex Contrast S.E.  Lower Upper     Z Pr(>|z|)
1     2      46  56   F    -6.89 2.07 -10.96 -2.83 -3.32   0.0009
2     4      46  56   F    -6.64 1.79 -10.15 -3.13 -3.70   0.0002
3     8      46  56   F    -5.49 2.00  -9.42 -1.56 -2.74   0.0061
4*   12      46  56   F    -1.76 1.74  -5.17  1.65 -1.01   0.3109
5*   16      46  56   F     2.62 2.09  -1.47  6.71  1.25   0.2099

Redundant contrasts are denoted by *

Confidence intervals are 0.95 individual intervals
```

```
k1 ← as.data.frame(k1[c('week', 'Contrast', 'Lower',
                        'Upper')])
p1 ← ggplot(k1, aes(x=week, y=Contrast)) + geom_point() +
      geom_line() + ylab('Low Dose - Placebo') +
      geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
k2 ← as.data.frame(k2[c('week', 'Contrast', 'Lower',
                        'Upper')])
p2 ← ggplot(k2, aes(x=week, y=Contrast)) + geom_point() +
      geom_line() + ylab('High Dose - Placebo') +
      geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
gridExtra::grid.arrange(p1, p2, ncol=2)    # Figure 7.7
```
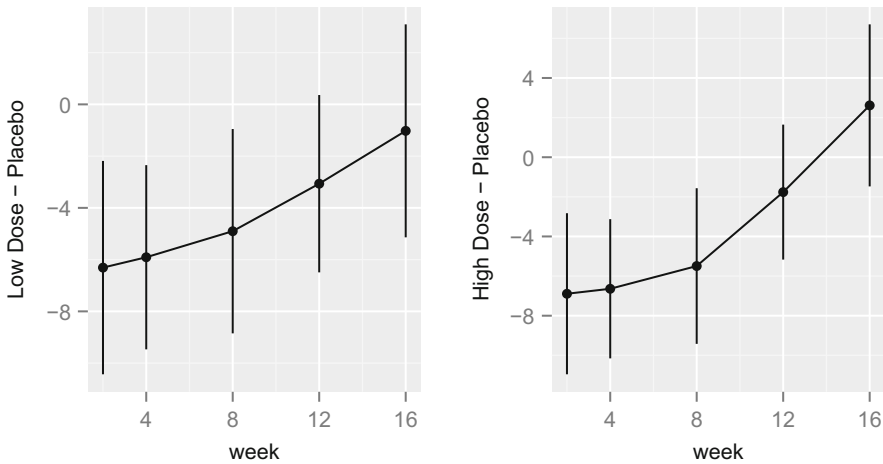

Fig. 7.7 Contrasts and 0.95 confidence limits from GLS fit

Although multiple d.f. tests such as total treatment effects or treatment $\times$ time interaction tests are comprehensive, their increased degrees of freedom can dilute power. In a treatment comparison, treatment contrasts at the last time point (single d.f. tests) are often of major interest. Such contrasts are informed by all the measurements made by all subjects (up until dropout times) when a smooth time trend is assumed. They use appropriate extrapolation past dropout times based on observed trajectories of subjects followed the entire observation period. In agreement with the top left panel of Figure 7.6, Figure 7.7 shows that the treatment, despite causing an early improvement, wears off by 16 weeks at which time no benefit is seen.

A nomogram can be used to obtain predicted values, as well as to better understand the model, just as with a univariate Y.

```
n ← nomogram(a, age=c(seq(20, 80, by=10), 85))
plot(n, cex.axis=.55, cex.var=.8, lmgp=.25)  # Figure 7.8
```
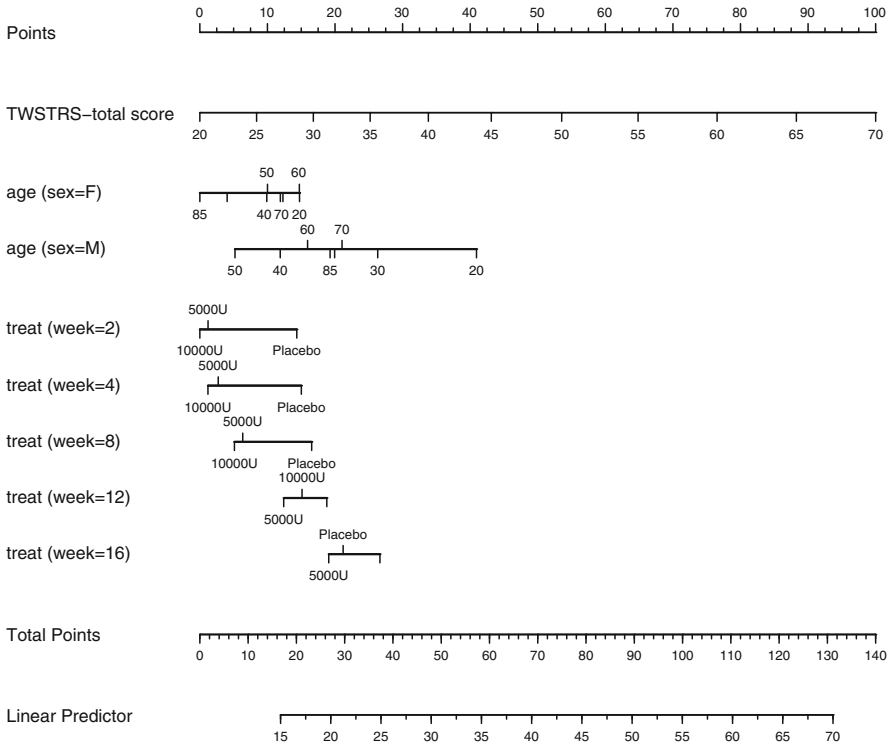
Fig. 7.8 Nomogram from GLS fit. Second axis is the baseline score.

7.10 Further Reading

[1]   Jim Rochon (Rho, Inc., Chapel Hill NC) has the following comments about using the baseline measurement of Y as the first longitudinal response.

> For RCTs [randomized clinical trials], I draw a sharp line at the point when the intervention begins. The LHS [left hand side of the model equation] is reserved for something that is a response to treatment. Anything before this point can potentially be included as a covariate in the regression model. This includes the "baseline" value of the outcome variable. Indeed, the best predictor of the outcome at the end of the study is typically where the patient began at the beginning. It drinks up a lot of variability in the outcome; and, the effect of other covariates is typically mediated through this variable.

> I treat anything after the intervention begins as an outcome. In the western scientific method, an "effect" must follow the "cause" even if by a split second.

> Note that an RCT is different than a cohort study. In a cohort study, "Time 0" is not terribly meaningful. If we want to model, say, the trend over time, it would be legitimate, in my view, to include the "baseline" value on the LHS of that regression model.

Now, even if the intervention, e.g., surgery, has an immediate effect, I would include still reserve the LHS for anything that might legitimately be considered as the response to the intervention. So, if we cleared a blocked artery and then measured the MABP, then that would still be included on the LHS.

Now, it could well be that most of the therapeutic effect occurred by the time that the first repeated measure was taken, and then levels off. Then, a plot of the means would essentially be two parallel lines and the treatment effect is the distance between the lines, i.e., the difference in the intercepts.

If the linear trend from baseline to Time 1 continues beyond Time 1, then the lines will have a common intercept but the slopes will diverge. Then, the treatment effect will the difference in slopes.

One point to remember is that the estimated intercept is the value at time 0 that we predict from the set of repeated measures post randomization. In the first case above, the model will predict different intercepts even though randomization would suggest that they would start from the same place. This is because we were asleep at the switch and didn't record the "action" from baseline to time 1. In the second case, the model will predict the same intercept values because the linear trend from baseline to time 1 was continued thereafter.

More importantly, there are considerable benefits to including it as a covariate on the RHS. The baseline value tends to be the best predictor of the outcome post-randomization, and this maneuver increases the precision of the estimated treatment effect. Additionally, any other prognostic factors correlated with the outcome variable will also be correlated with the baseline value of that outcome, and this has two important consequences. First, this greatly reduces the need to enter a large number of prognostic factors as covariates in the linear models. Their effect is already mediated through the baseline value of the outcome variable. Secondly, any imbalances across the treatment arms in important prognostic factors will induce an imbalance across the treatment arms in the baseline value of the outcome. Including the baseline value thereby reduces the need to enter these variables as covariates in the linear models.

Stephen Senn[563] states that temporally and logically, a "baseline cannot be a *response* to treatment", so baseline and response cannot be modeled in an integrated framework.

> ...one should focus clearly on 'outcomes' as being the only values that can be influenced by treatment and examine critically any schemes that assume that these are linked in some rigid and deterministic view to 'baseline' values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way.

The final reason that baseline cannot be modeled as the response at time zero is that many studies have inclusion/exclusion criteria that include cutoffs on the baseline variable yielding a truncated distribution. In general it is not appropriate to model the baseline with the same distributional shape as the follow-up

measurements. Thus the approach recommended by Liang and Zeger[405] and Liu et al.[423] are problematic[f].

2 Gardiner et al.[211] compared several longitudinal data models, especially with regard to assumptions and how regression coefficients are estimated. Peters et al.[500] have an empirical study confirming that the "use all available data" approach of likelihood–based longitudinal models makes imputation of follow-up measurements unnecessary.

3 Keselman et al.[347] did a simulation study to study the reliability of AIC for selecting the correct covariance structure in repeated measurement models. In choosing from among 11 structures, AIC selected the correct structure 47% of the time. Gurka et al.[247] demonstrated that fixed effects in a mixed effects model can be biased, independent of sample size, when the specified covariate matrix is more restricted than the true one.

---

[f] In addition to this, one of the paper's conclusions that analysis of covariance is not appropriate if the population means of the baseline variable are not identical in the treatment groups is arguable[563]. See[346] for a discussion of[423].