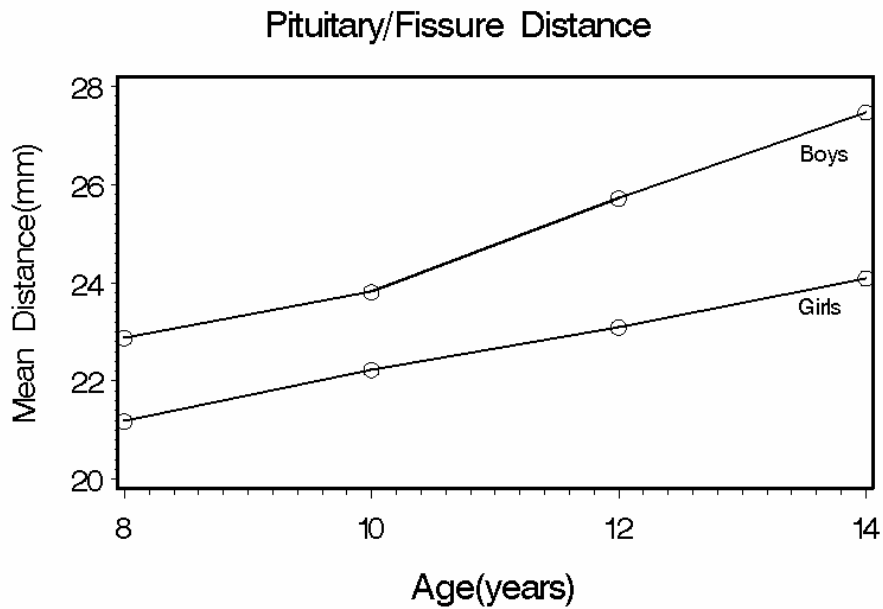
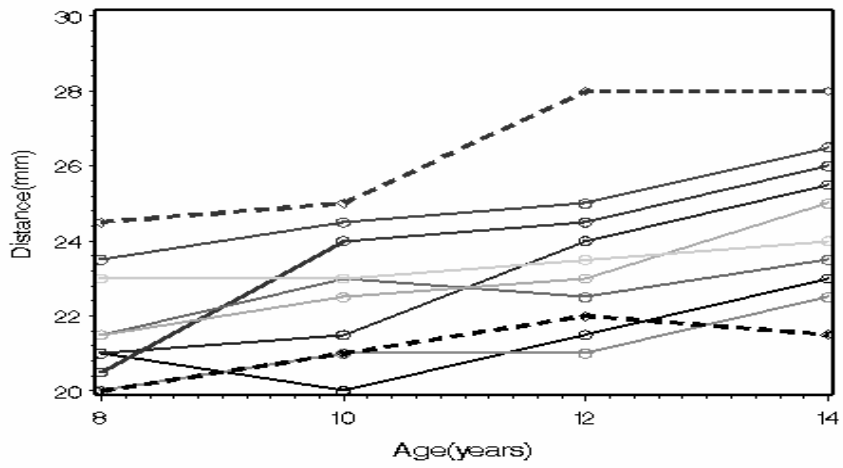


1. A. Explore the data with plots
 - i. Create a profile plot of the data. Use different types of lines for boys and girls. Describe any patterns that you see in the plot. What does this plot reveal about differences between boys and girls?

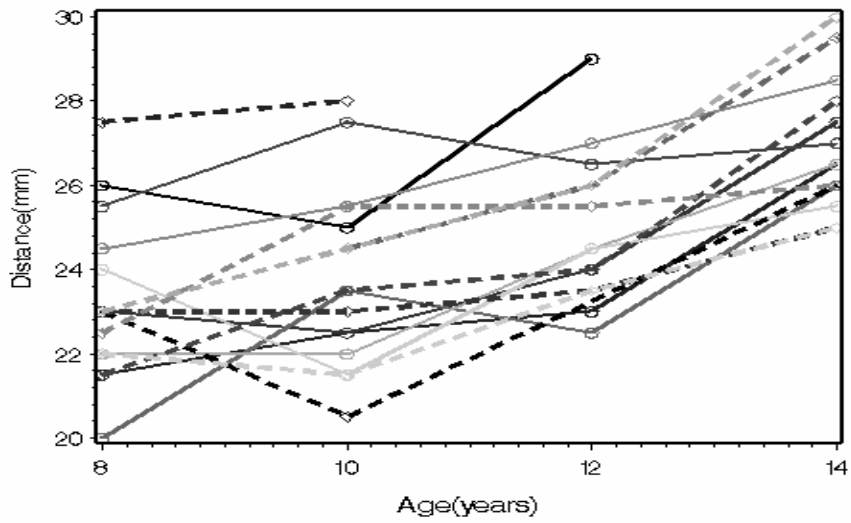


Mean distances are larger for boys at every age. Between the ages of 8 and 14, the mean distance seems to increase in a straight line fashion for both boys and girls. The following plots of profiles for individual children suggest that there is more variability among boys than among girls. Variability appears to be about the same at each age.

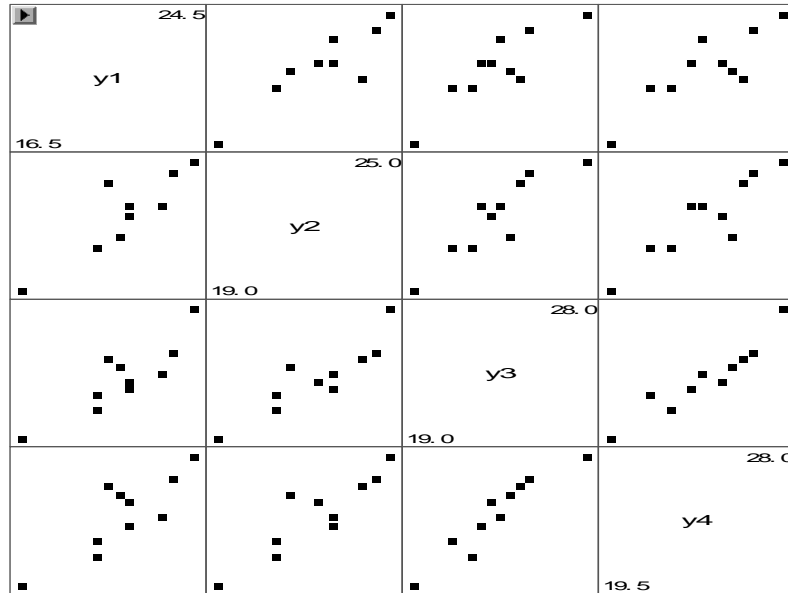
Pituitary/Fissure Distances (Girls)



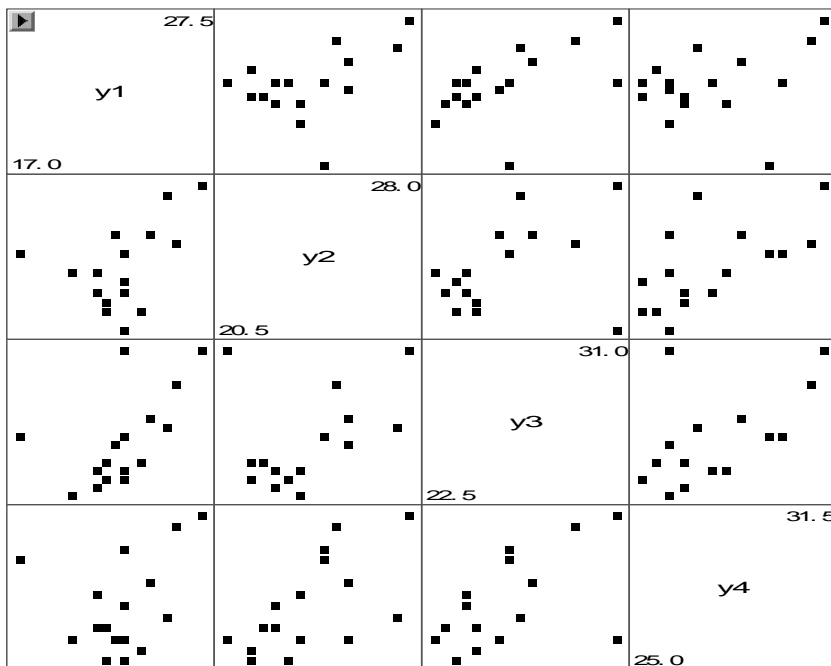
Pituitary/Fissure Distances (Boys)



ii. The following scatterplot matrix for girls shows positive correlations for all ages. It is not clear if the correlations become weaker as the distance between the ages increases. Correlations may be stronger among measurements at older ages.



The following scatterplot matrix for boys shows weaker positive correlations than for girls. It is not clear if the correlations change as the distances between the ages increase. The correlation seems to be strongest between the 12 and 14 year measurements



The correlation matrices for boys and girls are shown below. They corroborate what was seen in the scatterplots.

| Girls | | | | Boys | | | |
|---------|----------|----------|----------|---------|----------|----------|----------|
| 8 weeks | 10 weeks | 12 weeks | 14 weeks | 8 weeks | 10 weeks | 12 weeks | 14 weeks |
| 1 | 0.830 | 0.862 | 0.841 | 1 | 0.437 | 0.558 | 0.316 |
| 0.830 | 1 | 0.895 | 0.879 | 0.437 | 1 | 0.387 | 0.631 |
| 0.862 | 0.895 | 1 | 0.948 | 0.558 | 0.387 | 1 | 0.586 |
| 0.841 | 0.879 | 0.948 | 1 | 0.316 | 0.631 | 0.586 | 1 |

B. The following model was fit to the data,

$$Y_{ijk} = \beta_{0j} + \beta_{1j}X_{ijk} + \beta_{2j}X_{ijk}^2 + \beta_{3j}X_{ijk}^3 + \varepsilon_{ijk},$$

where Y_{ijk} is the i -th measurement on the k -th child in the j -th gender group, and

X_{ijk} is the age of the child when the measurement was taken. An arbitrary

(unstructured) covariance matrix was assumed for $(\varepsilon_{1jk} \ \varepsilon_{2jk} \ \varepsilon_{3jk} \ \varepsilon_{4jk})$, the random errors for the four measurements taken on a single subject. The estimated coefficients are as follows:

Solution for Fixed Effects

| Effect | sex | Estimate | Standard Error | DF | t Value | Pr > t |
|-----------------|-----|----------|----------------|----|---------|---------|
| sex | 0 | 8.8182 | 55.8701 | 25 | 0.16 | 0.8759 |
| sex | 1 | 51.3125 | 46.3251 | 25 | 1.11 | 0.2786 |
| age*sex | 0 | 2.8939 | 15.8708 | 25 | 0.18 | 0.8568 |
| age*sex | 1 | -8.6484 | 13.1594 | 25 | -0.66 | 0.5171 |
| age*age*sex | 0 | -0.2216 | 1.4724 | 25 | -0.15 | 0.8816 |
| age*age*sex | 1 | 0.8242 | 1.2208 | 25 | 0.68 | 0.5058 |
| age*age*age*sex | 0 | 0.006629 | 0.04457 | 25 | 0.15 | 0.8830 |
| age*age*age*sex | 1 | -0.02344 | 0.03695 | 25 | -0.63 | 0.5317 |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|-----------------|--------|--------|---------|--------|
| sex | 2 | 25 | 0.63 | 0.5429 |
| age*sex | 2 | 25 | 0.23 | 0.7942 |
| age*age*sex | 2 | 25 | 0.24 | 0.7890 |
| age*age*age*sex | 2 | 25 | 0.21 | 0.8102 |

After adjusting for other effects in the model, none of the effects appear to be significant. The systematic part of this model, however, is as large as possible because it allows for a different mean for each different combination of sex and age. Backward elimination can be used to find a more parsimonious model. For eliminate the interaction between sex and the cubic trend in age and fit a common slope on the cubic trend. Next delete the common cubic term because it is not significant. Continue this process with the quadratic term. The most parsimonious model that fits the trends across age in the mean responses has the following estimated coefficients:

| Effect | sex | Estimate | Standard Error | DF | t Value | Pr > t |
|---------|-----|----------|----------------|----|---------|---------|
| sex | 0 | 17.4254 | 1.2612 | 25 | 13.82 | <.0001 |
| sex | 1 | 15.8423 | 1.0457 | 25 | 15.15 | <.0001 |
| age*sex | 0 | 0.4764 | 0.1066 | 25 | 4.47 | 0.0001 |
| age*sex | 1 | 0.8268 | 0.08843 | 25 | 9.35 | <.0001 |

These estimates indicate that the straight line relationship in the mean response for boys has a steeper positive slope than the corresponding line for girls. The estimated covariance matrix for the random coefficients is:

| Row | Col1 | Col2 | Col3 | Col4 |
|-----|--------|--------|--------|--------|
| 1 | 5.4252 | 2.7092 | 3.8411 | 2.7151 |
| 2 | 2.7092 | 4.1906 | 2.9745 | 3.3137 |
| 3 | 3.8411 | 2.9745 | 6.2632 | 4.1332 |
| 4 | 2.7151 | 3.3137 | 4.1332 | 4.9862 |

The estimated correlation matrix for the estimated coefficients is

| Row | Col1 | Col2 | Col3 | Col4 |
|-----|--------|--------|--------|--------|
| 1 | 1.0000 | 0.5682 | 0.6589 | 0.5220 |
| 2 | 0.5682 | 1.0000 | 0.5806 | 0.7249 |
| 3 | 0.6589 | 0.5806 | 1.0000 | 0.7396 |
| 4 | 0.5220 | 0.7249 | 0.7396 | 1.0000 |

- C. To search for an appropriate covariance model, fit the most complex regression model in part B,

$$Y_{ijk} = \beta_{0j} + \beta_{1j}X_{ijk} + \beta_{2j}X_{ijk}^2 + \beta_{3j}X_{ijk}^3 + \varepsilon_{ijk}$$

was used. This model fits a different mean response for each of the eight combinations of gender and age. You could also search for a good covariance structure using the more parsimonious model with different intercepts and different straight line trends in age for boys and girls. Values of REML log-likelihood and corresponding AIC and BIC values are shown in the following table for various covariance models for $(\varepsilon_{1jk} \ \varepsilon_{2jk} \ \varepsilon_{3jk} \ \varepsilon_{4jk})$. The values for the more parsimonious model appear beneath the

values for the most general fixed effects model. (Do not compare AIC, BIC or REML likelihood values for different fixed effects models.) There is no indication of heterogeneous variances across age levels. The smallest AIC value occurs for the compound symmetry model and the smallest BIC value occurs for the slightly more complex Toeplitz model. Both models lead to similar estimates of the regression coefficients and similar standard errors for the estimated coefficients. The unstructured model is also a reasonable choice in this case, and it leads to the same inferences about the regression parameters.

| | -2(REML Log-likelihood) | AIC | BIC |
|--|-----------------------------|--------------|--------------|
| (1) Independence with homogeneous variances | 497.1 | 499.1 | 501.7 |
| | 483.6 | 485.6 | 486.9 |
| (2) Compound symmetry | 450.0 | 454.0 | 456.6 |
| | 433.8 | 437.8 | 440.3 |
| (3) Antedependence | 457.6 | 471.1 | 480.7 |
| | 441.1 | 455.1 | 464.2 |
| (4) Autoregressive(1) | 461.1 | 465.1 | 467.7 |
| | 444.6 | 448.6 | 451.2 |
| (5) Toeplitz model with homogeneous variances | 445.5 | 453.5 | 458.7 |
| | 429.4 | 437.4 | 442.6 |
| (6) Independence with heterogeneous variances | 497.1 | 499.1 | 500.4 |
| | 483.6 | 485.6 | 486.9 |
| (7) Equal correlations with heter. Variances | 448.0 | 458.0 | 464.5 |
| | 432.0 | 442.0 | 448.5 |
| (8) AR(1) with heterogeneous variances | 459.1 | 469.1 | 475.6 |
| | 442.8 | 452.8 | 459.3 |
| (9) Toeplitz with heterogeneous variances | 443.3 | 457.3 | 466.3 |
| | 427.4 | 441.4 | 450.5 |
| (10) Unstructured covariance matrix | 440.6 | 460.6 | 473.6 |
| | 424.5 | 444.5 | 457.5 |

The previous analysis assumed that the covariance matrices for the repeated measurements were the same for boys and girls. The plots suggest that variance may be larger and the correlations may be smaller for boys. You could fit models with different covariance matrices for boys and girls. You could first fit separate models for boys and girls and search for the best covariance structure for boys and the best covariance structure for repeated measures on girls. This can be done by running the MIXED procedure twice, once for each gender, or by using the Group= option to the REPEATED statement in the MIXED procedure. AIC, BIC and -2log(REML likelihood) values are shown below for the cubic regression model with the same form of covariance structure but different parameter estimates for boys and girls. Separate compound symmetry models for boys and girls seem to be adequate.

| | -2(REML Log-likelihood) | AIC | BIC |
|--|-----------------------------|--------------|--------------|
| (1) Independence with homogeneous variances | 496.9 | 500.9 | 503.5 |
| (2) Compound symmetry | 432.9 | 440.9 | 446.1 |
| (3) Antedependence | 432.2 | 460.2 | 478.4 |
| (4) Autoregressive(1) | 439.1 | 447.1 | 452.2 |
| (5) Toeplitz model with homogeneous variances | 427.8 | 443.8 | 454.2 |
| (6) Independence with heterogeneous variances | 495.0 | 511.0 | 521.4 |
| (7) Equal correlations with heter. variances | 428.7 | 448.7 | 461.6 |
| (8) AR(1) with heterogeneous variances | 435.7 | 455.7 | 468.6 |
| (9) Toeplitz with heterogeneous variances | 423.9 | 451.9 | 470.1 |
| (10) Unstructured covariance matrix | 419.4 | 459.4 | 485.3 |

Use the method=ML option in the model statement for the MIXED procedure in SAS to obtain maximum likelihood estimates for the combined set of parameters in the systematic and covariance parts of the model. Using the unstructured covariance matrix for the repeated measures, maximum likelihood estimates of the regression line for boys is

$$\hat{Y} = 15.8282 + 0.8340(\text{age})$$

(1.1258) (0.1044)

Standard errors are displayed beneath the estimated coefficients. The maximum likelihood estimate of the regression line for girls is

$$\hat{Y} = 17.4220 + 0.4824(\text{age})$$

(0.8099) (0.0718)

The values of $-2\log(\text{likelihood})$ for boys and girls are 264.4 and 130.6, respectively. The value of $-2\log(\text{likelihood})$ for fitting the separate regression lines to boys and girls with a common unstructured covariance matrix for the repeated measure is 419.5. A likelihood ratio test of the null hypothesis of common covariance matrices for boys and girls is

$X^2 = (419.5) - (264.4 + 130.6)$. When the null hypothesis is true, this test approximately has a central chi-squared distribution with 10 df. Using this distribution, the p-value is 0.0064, which supports the indication in the plots that the covariance matrices are not the same for boys and girls. The boys exhibit larger variation and weaker correlations. However, there is very little difference in the estimates of the regression parameters if separate covariance matrices (using either the unstructured or compound symmetry covariance structure) are estimated or a common covariance matrix is estimated for boys and girls. Confidence intervals and standards errors are somewhat affected, but inferences do not change. (The compound symmetry model appeared to be adequate for both boys and girls. Note that it is only necessary to get the covariance structure approximately correct to make proper inferences about the regression coefficients. In this case there are only 4 repeated measures on each subject, so simply using the unstructured covariance matrix does not add many more covariance parameters than the compound symmetry covariance structure, and about the same results for the regression coefficients are

achieved with either covariance structure. In situations where there are many repeated measures and not very many subjects, it will be beneficial to seek a simple covariance structure because there will not be sufficient information in the data to estimate a large number of covariance parameters with reasonable accuracy. Large variances for estimated covariance parameters will unduly inflate the variances of the generalized least squares estimates of the regression coefficients.)

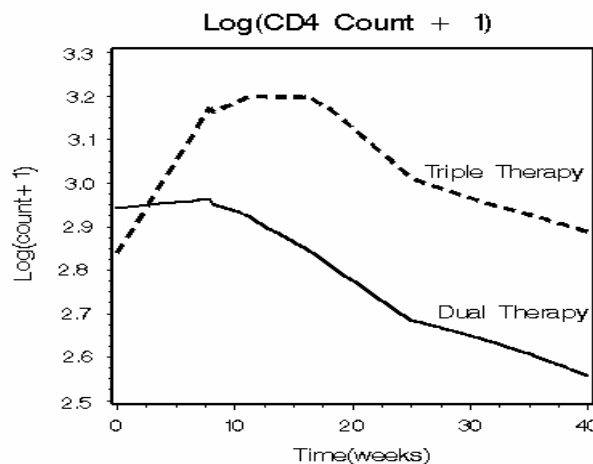
- D. Considering everything done above, a good model appears to be separate straight lines for boy and girls to model age trends in the mean response. I would use different unstructured covariance matrices for the repeated measures for boys and girls, but separate compound symmetry or Toeplitz models would be a little more parsimonious and give essentially the same results.

2. The data posted in the file cd4data.txt are from a randomized, double blind, study of AIDS patients with advanced immune suppression (CD4 counts less than 50 cells/mm³) (Henry, et al 1998). The patients in this study were randomized to either dual or triple combinations of HIV-1 reverse transcriptase inhibitors. The four treatment groups correspond to daily regimens containing 600 mg of zidovudine: zidovudine alternating monthly with 400 mg didanosine, zidovudine plus 2.25 mg of zalcitabine, zidovudine plus 400 mg of didanosine, or zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine (triple therapy). There are 1313 lines on the data file. On each line, the variables appear in the following order:

- id patient identification number
- group treatment group: coded 1 for triple therapy and 0 for dual therapy
- age baseline patient age in years
- sex coded 0 for female and 1 for male.\
- week number of weeks since baseline
- logcd4 log(cd4 count + 1)

The first three treatment groups into a single group (dual therapy) and comparisons are made with the other treatment group (triple therapy).

A. The loess smoothing technique was used to plot a smooth curve for log(CD4 count + 1) against time since baseline for the dual and triple therapy groups.



This plot reveals an increasing trend in $\log(\text{CD4 counts} + 1)$ for patients given triple therapy that peaks between 10 and 18 weeks and then declines. The mean response for dual therapy appears to decline after about 9 weeks.

- B. A piecewise linear model with a knot at 16 weeks was fit to the data. Letting t_{ij} denote the time since baseline for the j -th measurement on the i -th patient (with $t_{ij} = 0$ at baseline), and including random coefficients for each patient, the model can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)_+ + \beta_4 z_i t_{ij} + \beta_5 z_i (t_{ij} - 16)_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij} - 16)_+ + \varepsilon_{ij}$$

where

Y_{ij} is the j -th measurement of $\log(\text{CD4 count} + 1)$ on the i -th patient

$$(t_{ij} - 16)_+ = \begin{cases} t_{ij} - 16 & \text{if } t_{ij} \geq 16 \\ 0 & \text{if } t_{ij} < 16 \end{cases}$$

$$z_i = \begin{cases} 0 & \text{for dual therapy} \\ 1 & \text{for triple therapy} \end{cases}$$

$$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim \text{NID} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix} \right) \quad \text{and} \quad \varepsilon_{ij} \sim \text{NID}(0, \sigma^2) \quad \text{and}$$

any ε_{ij} is independent of any \mathbf{b}_i .

- i. The REML estimates of g_{11} g_{12} g_{13} g_{22} g_{23} g_{33} and σ^2 and the standard errors of the estimates are

| Covariance Parameter Estimates | | | | |
|--------------------------------|----------|----------------|---------|---------|
| Cov Parm | Estimate | Standard Error | Z Value | P-value |
| g_{11} | 0.5857 | 0.03475 | 16.85 | <.0001 |
| g_{12} | 0.007254 | 0.001805 | 4.02 | <.0001 |
| g_{22} | 0.000923 | 0.000160 | 5.76 | <.0001 |
| g_{13} | -0.01235 | 0.002730 | -4.52 | <.0001 |
| g_{23} | -0.00092 | 0.000236 | -3.90 | <.0001 |
| g_{33} | 0.001240 | 0.000395 | 3.14 | 0.0008 |
| Residual | 0.3062 | 0.01007 | 30.39 | <.0001 |

The variances of all random coefficients are significantly larger than zero, so it appears that all three are needed in the model. It appears that an unstructured covariance matrix is needed for the random coefficients. The variation in the intercepts is much larger than the variation in the regression coefficients for the time effects which indicates that the baseline CD4 counts are much more variable across subjects than the rates of change in the CD4 counts.

- ii. Generalized least squares estimates of $\beta_0, \beta_1, \beta_2, \beta_4, \beta_5$ and their standard errors are shown below with corresponding t-values and p-values.

| Effect | Estimate | Standard Error | DF | t Value | Pr > t |
|-------------|----------|----------------|------|---------|---------|
| Intercept | 2.9415 | 0.02562 | 1308 | 114.81 | <.0001 |
| week | -0.00734 | 0.001987 | 1185 | -3.70 | 0.0002 |
| week_16 | -0.01204 | 0.003174 | 1004 | -3.79 | 0.0002 |
| week*trt | 0.02685 | 0.003847 | 1534 | 6.98 | <.0001 |
| week_16*trt | -0.02774 | 0.006198 | 1534 | -4.47 | <.0001 |

All of the regression coefficients are significantly different from zero.

- iii. There is a significant difference in the trend of $\log(\text{count}+1)$ against time before week 16 ($t=6.98$ with $p\text{-value}<.0001$). Triple therapy subjects initially have an increasing trend in $\log(\text{count} + 1)$ of about 0.02 per week over the first 16 weeks, but the dual therapy patients tend to exhibit a small decline CD4 counts during the first 16 weeks. There is also a significant difference in the increment to the trend of $\log(\text{count}+1)$ against time after 16 weeks, so that the trends are decreasing and nearly parallel for the two therapies after 16 weeks.

(Note that $\hat{\beta}_4 + \hat{\beta}_5 = 0.02685 - 0.02774 = -0.00089$ with standard error

$$\sqrt{\text{var}(\hat{\beta}_4) + \text{var}(\hat{\beta}_5) + 2 \text{cov}(\hat{\beta}_4, \hat{\beta}_5)} =$$

$\sqrt{.000015 + .000038 + (2)(-.000020)} = .003606$) Between 16 and 40 weeks after treatment, the rate of decrease in $\log(\text{CD4 count} + 1)$ is about the same for dual and triple therapy. The main difference between the results for dual and triple therapy is in the response during the first 16 weeks of treatment when $\log(\text{CD4 count}+1)$ tends to increase and reach a higher peak for triple therapy.

- iv. Interpretations of fixed and random parameters
- $\beta_0 + b_{0i}$ is the $\log(\text{CD4 count} + 1)$ at baseline for the i -th subject
 - $\beta_1 + b_{1i}$ is the average weekly change in $\log(\text{CD4 count} + 1)$ during the first 16 weeks after the start of treatment experienced by the i -th subject who received dual therapy
 - $\beta_1 + \beta_4 + b_{1i}$ is the average weekly change in $\log(\text{CD4 count} + 1)$ during the first 16 weeks after the start of treatment experienced by the i -th subject who received triple therapy

$\beta_1 + \beta_2 + b_{1i} + b_{2i}$ is the average weekly change in $\log(\text{CD4 count} + 1)$ between 16 and 40 weeks after the start of treatment experienced by the i -th subject who received dual therapy

$\beta_1 + \beta_2 + \beta_4 + \beta_5 + b_{1i} + b_{2i}$ is the average weekly change in $\log(\text{CD4 count} + 1)$ between 16 and 40 weeks after the start of treatment experienced by the i -th subject who received triple therapy

iv. It is reasonable to use the same intercept for each treatment group because the subjects are randomly assigned to the treatments. Consequently, the expected value of $\log(\text{CD4 count} + 1)$ at baseline (before the treatment has any effect) would be the same for each treatment and any observed difference would be the result of random assignment of subjects to treatments.

C. The following model that accounts for associations with gender and age:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)_+ + \beta_4 z_i t_{ij} + \beta_5 z_i (t_{ij} - 16)_+ + \beta_6 \text{Age}_i + \beta_7 \text{Gender}_i + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij} - 16)_+ + \varepsilon_{ij}$$

i. With respect to this model, the average values of $\log(\text{CD4 count} + 1)$ are not significantly different at baseline (p -value=0.212). There is a significant affect of age (p -value=.0009) on the mean baseline value for $\log(\text{CD4 count} + 1)$. The estimated coefficient for age indicates that the mean baseline value for $\log(\text{CD4 count} + 1)$ increases by 0.01 for each year increase in age.

ii. Generalized least squares estimates of the regression coefficients are shown below. After adjusting for the effects of sex and age on baseline values of $\log(\text{CD4 count} + 1)$, inferences about the trends across time in the $\log(\text{CD4 count} + 1)$ do not change in any substantial way. Except for the intercept which was reduced from 2.9415 to 2.6455 by adjusting for possible age and sex effects on baseline values, the values of the other estimated regression coefficients are essentially unaffected by the addition of the age and sex effects to the model in part C?

| Effect | Estimate | Standard Error | DF | t Value | Pr > t |
|-------------|----------|----------------|------|---------|---------|
| Intercept | 2.6455 | 0.1280 | 1306 | 20.67 | <.0001 |
| age | 0.009996 | 0.003016 | 1534 | 3.31 | 0.0009 |
| sex | -0.09269 | 0.07537 | 1534 | -1.23 | 0.2190 |
| week | -0.00730 | 0.001986 | 1185 | -3.67 | 0.0003 |
| week*trt | 0.02679 | 0.003843 | 1534 | 6.97 | <.0001 |
| week_16 | -0.01206 | 0.003174 | 1004 | -3.80 | 0.0002 |
| trt*week_16 | -0.02772 | 0.006197 | 1534 | -4.47 | <.0001 |

- D. In addressing this model fitting exercise, you could consider the following issues:
- a. Does subject age affect the trend in $\log(\text{CD4 count} + 1)$ over time and is the effect the same for both therapies?
 - b. Does subject gender affect the trend in $\log(\text{CD4 count} + 1)$ over time and is the effect the same for both therapies?
 - c. Was week 16 a good place to place a knot for this linear spline? If you use a linear spline should the knot be placed at different time points for the two therapies?
 - d. Would a polynomial model in time or some other curve (possibly a non-linear curve) fit the data as well as a linear spline with one knot? Such curves would have to allow for a difference in trends for the two therapies for the first 16 weeks of treatment and then provide nearly parallel curves between 16 and 40 weeks. If some other model is selected for changes in $\log(\text{CD4 count} + 1)$ across time, how should random effects for individual subjects be incorporated into the new model?
 - e. Consider an analysis in which the three dual therapy groups are not combined.