Prof. Brent Coull

BIO 226: Applied Longitudinal Analysis

Homework 4 Solutions

Due Thursday, March 26, 2015

[100 points]

Purpose:
To provide an introduction to the joint parametric modeling of the mean response and covariance for longitudinal data.

Instructions:
1. For each question requiring data analysis, support your conclusions by including the relevant SAS output in your answer.
2. Include your SAS program (but not your SAS output) as an appendix to your solutions. In general, this will only be reviewed during grading to help identify a major problem affecting your answers to questions so please do not cross-reference the appendix in your answers to questions.
Late homework will not be graded unless you make arrangements with the Instructor prior to the due date/time.

Joint Modeling of Mean and Covariance in the a Study on the Effect of Ozone on Pulmonary Function:
In a study designed to examine the acute responses of normal subjects to ozone exposure, Follinsbee and colleagues randomized 20 subjects to exposure to either room air or 0.12 ppb ozone. The exposure lasted six hours. A baseline and 6 other measures of FEV1 (a measure of pulmonary function) made at hourly intervals were recorded for all study participants in the study. Subjects cycled on an exercise bike before a measure of FEV1 was obtained. The investigators were interested in determining whether changes in pulmonary function during the 6 hours of exposure were different in the ozone and room air exposed groups.
In the analyses of the data from this study, the response variable of interest was FEV1 (ml)/100. From here on out when we refer to FEV1 we mean this rescaled response. The measurement times were coded 0-6, with time=0 for the baseline measurement, time=1 for the measurement at hour 1,,time=6 for the measurement at hour 6. For this homework, we will consider the data from hours 0, 2, 4, 6. The two exposure groups were coded 0 and 1, with group=1 denoting exposure to ozone and group=0 denoting exposure to room air.
The data are in the file "ozone0246.txt" on the course web page. Each row of the data set contains the following four variables: subject ID, hour, group and the y = FEV1/100 measurements, respectively.

Problem 1

[15 points: 5 for each part] Descriptive Analyses:

(a) Describe key aspects of the longitudinal design and completeness of data.

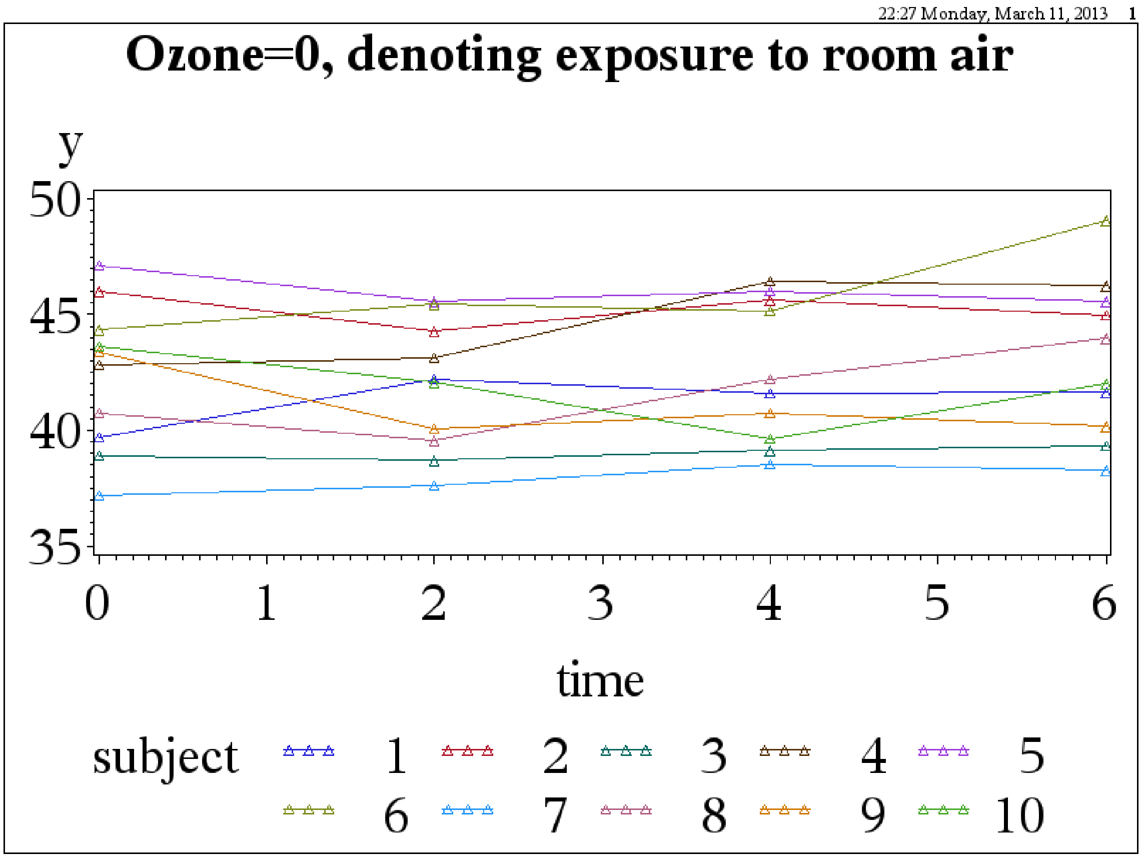
```
                                  N
        time          group     Obs      N           Mean
        --------------------------------------------------------
```

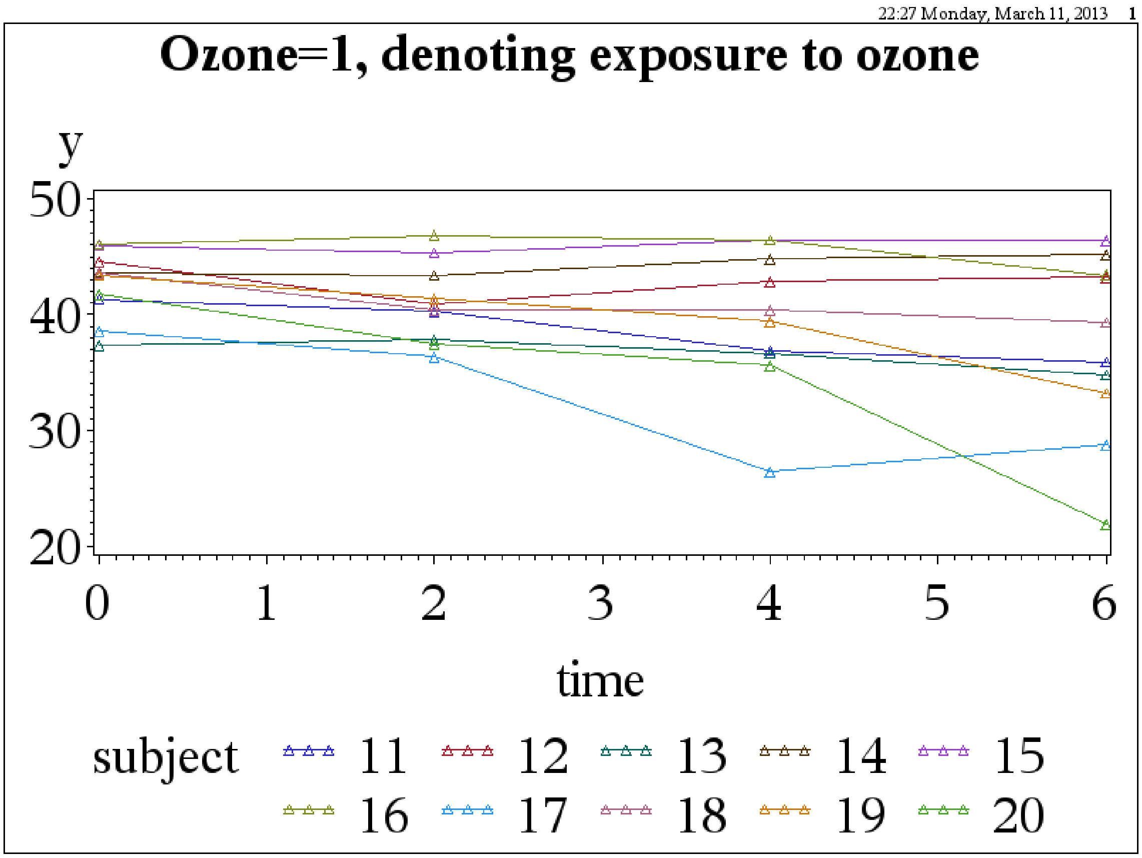
```
           0              0    10    10     42.3830000

                          1    10    10     42.6170000

           2              0    10    10     41.8680000

                          1    10    10     41.0090000

           4              0    10    10     42.5150000

                          1    10    10     39.6000000

           6              0    10    10     43.1240000

                          1    10    10     37.2110000
------------------------------------------------------
```

There are 10 observations in each group in each time point, with no missingness. The time points are evenly spaced at intervals of 2 hours.

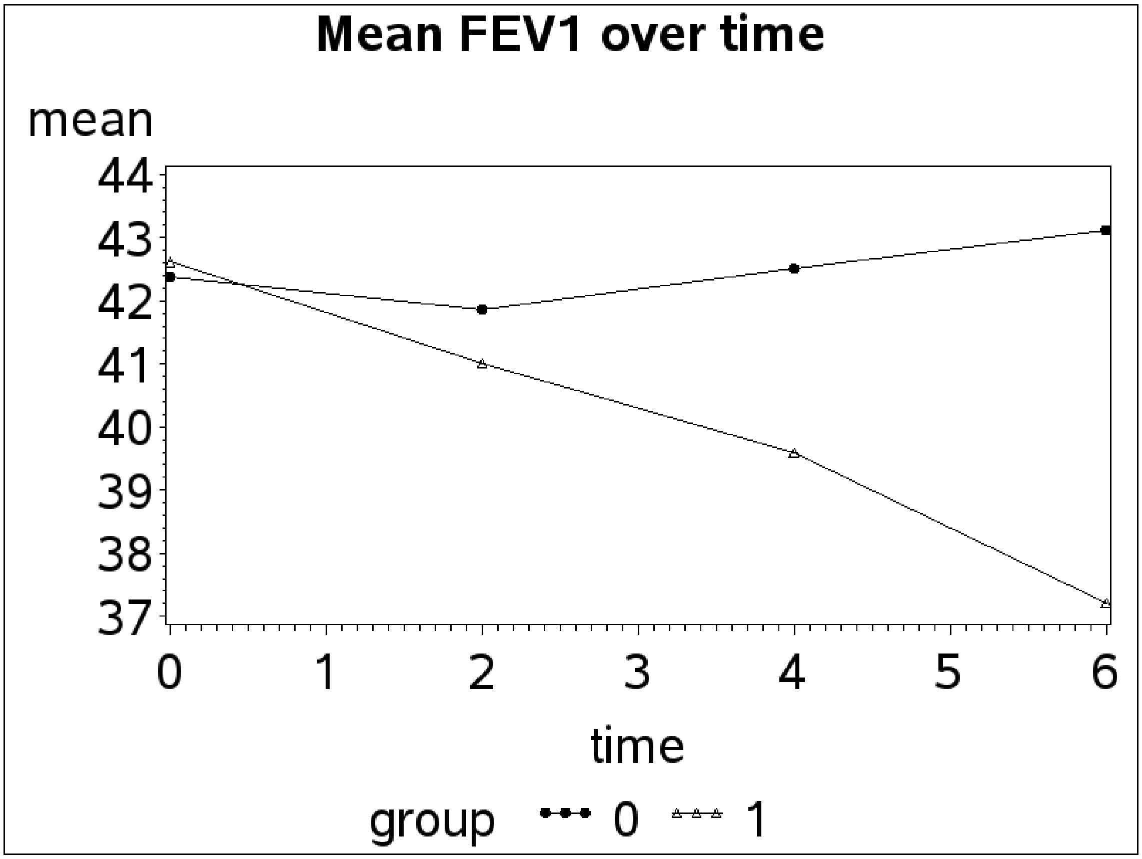
(b) Plot the FEV1 response against hour for each subject exposed to room air (overlaid on the same plot) and for each subject exposed to ozone (overlaid on a second plot). Comment on any patterns in the data or other notable aspects of the data.

There is a lot of variability from person to person in their FEV1 patterns over time. There are a few people whose FEV1 decreased a lot over time when exposed to ozone, and a few people in the room temperature group that increased their FEV1 over time.


22:27 Monday, March 11, 2013 1
Ozone=0, denoting exposure to room air
y
50
45
40
35
0
1
2
3
4
5
6
time
subject
1
2
3
4
5
6
7
8
9
10


22:27 Monday, March 11, 2013 1
Ozone=1, denoting exposure to ozone
y
50
40
30
20
0
1
2
3
4
5
6
time
subject
11
12
13
14
15
16
17
18
19
20

(c) Obtain the mean FEV1 value at each hour of measurement for ozone and room air subjects separately. Plot the means against hour. Comment on the pattern of change in mean FEV1 with hour for ozone and room air subjects.



Mean FEV1 decreases over time for individuals exposed to ozone, and increases slightly for those exposed to room air. The two groups have very similar means at baseline (as would be expected in a randomized study).

Problem 2

[30 points: 5 points for defining the maximal model, 5 for fitting the maximal model and getting the variance-covariance matrix, 3 points for suggesting a reasonable structure, 12=4*3 for fitting each of the four covariance structures. I did the LRT for each against the unstructured, but that is not totally necessary. They can instead compare AIC. All evidence should point to AR1-heterogeneous as being the best covariance structure. 5 points for concluding AR1-hetero.] Fitting a Maximal Model and Evaluating Variance-Covariance Structure:

(a) Define a reasonable "maximal" mean model for this study. Fit this model using an unstructured variance-covariance matrix. Comment on the variance structure and on the correlation structure. What simplified variance-covariance structure(s) might be reasonable? Justify your answer.

Let

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i0} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix} = \begin{pmatrix} \text{individual } i\text{'s FEV1 at hour 0} \\ \text{individual } i\text{'s FEV1 at hour 2} \\ \text{individual } i\text{'s FEV1 at hour 4} \\ \text{individual } i\text{'s FEV1 at hour 6} \end{pmatrix}$$

Lets consider the different covariates:

$$\begin{aligned} X_{1ij} &= 1 \text{ for all } i \text{ and } j, \\ X_{2ij} &= \begin{cases} 1 & \text{if corresponding measure at hour 2} \\ 0 & \text{otherwise,} \end{cases} \\ X_{3ij} &= \begin{cases} 1 & \text{if corresponding measure at hour 4} \\ 0 & \text{otherwise,} \end{cases} \\ X_{4ij} &= \begin{cases} 1 & \text{if corresponding measure at hour 6} \\ 0 & \text{otherwise.} \end{cases} \\ X_{5ij} &= \begin{cases} 1 & \text{if treatment is ozone} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + e_{ij} + \beta_6 X_{2ij} * X_{5ij} + \beta_7 X_{3ij} * X_{5ij} + \beta_8 X_{4ij} * X_{5ij} + e_{ij}$$

where $i = 1, \ldots, 20, \quad j = 0, \ldots, 3$. We assume $\mathbf{e}_i = \begin{bmatrix} e_{i0} \\ e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is unstructured.

Or if you prefer to replace the X's with something more descriptive:

$$Y_{ij} = \beta_1 + \beta_2 \text{hour2}_{ij} + \beta_3 \text{hour4}_{ij} + \beta_4 \text{hour6}_{ij} + \beta_5 \text{ozone}_{ij} + \beta_6 \text{hour2}_{ij} * \text{ozone}_{ij} + \beta_7 \text{hour4}_{ij} * \text{ozone}_{ij} + \beta_8 \text{hour6}_{ij} * \text{ozone}_{ij} + e_{ij}$$

where $i = 1, \ldots, 58, \quad j = 0, \ldots, 3$.

We could alternatively write the model in matrix form:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{e}_i \qquad i = 1, \ldots, 58$$

where

$$\mathbf{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & X_{5i0} & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & X_{5i1} & X_{5i1} & 0 & 0 \\ 1 & 0 & 1 & 0 & X_{5i2} & 0 & X_{5i2} & 0 \\ 1 & 0 & 0 & 1 & X_{5i3} & 0 & 0 & X_{5i3} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{pmatrix}$$

We assume $\mathbf{e}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$, so $\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is unstructured.

Looking at the correlation matrix, we see that a compound symmetry correlation would not fit well, so maybe an AR1 would be appropriate. Looking at the diagonal of the estimated variance matrix, we see that the variance may not be homogeneous across time.

```
proc sort data=ozone;
    by descending time descending group;
run;


proc mixed data=ozone;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=un r rcorr;
run;
```

```
-------------------------------------------------------------------------------------------------------------------
                        Estimated R Matrix for subject 1

               Row        Col1        Col2        Col3        Col4

                 1     36.8041     24.2632     14.9012     11.1966
                 2     24.2632     23.4591     12.9840     11.1510
                 3     14.9012     12.9840      9.7035      8.0468
                 4     11.1966     11.1510      8.0468      9.3296

                  Estimated R Correlation Matrix for subject 1

               Row        Col1        Col2        Col3        Col4

                 1      1.0000      0.8257      0.7885      0.6042
                 2      0.8257      1.0000      0.8606      0.7537
                 3      0.7885      0.8606      1.0000      0.8457
                 4      0.6042      0.7537      0.8457      1.0000
-------------------------------------------------------------------------------------------------------------------


                                Fit Statistics

                     -2 Res Log Likelihood          355.2
                     AIC (smaller is better)        375.2
                                  Dimensions

                      Covariance Parameters           10
```

(b) Keeping the same maximal mean model, evaluate whether your suggestion(s) for the variance-covariance structure from question 2a as well as the following models for the variance-covariance structure provide an adequate fit to the data compared with an unstructured variance-covariance:

i) compound symmetry

```
proc mixed data=ozone;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=cs r rcorr;
run;
                                Fit Statistics

                     -2 Res Log Likelihood           394.1
                     AIC (smaller is better)         398.1

                                  Dimensions

                      Covariance Parameters             2


                      394.1- 355.2 = 38.9

                      10 - 2 = 8


DATA pvalues;
chsq = SDF('chisquare',38.9,8);
RUN;
PROC PRINT DATA=pvalues;
RUN;

                     Obs          chsq

                                1     .000005130
```

ii) heterogeneous compound symmetry

```
proc mixed data=ozone;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=csh r rcorr;
run;

                                Fit Statistics
```

```
                     -2 Res Log Likelihood          365.3
                     AIC (smaller is better)        375.3
                                  Dimensions

                      Covariance Parameters            5

                     365.3- 355.2 = 10.1

                      10 - 5 = 5

DATA pvalues;
chsq = SDF('chisquare',10.1,5);
RUN;
PROC PRINT DATA=pvalues;
RUN;

                                Obs      chsq

                                 1     0.072451
```

iii) 1st-order autoregressive

```
proc mixed data=ozone;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=ar(1) r rcorr;
run;

                                Fit Statistics

                     -2 Res Log Likelihood          381.1
                     AIC (smaller is better)        385.1

                                  Dimensions

                      Covariance Parameters            2

                     381.1- 355.2 =25.9

                      10 - 2 = 8

DATA pvalues;
chsq = SDF('chisquare',25.9,8);
RUN;
PROC PRINT DATA=pvalues;
RUN;
```

```
                                   Obs          chsq

                                    1     .001092488
```

iv) heterogeneous 1st-order autoregressive

```
proc mixed data=ozone;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=arh(1) r rcorr;
run;
```

```
                                    Fit Statistics

                     -2 Res Log Likelihood           358.2
                     AIC (smaller is better)         368.2

                                      Dimensions

                      Covariance Parameters             5

                      358.2 - 355.2 = 3

                      10 - 5 = 5
```

```
DATA pvalues;
chsq = SDF('chisquare',3,5);
RUN;
PROC PRINT DATA=pvalues;
RUN;
```

```
                                    Obs      chsq

                                     1     0.69999
```

```
iii vs iv nested LRT:

                     381.1- 358.2 =25.9

                     5 - 2 = 3
```

```
DATA pvalues;
chsq = SDF('chisquare',25.9,3);
RUN;
PROC PRINT DATA=pvalues;
```

```
RUN;

                               Obs          chsq

                                1     .000010008

iv  heterogeneous 1st-order autoregressive parameters:

                        Estimated R Matrix for subject 11

               Row        Col1        Col2        Col3        Col4

                 1     37.6514     25.0563     13.4381     11.1475
                 2     25.0563     23.4532     12.5783     10.4343
                 3     13.4381     12.5783      9.4883      7.8710
                 4     11.1475     10.4343      7.8710      9.1838


                  Estimated R Correlation Matrix for subject 11

               Row        Col1        Col2        Col3        Col4

                 1      1.0000      0.8432      0.7110      0.5995
                 2      0.8432      1.0000      0.8432      0.7110
                 3      0.7110      0.8432      1.0000      0.8432
                 4      0.5995      0.7110      0.8432      1.0000

                         Covariance Parameter Estimates

                         Cov
                         Parm       Subject    Estimate

                         Var(1)     subject     37.6514
                         Var(2)     subject     23.4532
                         Var(3)     subject      9.4883
                         Var(4)     subject      9.1838
                         ARH(1)     subject      0.8432
```

Using likelihood ratio tests and the AIC criterion as appropriate, identify a model for the variance-covariance structure that provides a good fit to the data. Provide estimates for the parameters used in defining this variance-covariance model. Also provide estimates of the variance-covariance and correlation matrices.

We see that the heterogeneous 1st-order autoregressive model has the lowest AIC among all structures considered (including the unstructured) and the highest p-value for a likelihood ratio test against the unstructured covariance.

Problem 3

[20 points: 10 for testing coincident lines and concluding that we fail to reject the null hypothesis that the mean profiles do not coincide, 10 for testing that groups are parallel and concluding appropriately (it is borderline so it depends on what method they used, see below)] Fit the usual model for the analysis of mean profiles using room air exposure as the reference level for group and baseline as the reference group for time. Use the variance-covariance structure identified in your answer to question 2b. Based on this model:

(a) Test the null hypothesis that the pattern of means over hours is identical (coincides) for the two exposure groups. What do you conclude?

```
proc mixed data=ozone noclprint method=ml;
class time group t subject;
model y = group time group*time / solution;
repeated  t / subject=subject type=arh(1) r rcorr;
        title ’arh(1) Covariance Structure, Full Model, ML’;
run;

                                Fit Statistics

                     -2 Log Likelihood               369.1

proc mixed data=ozone noclprint method=ml;
class time group t subject;
model y = time / solution;
repeated  t / subject=subject type=arh(1) r rcorr;
        title ’arh(1) Covariance Structure, Reduced Model, ML’;
run;

                                Fit Statistics

                     -2 Log Likelihood               377.2

DATA pvalues;
chsq = SDF(’chisquare’,8.1,4);
RUN;
PROC PRINT DATA=pvalues;
RUN;

                                Obs      chsq

                                 1     0.087983
```

Testing that the mean profiles coincide

Structure	-2 Log Likelihood	Number of Parameters
Time only model	377.2	4
Saturated model	369.1	8
Difference	8.1	4

LRT yields $G^2 = 8.1$ with 4 df ($p = 0.08$), so we fail to reject the null hypothesis at $\alpha = 0.05$ that the mean profiles do not coincide.

(b) Test the null hypothesis that the mean response profiles of the two groups are parallel. What do you conclude?

```
with REML I get

                         Type 3 Tests of Fixed Effects

                              Num     Den
                Effect         DF      DF    F Value    Pr > F

                group           1      18       1.86    0.1900
                time            3      54       3.00    0.0382
                time*group      3      54       2.60    0.0615

with ML I get

                         Type 3 Tests of Fixed Effects

                              Num     Den
                Effect         DF      DF    F Value    Pr > F

                group           1      18       2.06    0.1682
                time            3      54       3.34    0.0259
                time*group      3      54       2.89    0.0438

the parallel model:

proc mixed data=ozone order=data method=ml;
class time group t subject;
model y = group time / solution;
repeated  t / subject=subject type=arh(1) r rcorr;
        title 'arh(1) Covariance Structure, Parallel Model, ML';
run;

                                Fit Statistics


                     -2 Log Likelihood              376.8
```

```
DATA pvalues;
chsq = SDF('chisquare',7.7,3);
RUN;
PROC PRINT DATA=pvalues;
RUN;
```

```
                                Obs      chsq

                                 1     0.052636
```

Testing that the mean profiles coincide

Structure	-2 Log Likelihood	Number of Parameters
Parallel model	376.8	5
Saturated model	369.1	8
Difference	7.7	3

LRT yields $G^2 = 7.7$ with 3 df ($p = 0.053$), so we fail to reject the null hypothesis at $\alpha = 0.05$ that the mean profiles are parallel. However, using the Type 3 F tests, we see that time*group have a significance at $p = 0.0438$ if we use ML and $p = 0.0615$ if we use REML. We see that this is a borderline case.

Problem 4

[12 points: 3 points for each part. It is ok here it they fit with REML or ML. No need to provide CI except on 4d.] Fitting a Linear Model in Time:

Fit a model that includes hour as a continuous variable, group and their interaction. Use the model for the variance-covariance structure that you identified in question 2b.

```
                         Solution for Fixed Effects


                                         Standard
Effect        group    Estimate         Error      DF    t Value    Pr > |t|

Intercept               41.8620        0.8300      18      50.44      <.0001
group         1          0.7328        1.1738      18       0.62      0.5403
group         0               0             .       .          .           .
time                   -0.05420        0.2162      58      -0.25      0.8029
time*group    1         -0.7911        0.3058      58      -2.59      0.0122
time*group    0               0             .       .          .           .
```

(a) What is the estimated rate of change in mean response for the room air group?

The estimated decrease in mean FEV1 is 0.05420 per hour for the room air group (CI: -0.05420 - 1.96*0.2162 = -0.477952, -0.05420 + 1.96*0.2162 = 0.369552).

(b) What is the estimated rate of change in mean response for the ozone group?

The estimated decrease in mean FEV1 is $0.05420 + 0.7911 = 0.8453$ per hour for the ozone group.

(c) Test the hypothesis that the rates of change in mean response are identical in the two groups. What do you conclude? Give a possible reason for any difference in conclusion you make from this test and the analogous test based on the mean profiles analysis from 3b.

To test that the rates of change in mean response are identical in the two groups, we only have to look to the parameter for time*group. We see that the parameter is significant, so we conclude that the change in mean response is different between the two groups. In 3b the test was only borderline significant, possibly because there was less power in 3b because we were estimating many parameters.

(d) What is the estimated difference in rate of mean change between the two groups? By calculating a 95% confidence interval for this difference, identify what are plausible values for the underlying true difference.

The estimated difference in rate is the time*group parameter, and it is -0.7911.

$(-0.7911 - 1.96 * 0.3058, -0.7911 + 1.96 * 0.3058) = (-1.39, -0.19)$

We believe with 95% certainty that the true difference in rates of mean change in FEV1 over time is between -1.39 and -0.19, where the ozone group has a more negative change in FEV1 over time.

Problem 5

[13 points: 10 for trying some other model (splines, quadratic), 3 for concluding] Evaluating the Fit of a Linear Model in Time:
Does a model with a linear trend in hour for each exposure group adequately describe the pattern of change in the two groups? Justify your answer with appropriate statistical analysis.

We see that we fail to reject the linear trend model with both the quadratic model and the saturated model as alternatives. Thus, we conclude that the linear trend adequately describes the pattern of change in the two groups.

```
---------- Uncentered time, linear  -----------------

                     -2 Log Likelihood              371.8

---------- Centered time, linear  -----------------

                                Fit Statistics

                     -2 Log Likelihood              371.8

---------- Uncentered time, quadratic -----------------

                     -2 Log Likelihood              369.9

---------- Centered time, quadratic -----------------

                     -2 Log Likelihood              369.9

------------ Saturated mean model ---------------------
```

```
                     -2 Log Likelihood               369.1

Quadratic vs linear

 371.8 -   369.9 = 1.9

DATA pvalues;
chsq = SDF('chisquare',1.9,2);
RUN;
PROC PRINT DATA=pvalues;
RUN;
                                 Obs      chsq

                                  1     0.38674

Saturated vs linear

371.8 - 369.1 = 2.7

DATA pvalues;
chsq = SDF('chisquare',2.7,4);
RUN;
PROC PRINT DATA=pvalues;
RUN;


                                 Obs      chsq

                                  1     0.60921
```

Problem 6

[10 points] Summarizing the Key Results and Conclusions:
Write a brief structured abstract (maximum 200 words) summarizing the objective, methods, results and conclusions that might be drawn concerning exposure differences in patterns of pulmonary function over time.

Background: In a study designed to examine the acute responses of normal subjects to ozone exposure, Follinsbee and colleagues randomized 20 subjects to exposure to either room air or 0.12 ppb ozone. The exposure lasted six hours. A baseline and 6 other measures of FEV1 (a measure of pulmonary function) made at hourly intervals were recorded for all study participants in the study. Subjects cycled on an exercise bike before a measure of FEV1 was obtained. The investigators were interested in determining whether changes in pulmonary function during the 6 hours of exposure were different in the ozone and room air exposed groups.

Methods: Multivariate linear regressions were used to study the association of FEV, time, and group.

Results: We see that we fail to reject the linear trend model with both the quadratic model and the saturated model as alternatives. Thus, we conclude that the linear trend adequately describes the pattern of change in the two groups. We see that in the linear model, the interaction of group and time is significant, indicating that the change in mean response is different between the two groups. The estimated decrease in mean FEV1 is 5.420 ml per hour for the room air group (95% CI: 47.8 ml decrease, to 37 ml increase). The estimated decrease in mean FEV1 is 84.53 ml per hour for the ozone group. We believe with 95% certainty that the true difference in rates of mean change in FEV1 over time is between -139 and -19, where the ozone group has a more negative change in FEV1 over time.

Conclusion: Exposure to ozone is associated with decreased lung function over time.