R Answers to Exercises: Module 4

If you have any questions as you go through these, feel free to ask them in the forum. For all of these, **you will need the long version of the data set**.

1. If you haven't done so, read the references listed on the module 4 page.

2. Using the Physical Training Data, run a marginal model to test if mean LDL levels change from pre- to post-training equally in the three training regimen groups. Support your answer. What do you conclude about the effects of training regimen on LDL?

Which covariance structure did you choose, and why?

All three training regimens have a significant effect on LDL—the mean value of LDL decreased across the three groups (χ^2=13.0539 or F = 33.19, p < .001), but there was no significant interaction.

I used an unstructured covariance structure because the variances are clearly different from pre- to post-training. Since there are only two variances and one covariance, there is no way to use fewer parameters with any other structure with unequal variance.

```
Analysis of Deviance Table (Type III tests)

Response: LDL
                             Df   Chisq Pr(>Chisq)
(Intercept)                   1 74.8941  < 2.2e-16 ***
as.factor(time)               1 13.0539  0.0003027 ***
as.factor(group)              2  0.3742  0.8293460
as.factor(time):as.factor(group)  2  0.1693  0.9188291
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Var/Covar matrix

```
          [,1]      [,2]
[1,] 0.9117772 0.1799371
[2,] 0.1799371 0.3261424
```

Means

```
time   lsmean        SE df lower.CL upper.CL
    1 2.850404 0.1949121 48 2.458507 3.242301
    2 1.748451 0.1165727 48 1.514066 1.982836
```

3. Do the same for HDL. Are the effects the same as on LDL? Support your answer. What do you conclude about the effects of training regimen on HDL?

The results are exactly the same as for LDL, but in the opposite direction. All three training regimens had a positive effect (χ^2=13.0539 or F = 20.9, p<.001).

```
Analysis of Deviance Table (Type III tests)

Response: HDL
                                Df    Chisq Pr(>Chisq)
(Intercept)                      1 313.6899  < 2.2e-16 ***
as.factor(time)                  1   8.9903   0.002714 **
as.factor(group)                 2   0.1863   0.911061
as.factor(time):as.factor(group)  2   0.3905   0.822629
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
            [,1]        [,2]
[1,] 0.041556326 0.005479427
[2,] 0.005479427 0.014864685
```

```
time   lsmean         SE df lower.CL upper.CL
    1 1.292628 0.04161146 48 1.208962 1.376293
    2 1.732929 0.09649973 48 1.538903 1.926954
```

4. Using the County data, test whether the mean number of jobs in Alabama changed across the 5 decades of the study, and whether the change differed for counties classified as rural and non-rural. Include a plot. Describe the findings and support your answer. (Note: It may be easier to read the output if you change the scale of the outcome variable to Thousands of Jobs).

Which covariance structure best fits these data?

Unstructured. The times clearly have different variances, and this is the model fit for the three covariance structures with non-constant variance:

UN:

```
     AIC      BIC    logLik
2166.275 2260.871 -1058.138
```

ARH:

```
     AIC      BIC    logLik
2300.461 2361.003 -1134.231
```

CSH:

```
     AIC      BIC    logLik
2440.214 2500.756 -1204.107
```

The unstructured is by far the best fit.

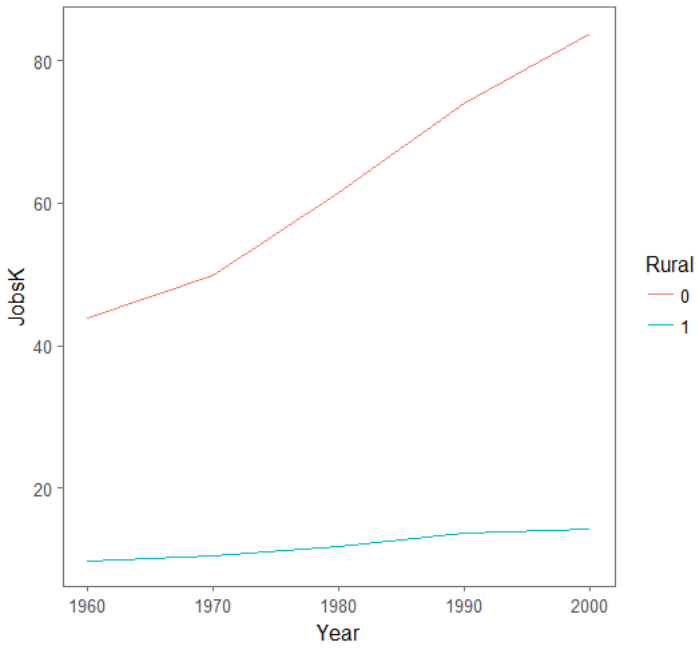
Once you choose a covariance structure, rerun the model using ML estimation (not REML). Now test the same model, but treat time (Decade) as continuous. Which model fits better? Would you reach the same conclusions using these two models?

Decade as categorical:

```
    AIC      BIC    logLik
2179.07 2274.423 -1064.535

    Response: JobsK
                  Df   Chisq Pr(>Chisq)
    (Intercept)    1 14.4879  0.0001411 ***
    decadeF        4 32.4273  1.564e-06 ***
    RuralF         1  1.5082  0.2194126
    decadeF:RuralF 4 16.6886  0.0022216 **
```

Decade as continuous:

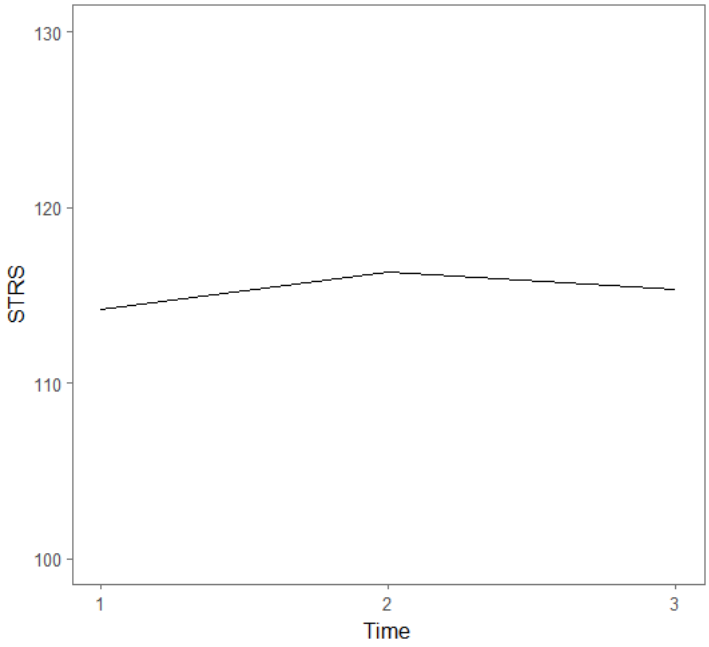
```
     AIC     BIC    logLik
2191.252 2263.72 -1076.626

    Response: JobsK
                 Df  Chisq Pr(>Chisq)
    (Intercept)   1 0.3371    0.56152
    decade        1 6.0077    0.01424 *
    RuralF        1 6.4944    0.01082 *
    decade:RuralF 1 3.6212    0.05705 .
```

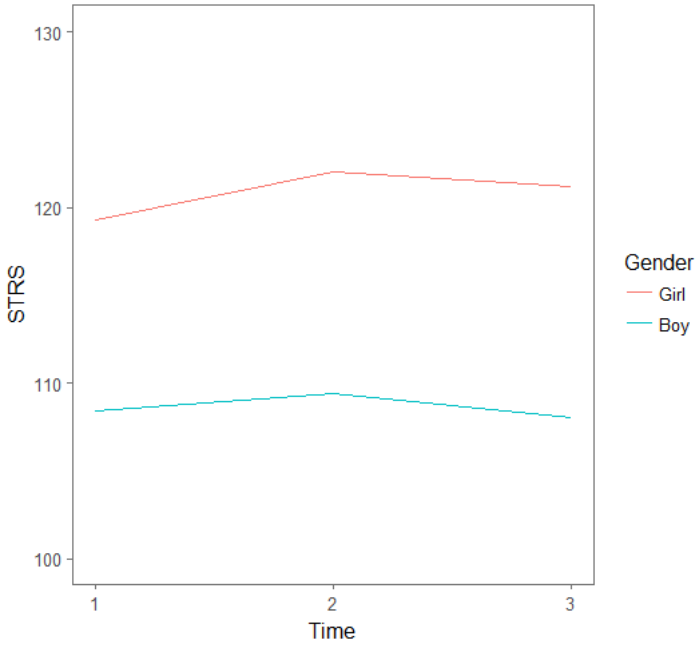


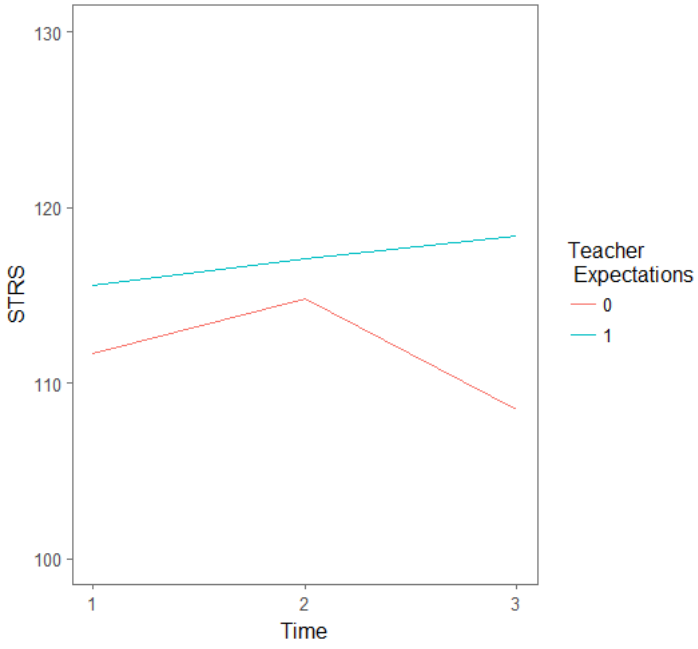
The categorical data fit better as per the -2LL and the AIC, but not by the BIC, which favors a more parsimonious model. That is because fitting four dummy variables for Decade is a much less parsimonious model than one slope and one intercept (especially b/c decade is involved with an interaction). Because of that and because the graph indicates a near linear form, I would use Decade as a continuous variable.

The two models actually led to different conclusions about the effects of time and rural, and I am more inclined to trust the second model, especially given the graph. However, it seems that this model needs a bit of refinement, and I would continue to pursue a better fit (perhaps through a change in slope or even a cubic model—both curves have a very slight S-shape).

5. Using the Teacher data, plot teacher's ratings of rapport with students (STRS) across time, with:
    a. An overall average trajectory over time.
    b. A separate trajectory for each gender.
    c. A separate trajectory for high and low values of Teacher Expectancy.
    d. What do the trajectories look like? Are they linear? Do they tend to vary much in height or slope over time?







They are all close to linear. The one exception is the trajectory for students with low teacher expectations, and even that is not far off.

6. Now test whether children's summer expectancies and gender predict teacher's ratings of rapport with each student. Treat student as the subject. Does it make sense to treat time as continuous or categorical? Which covariance structure for the residuals fits best? Why?

I treated Time as continuous to fit a line, despite the slightly off-linear trend above. Since teacher expectancy is really a continuous variable that we have categorized only so we could graph it, and it's not that far off linear, it seems quite reasonable to treat Time as a continuous variable, so that we can fit a line for each of these predictors.

Unstructured:

```
       AIC      BIC    logLik
  1800.059 1837.878 -889.0297


Response: STRS
              Df   Chisq Pr(>Chisq)
(Intercept)    1 84.5385  < 2.2e-16 ***
time           1  3.4816    0.06205 .
tOTchExp       1  1.7335    0.18797
Gender         1 16.9103  3.919e-05 ***
time:tOTchExp  1  4.1086    0.04266 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Variance matrix:

```
         [,1]     [,2]     [,3]
[1,] 181.8369 124.8857 120.3648
[2,] 124.8857 192.9108 122.8048
[3,] 120.3648 122.8048 189.1831
```

Compound symmetry:

```
       AIC      BIC    logLik
  1792.299 1816.365 -889.1493



Response: STRS
              Df   Chisq Pr(>Chisq)
(Intercept)    1 84.5071  < 2.2e-16 ***
time           1  3.7056    0.05423 .
tOTchExp       1  1.8244    0.17679
Gender         1 17.0050  3.728e-05 ***
time:tOTchExp  1  4.3035    0.03803 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Variance matrix:

```
       [,1]   [,2]   [,3]
[1,] 187.27 122.42 122.42
[2,] 122.42 187.27 122.42
[3,] 122.42 122.42 187.27
```

I started with an unstructured covariance structure to see the pattern without any constraints. This is a perfect example of one with near-equal variances and near-equal covariances—a perfect candidate for Compound Symmetry. The -2RLL is basically identical in the two models and the other

information criteria that take model complexity into account are better in the CS model.

The Fixed effects are mostly unchanged in the two models. In both we have a significant Time*Expectancy interaction such that teacher's reports of rapport have the largest decrease over time for kids with the lowest expectancies. As expectancy rises, so does the slope of the trajectory. Likewise, there is a large Gender effect, such that teachers rate their rapport with girls higher than their rapport with boys.