Classical Methods in Longitudinal (Repeated Measures) Analysis

2018-04-04

Preparation

We will use the data in the *seasonal-depression.csv* file. These data include the Beck depression scores in four different seasons for 14 males under 35 years of age. The source of these data is: Myers & Well (2003). We will use these data to explore seasonal depression. The attributes in the dataset include:

- `subject`: The subject ID number for each male
- `s1`: The Beck depression score in winter (season 1)
- `s2`: The Beck depression score in spring (season 2)
- `s3`: The Beck depression score in summer (season 3)
- `s4`: The Beck depression score in fall (season 4)

The data are displayed below:

subject	s1	s2	s3	s4
1	7.50	11.55	1.000	1.21
2	7.00	9.00	5.000	15.00
3	1.00	1.00	0.000	0.00
4	0.00	0.00	0.000	0.00
5	1.06	0.00	1.097	4.00
6	1.00	2.50	0.000	2.00
7	2.50	0.00	0.000	2.00
8	4.50	1.06	2.000	2.00
9	5.00	2.00	3.000	5.00
10	2.00	3.00	4.208	3.00
11	7.00	7.35	5.877	9.00
12	2.50	2.00	0.009	2.00
13	11.00	16.00	13.000	13.00
14	8.00	10.50	1.000	11.00

Exploring Seasonal Depression

We might explore seasonal depression by examining the mean depression score across seasons. If they vary, it is evidence supporting seasonal depression. If not, it is evidence against seasonal depression.

season	M	SD
s1	4.29	3.36
s2	4.71	5.18
s3	2.58	3.60
s4	4.94	4.97

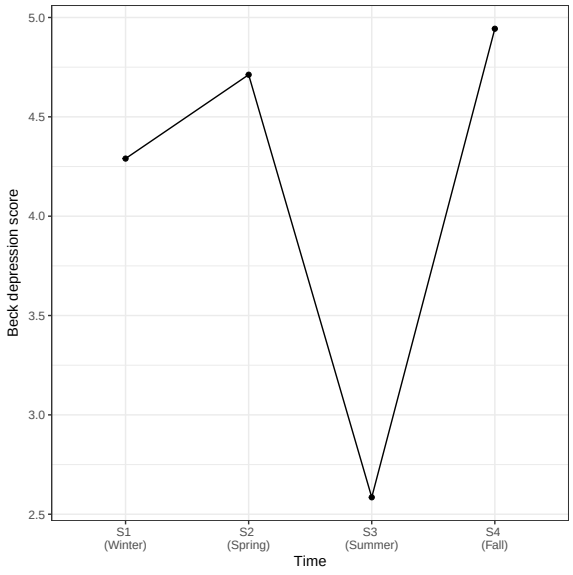

Figure 1: Plot of the mean depression score by season.

We can also explore this graphically.

Both the summary values and the plot suggest that there are sample differences in the average depression score across seasons. There is, however, a great deal of variation in the scores within each season. Are these differences only due to sampling error? Or are they more than we would expect because of chance?

If the cases were independent, we could fit an ANOVA model (or a regression model with three indicators to represent season) and examine the results for significance. However, with repeated measures data, the observations (and thus the model errors) would not be independent. So, we have to fit a model that accounts for this non-independence.

Classical Methods for Modeling Repeated Measures Data

There are two classical methods for modeling repeated measures data: (1) Repeated Measures ANOVA (RM-ANOVA); and (2) Multivariate ANOVA (MANOVA). I will not teach you these methods, but I will show you the output from fitting these procedures in case you encounter them in the research literature you read.

It is recommended that you fit a mixed-effects regression model (LMER) rather than use either one of the classical methods to model repeated measures data. The LMER models are far more flexible than either of these methods. The LMER models allow for continuous or categorical predictors, and missing data. They also have a less stringent set of assumptions than either of the classical methods.

RM-ANOVA

The key to modeling repeated measures data is to include subject in the model explicitly. RM-ANOVA fits a model that includes a main-effect of time, a main-effect of subject, and an interaction effect between subject and time. The ANOVA table for this model fitted to our seasonal depression data looks like this:

Table 3: ANOVA Table for the Effects of Seasons and Subjects

Source	df	SS	MS	F	p
Subjects	13	779.0	59.93	11.30	<.001
Seasons	3	47.8	15.93	3.00	.042
Subjects x Seasons	39	207.0	5.31		

In this table there is no "Error" line. That is because there is only one measurment per cell in the table that divides up subjects (rows) and seasons (columns). The error measures the within-cell variation, but with only one measurment per cell, there is no within-cell variation. In a repeated measures design, the interaction between subject and time takes the place of the error variance in the calculations of F.

The hypotheses associated with the main-effect being tested by the F-tests are, respectively,

- H_0: The mean depression score (collapsing across seasons) is the same for every subject.
- H_0: The mean depression score (collapsing across subjects) is the same for every season.

The first of these we reject. The significant p-value ($p < .001$) indicates that the mean depression score (collapsing across seasons) is not the same for every subject. While this might be interesting in some research problems, it is not of interest here.

The more important hypothesis is whether the The mean depression score (collapsing across subjects) is the same for every season. We also reject this hypothesis ($p = .042$). There are likely seasonal differences in the mean depression scores. This is evidence that supports the scientific hypothesis of seasonal depression.

If you were interested in which seasons differed, you would need to go in and test each pairwise contrast (Winter vs. Fall, etc.) and then adjust the p-values to account for the number of tests you carried out.

Assumptions for Using RM-ANOVA

One of the major assumptions for valid output from the RM-ANOVA is *sphericity*. This assumption says that the difference scores between measurements at any two time points are the same in the population. To examine this, we would need to compute the difference scores between every two timepoints and evaluate the variances of those differences. With four time points (seasons) there are six different difference scores to compute.

d1	d2	d3	d4	d5	d6
4.054	-10.55	-6.500	0.208	-10.35	-6.29
2.000	-4.00	-2.000	10.000	6.00	8.00
0.000	-1.00	-1.000	0.000	-1.00	-1.00
0.000	0.00	0.000	0.000	0.00	0.00
-1.059	1.10	0.038	2.903	4.00	2.94
1.500	-2.50	-1.000	2.000	-0.50	1.00
-2.500	0.00	-2.500	2.000	2.00	-0.50
-3.440	0.94	-2.500	0.000	0.94	-2.50
-3.000	1.00	-2.000	2.000	3.00	0.00
1.000	1.21	2.208	-1.208	0.00	1.00
0.354	-1.48	-1.123	3.123	1.65	2.00
-0.500	-1.99	-2.491	1.991	0.00	-0.50
5.000	-3.00	2.000	0.000	-3.00	2.00
2.500	-9.50	-7.000	10.000	0.50	3.00

The variance for each of the six sets of difference scores are displayed below:

```
##    d1    d2    d3    d4    d5    d6
##  6.23 13.94  6.88 12.14 14.26 10.23
```

This analysis suggests that the sphericity assumption is not met. (Note: Some data analysts use a test, Mauchley's Test, to examine the sphericity assumption. This test is known to yield signiicant results, saying that sphericity has been violated, when it shouldn't. This is especially true when the distributions of difference scores are non-normal. The methodological advice is: DO NOT USE MAUCHLEY'S TEST to examine the assumption of sphericity.)

Because computing the difference scores can be a pain, especially with a lot of time points, data analysts often examine a related property called *compound symmetry*. Compound symmetry looks at the raw data rather the difference scores. In order for compound symmetry to be met, we must have:

- Homogeneity of variance of the raw data at each time point
- Equal correlations between time points

To check this we compute the sample variances at each time point and we also compute the correlations between time points.

season	Var
s1	11.3
s2	26.9
s3	13.0
s4	24.7

rowname	s1	s2	s3	s4
s1				
s2	.917			
s3	.718	.694		
s4	.772	.724	.714	

There is evidence that the property of compound symmetry is not met: the sample variances are not the same, and the correlations between time points is not the same. This tells us the same story as examining the variances of the differences—the assumption of sphericity is likely violated!

Epsilon-Adjusted Tests

When the sphericity assumption has been violated, the F-distribution we use to compute the p-value does not have the dfs that we calculated in the ANOVA output. One attempt to remedy this is to adjust the degrees of freedom, called an epsilon-adjustment. There are several ways to amake these adjustments. Two common methods employed inthe social sciences are the (1) Greenhouse–Geiser epsilon-adjustment; and (2) Hyunh–Feldt epsilon-adjustment (named after the statisticians who derived them).

It is typical to present the p-values based on the unadjusted F-test, and the epsilon-adjusted tests for the time effect in an ANOVA table. (In these tables, the interaction effect is sometimes referred to as the Error term)

Table 7: ANOVA Table for the Within-Subject Effect of Season. Epsilon-Adjusted p-Values are also Provided (GG = Greenhouse–Geiser; HF = Hyunh–Feldt).

Source	df	SS	MS	F	p	GG	HF
Seasons	3	47.8	15.93	3.00	.042	.053	.042
Subjects x Seasons	39	207.0	5.31				

MANOVA

Another common method for modeling repeated measures data is to use a multivariate ANOVA. Multivariate analyses, whether regression or ANOVA, model more than one outcome simultaneously. In these types of analyses, we can account for the variation in each outcome, and the correlation between the outcomes. The following pseudo-syntax shows the idea behind multivariate analysis:

```
manova( c(s1, s2, s3, s4) ~ 1 )
```

In this multivariate analysis, we use time to predict a vector of outcomes. The subject is included in the model not as a predictor, but by connecting all of their repeated measures in the outcome. (Note: This syntax won't work in R, but is presented to illustrate the idea of multivariate analyses.)

There are many different multivariate statistics that have been derived over the years. Social scientists tend to use one of the following in repeated measures analsysis: (1) Pillai's trace; (2) Wilk's lambda; (3) Hotelling's T^2; or (4) Roy's largest root. The results of fitting a MANOVA to the depression data are displayed below.

```
Multivariate Tests: Season
                 Df test stat approx F num Df den Df Pr(>F)
Pillai            1     0.392     2.37      3     11   0.13
Wilks             1     0.608     2.37      3     11   0.13
Hotelling-Lawley  1     0.645     2.37      3     11   0.13
Roy               1     0.645     2.37      3     11   0.13
```

Based on the multivariate analysis, there is no statistically significant effect of time.

Example: Sleep Quality among Nurses

Hasson & Gustavsson (2010) carried out a longitudinal study to monitor the development of sleep quality in Swedish nurses. They collected data about sleep quality from 1,114 nurses using a self-rated Likert-scale item. The nurses answered this item at four diffeent times during the beginning of their career: during their last semester at university, and at three subsequent annual follow-ups once the nurses had entered working life.

The primary analytic method used in the study is RM-ANOVA and is described as follows:

> A repeated measures ANOVA was conducted to assess possible change in mean sleep quality over time, and effect size was calculated as eta squared statistics. The interpretations were based on established cut-offs [56]. ANCOVAs were then utilized to measure potential interaction between factors such as age groups (divided by quartiles) and sex, as well as factors such as whether or not the nurses were committed to a steady relationship, were living alone (or with parents), had children at home, and whether or not they had previous nursing assistant training, other previous experience in healthcare, or felt their education had prepared them well enough to work as nurses. The covariate in the ANCOVAs was baseline sleep quality.

Because of dropout, they only had complete data on only 846 participants. They report their results using these cases. (They also tried other analyses where they imputed missing values and analyzed more "complete" datasets.) The start by writing about the RM-ANOVA results with no covariates included in the model:

The repeated measure ANOVA indicates a general significant decrease in sleep quality over time (Figure 3, time effect: $F=19.147_{df=3}$, $p<.0001$, eta sq=.022, t3–t4 is ns). The most pronounced decline in sleep quality for the whole group occurred immediately after the transition from study life to working life. There was no significant change between t3–t4, although the change between t2–t3 was significant ($p<.05$).

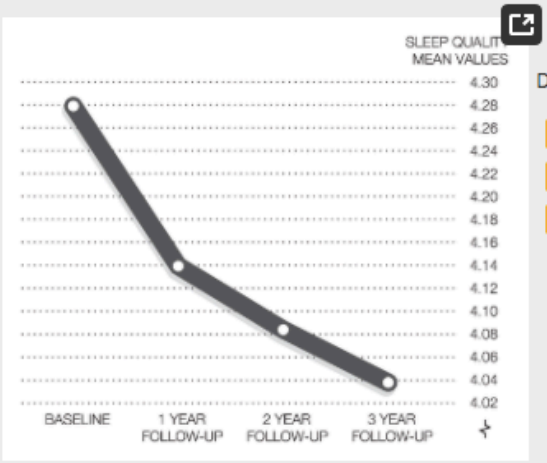

Figure 3. Repeated measures ANOVA illustrating the deterioration in sleep quality mean values over time. (n=846).
https://doi.org/10.1371/journal.pone.0014265.g003

Figure 2: Analytic results of fitting the Subject x Time RM-ANOVA.

After summarizing the longitudinal differences, they re-evaluate the mean differences over time by controlling for other covariates (sex, age, etc.)

There were statistically significant sex ($t=-2.06_{df=1}$, $p<.05$) and age ($F=5.731_{df=2}$, $p=.001$) related differences at baseline, where men (mean =4.42±.78) rated better sleep quality than women (mean =4.24±.90) and participants older than 29 (29–36 years mean =4.13±.95; >36 years mean =4.18±.95) rated worse sleep quality than those younger than 29 (24–28 years mean =4.34±.82; <24 years mean =4.41±.78). However, the two-way ANCOVA that adjusts for possible initial differences showed no significant interaction effects between the four age groups over time. There were no differences in sleep quality decline between women and men, even if there was a tendency ($p=.079$) for men to decline more immediately after the transition but have better recovery in the subsequent follow-ups. Furthermore, there were no differences in sleep quality decline regardless of whether the participants at baseline had partners, were living alone (or with parents), living with children at home, had previous experience from work within healthcare, or whether they felt their education had prepared them well enough to work as nurses. However, the nurses who had previous nursing assistant training exhibited marginally but consistently better sleep quality over time compared to those that did not (time effect: $F=84.691_{df=3}$, $p<.0001$, eta sq =.092; time x group effect: $F=2.851_{df=3}$, $p<.05$, eta sq =.003).

Figure 3: Analytic results of including various covariates in the Subject x Time RM-ANOVA.

Methodological Concerns

Despite doing many things well (e.g., examining selection bias, thinking about drop out), there are several methodological concerns directly related to the use of RM-ANOVA:

- The authors make no note of examining the spericity assumption underlying this model.
- The outcome is a single item that has a rating of 1–5. Is the distribution at each time point reasonably normally distributed? It is unclear.
- The listwise deletion removed 25% of the nurses from the study. In RM-ANOVA if a subject is missing data at even one time point they are eliminated from the analysis. This seems to be throwing away information.
- The plot of the mean sleep quality over time suggests that there is a log-linear relationship. RM-ANOVA treats each time point as discrete, not allowing for fitting functional forms (at least not easily). By treating time continuously, rather than discretely, we can reduce the number of predictors in the model (one continuous time predictor vs three dummy coded time points). We can also model more complex non-linear relationships.
- The RM-ANOVA only gives us information about mean differences in sleep quality. We do not get an indication about how individual nurses' sleep quality changes over time (or how variable their trajectories are).

References

Hasson, D., & Gustavsson, P. (2010). Declining sleep quality among nurses: A population-based four-year longitudinal study on the transition from nursing education to working life. *PLOS One*, *5*(12), e14265. Retrieved from https://doi.org/10.1371/journal.pone.0014265

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.