# Lab 7 - Longitudinal Analysis

In this lab, you will analyze longitudinal data from an educational research study.

## Setup

- Create a project in RStudio for this lab.
- Download the **BuCKETS_DATA.xlsx** data file from Canvas and move it to your project folder.
- Open a new R script for this lab and insert comments with your name and lab number.
- Load the `tidyverse`, `janitor`, `readxl`, and `dataedu` packages.
- Install and load the `nlme` and `emmeans` packages.

This lab contains example code that you will need to type into your R script and run. Additionally, there are exercises in green text that you should complete. Be sure to use appropriate comments in your R script so that it is easy to identify where your solutions to these exercises are located.

## Background

The data for this assignment were collected by Dr. Brooke Whitworth and colleagues as part of the first year of the NSF-funded grant *Building Content Knowledg of Elementary Teachers of Sciences* (BuCKETS). This quasi-experiment included two undergraduate general physics classes: a control class (*n* = 19) where a traditional lecture style of instruction was used and a treatment class (*n* = 21) where a conceptually rich curriculum was used. Both classes were measured at three time points (pretest, posttest, and delayed post) on three different scales:

- the *Making Sense of Science Assessment* (MSSA), a 32-item assessment of physics content knowledge consisting of four 8-item subscales (Waves, Forces & Motion, Energy, and Matter) with each item scored as 1 = correct or 0 = incorrect such that possible sum scores for the total scale range from a low of 0 to a high of 32,
- the *Preservice Elementary Teacher Affect Scale for Science* (PETAS-S), a 7-item survey with 5-point Likert scale response options such that possible sum scores range from a low of 7 to a high of 35, and
- a *pedagogical content knowledge* (PCK) assessment with 8 items, each scored as 1 = correct or 0 = incorrect such that possible sum scores range from a low of 0 to a high of 8.

The primary objective is to determine whether these data provide any statistical evidence that the conceptually rich instructional approach leads to improvements in physics content knowledge, affect for science, and/or pedagogical content knowledge over time as compared to traditional lecture classes, among undergraduate non-physics majors. To do this, we will fit linear mixed effects models to assess the significance of the treatment and time effects and their interaction in predicting the mean scale scores. You will be provided with example code for the model for MSSA scores and then fit the model for PETAS-S scores as a lab exercise and the model for the PCK scores as a homework exercise.

## Load and Process Data

To begin, use the `read_excel` function to import the data file. Note that many of the variable names contain spaces and are a mix of uppercase and lowercase letters. Use the `clean_names` function of the `janitor` package to create names that are easier to work with.

```
buckets <- read_excel("BuCKETS_DATA.xlsx")
buckets <- clean_names(buckets)
```

Next, there are several steps that need to be taken to process the data.

- This analysis will focus just on the MSSA total score, so we will drop the MSSA subscale scores to help focus our thinking.
- As part of the homework this week, you will be asked to consider some of the demographic variables collected. So, we will go ahead and convert these to the appropriate data type (character), re-label the response options to align with the information provided in the code book for this data, and create new variables from existing variables for both the participants' race and whether or not they were an in-state student.
- Note that there are 8 students who do not have any delayed-post measurements. In the homework, you will be asked to conduct a missing data analysis that considers whether our approach of using all the available information is warranted (as opposed to excluding these students and/or dropping all delayed-post measurements). To assist in this analysis, we will go ahead and create a missing_dp variable to indicate whether the participant is missing the delayed-post measurements or not.

```
buckets <- buckets %>%
  select(!starts_with(c("wav", "fm", "eg", "mtr")), -science_courses, -age) %>%
  mutate_at(vars(id_number:type_of_hs), ~as.character(.)) %>%
  mutate(treatment = if_else(treatment == "1", "treatment", "control")) %>%
  mutate(sex = if_else(sex == "1", "female", "male")) %>%
  mutate(race = case_when(black == "1" ~ "Black",
                          hispanic == "1" ~ "Hispanic",
                          TRUE ~ "White")) %>%
  mutate(year = case_when(year == "1" ~ "Freshman",
                          year == "2"  ~ "Sophomore",
                          year == "3" ~ "Junior",
                          TRUE ~ "Senior")) %>%
  mutate(major = case_when(major == "0" ~ "Elementary Ed",
                           major == "1" ~ "Secondary Ed",
                           major == "2" ~ "Business",
                           major == "3" ~ "Music Ed",
                           major == "4" ~ "Accounting",
                           major == "5" ~ "Mathematics",
                           major == "6" ~ "Undeclared",
                           TRUE ~ "Social Work")) %>%
  mutate(in_state = if_else(home_state == "MS", "yes", "no")) %>%
  mutate(type_of_hs = if_else(type_of_hs == "0", "public", "private")) %>%
  select(id_number, treatment, sex, race, year, major, in_state, type_of_hs,
mssa_sum_pre:petas_s_sum_dp) %>%
  mutate(missing_dp = if_else(is.na(mssa_sum_dp), "yes", "no"))
```

## Longitudinal Analysis of MSSA Scores

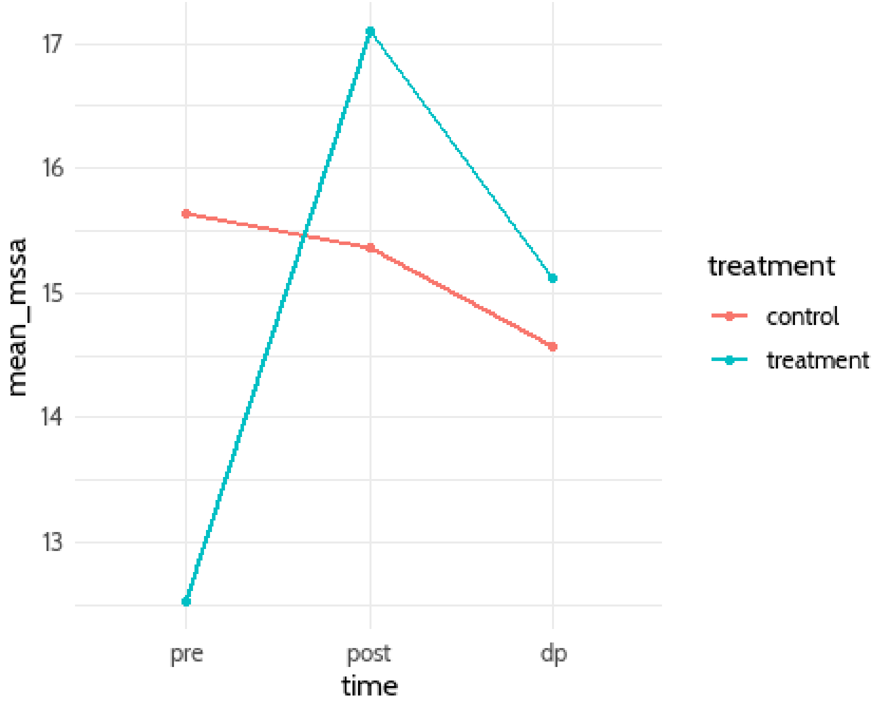
The longitudinal analysis will require that the data be in ‘long’ format with three rows per participant, one for each time point (pre, post, dp), with one column that records the time of the measurement and another column that records the MSSA score at that time point.

```
#pivoting data
buckets_mssa <- buckets %>%
  pivot_longer(cols = starts_with("mssa"),
               names_to = "time",
               values_to = "mssa_score") %>%
  select(!starts_with(c("pck", "petas"))) %>%
  mutate(time = str_replace(time, "mssa_sum_", ""))
```

### Visualize Data

Before fitting the model for MSSA scores, we will get a sense of the change over time for both groups by creating a line plot of mean MSSA scores over the three measurement time points separately for each group.

```
#line plot
buckets_mssa %>%
  group_by(treatment, time) %>%
  summarize(mean_mssa = mean(mssa_score, na.rm = T)) %>%
  ggplot(aes(x = time, y = mean_mssa, color = treatment)) +
  geom_line(aes(group = treatment), size = 1) +
  geom_point(pch = 19) +
  theme_dataedu() +
  scale_x_discrete(limits = c("pre", "post", "dp"))
```



The plot reveals that the treatment class had a much lower mean MSSA total content score than the control class at pretest. Then, the control class mean did not change much over time, whereas the treatment class mean saw a large increase from pretest to posttest and then a smaller decline from posttest to delayed post.

### Model Data

To determine if the treatment had a significant effect on content knowledge (as measured by the MSSA) over time, we can fit a linear mixed effects model with a subject-specific random effect and with fixed effects for time, treatment, and their interaction using the `lme` function of the `nlme` package. In the call to `lme`, the first term specifies the two-way factorial structure in the model for the mean. The random argument specifies a subject-specific random effect and should be of the form ~ t | s where t is the time covariate and s is an identifier for each subject. The `na.omit` option tells R to use all of the available information in fitting the model, rather than excluding individuals who are missing a measurement at any time point. Asking for a `summary` of the model gives estimates for the regression model with indicator variables, whereas the `anova` function will give an ANOVA table for the fitted model.

```
mssa_lme <- lme(mssa_score ~ treatment*time,
                data = buckets_mssa,
                random = ~time | id_number,
                na.action = na.omit)

anova(mssa_lme)

##                 numDF denDF  F-value p-value
## (Intercept)         1    68 413.8904  <.0001
## treatment           1    38   0.3561  0.5542
## time                2    68   5.9173  0.0043
## treatment:time      2    68   6.1746  0.0034
```

At this point, any model refinement should keep in mind the principle of hierarchy in factorial ANOVA: if you include an interaction term in a model, then you also include the main effects that created that term. Here, because the *treatment* x *time* interaction term is statistically significant, we will also keep both the *time* and the *treatment* main effects and consider this the final model for MSSA scores.

As with factorial ANOVA, follow-up confidence intervals can be used to determine which particular treatment by time combinations are significantly different from each other. To do this, we will use the `contrast` function of the `emmeans` package with a conservative tukey adjustment.

```
#follow-up confidence interval
contrast(emmeans(mssa_lme, c("time", "treatment")), 'tukey')

##  contrast                          estimate    SE df t.ratio p.value
##  dp control - post control           -0.986 1.225 68  -0.804  0.9658
##  dp control - pre control            -1.249 1.367 68  -0.913  0.9418
##  dp control - dp treatment           -0.703 1.878 38  -0.374  0.9990
##  dp control - post treatment         -2.712 1.794 38  -1.512  0.6587
##  dp control - pre treatment           1.859 1.756 38   1.059  0.8945
##  post control - pre control          -0.263 1.001 68  -0.263  0.9998
##  post control - dp treatment          0.283 1.820 38   0.155  1.0000
##  post control - post treatment    -1.727 1.733 38  -0.997  0.9162
##  post control - pre treatment      2.845 1.694 38   1.679  0.5533
##  pre control - dp treatment        0.546 1.779 38   0.307  0.9996
##  pre control - post treatment     -1.464 1.690 38  -0.866  0.9522
##  pre control - pre treatment       3.108 1.650 38   1.884  0.4273
##  dp treatment - post treatment    -2.009 1.215 68  -1.654  0.5664
##  dp treatment - pre treatment      2.562 1.345 68   1.905  0.4079
##  post treatment - pre treatment    4.571 0.952 68   4.803  0.0001
##
## Degrees-of-freedom method: containment
## P value adjustment: tukey method for comparing a family of 6 estimates
```

The meaningful contrasts will be those comparing two groups at the same time point (e.g., pre control - pre treatment) and those comparing the same group at two time points (e.g., post treatment - pre treatment). Not surprisingly, the noticeable increase from pretest to posttest in the treatment group was highly statistically significant.

---

## Exercise

Carry out a longitudinal analysis of the PETAS-S scores by completing the following steps.

- Create a data frame where the PETAS-S scores have been pivoted to long format, with one column for time and another column for PETS-S scores.
- Construct a line plot to visualize the change in mean PETAS-S score over time separately for each group (control and treatment).
- Fit a linear mixed effects model for the PETAS-S scores with a subject-specific random effect and fixed effects for time, treatment, and their interaction. Request an ANOVA table for the fitted model.

---

You should have found that in the model for the mean PETAS-S scores, the *treatment* x *time* interaction is not statistically significant.

```
anova(petas_lme)

##                 numDF  denDF   F-value p-value
## (Intercept)         1     68 1594.0077  <.0001
## treatment           1     38    0.7481  0.3925
## time                2     68    5.2816  0.0074
## treatment:time      2     68    0.9521  0.3910
```

We can remove the non-significant interaction term and reassess the significance of the two main effects.

```
petas_lme.me <- lme(petas_score ~ treatment + time,
                    data = buckets_petas,
                    random = ~time | id_number,
                    na.action = na.omit)
```

```
anova(petas_lme.me)

##             numDF denDF   F-value p-value
## (Intercept)     1    70 1574.7154  <.0001
## treatment       1    38    0.7143  0.4033
## time            2    70    5.1687  0.0081
```

We see that the treatment effect is still not statistically significant, so we can remove it and fit a model with time as the only effect.

```
petas_lme.time <- lme(petas_score ~ time,
                      data = buckets_petas,
                      random = ~time | id_number,
                      na.action = na.omit)

anova(petas_lme.time)

##             numDF denDF   F-value p-value
## (Intercept)     1    70 1578.6194  <.0001
## time            2    70    5.1895  0.0079
```

Now, we have a final model where only the main effect of time is statistically significant. The contrast statement reveals that the delayed post mean is significantly higher than both the pretest and posttest mean.

```
#follow-up confidence intervals
contrast(emmeans(petas_lme.time, "time"), 'tukey')

## contrast   estimate    SE df t.ratio p.value
## dp - post     2.152 0.846 70   2.546  0.0346
## dp - pre      2.327 0.781 70   2.979  0.0109
## post - pre    0.175 0.817 70   0.214  0.9750
##
## Degrees-of-freedom method: containment
## P value adjustment: tukey method for comparing a family of 3 estimates
```

Note how the conclusion might differ if the correlation between observations from the same individual were ignored by fitting a simple linear regression model.

```
#comparing results to linear model
petas_lm <- lm(petas_score ~ time, data = buckets_petas)
anova(petas_lm)

## Analysis of Variance Table
##
## Response: petas_score
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## time        2  137.11  68.557  2.5161 0.08546 .
## Residuals 109 2969.99  27.248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```