Classical Procedures for Analysis

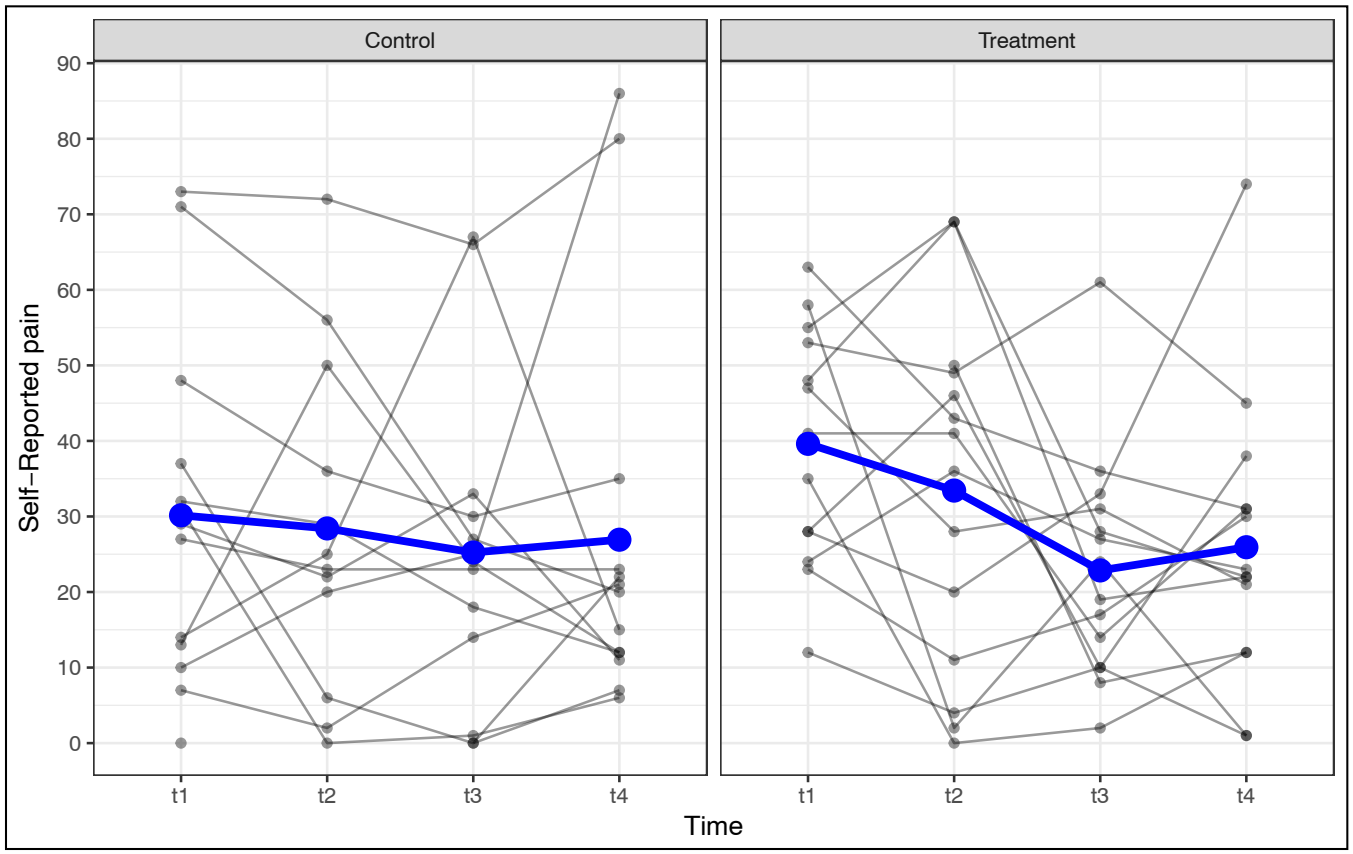
Andrew Zieffler

Introduction

There are three classical methods for analyzing repeated measures data:

- **Analysis of summary statistics:** Summary statistics are computed over the responses for each individual, and these summaries are then analyzed.
- **Univariate repeated measures ANOVA:** Analysis of variance model is fitted to the RM data. This method is reliable for data coming from designed experiments.
- **Multivariate ANOVA (MANOVA):** A multivariate (multiple outcome variables) analysis of variance is fitted to the RM data.

Example 1: The data in the file *backpain.csv* were collected from an experiment in which patients were treated for the discomfort of back pain. Half the subjects received an injection of pain medication at the site of the pain (treatment). The other half received an injection of morphine as a general pain treatment (control). Patients gave a verbal rating of discomfort (on a scale from 0–100, where lower scores indicate less pain) at four time points during the study.



Primary RQ:
Is there a treatment effect on pain?

	time	treatment	M	SD	N
1	t1	Control	30.15385	22.89777	13
2	t2	Control	28.41667	22.04317	12
3	t3	Control	25.23077	21.48315	13
4	t4	Control	26.92308	26.09426	13
5	t1	Treatment	39.61538	15.81423	13
6	t2	Treatment	33.42857	23.27995	14
7	t3	Treatment	22.85714	15.07855	14
8	t4	Treatment	25.92857	18.77586	14

Control

	rowname	t1	t2	t3	t4
1	t1				
2	t2	.647			
3	t3	.350	.629		
4	t4	.213	.412	.430	

Treatment

	rowname	t1	t2	t3	t4
1	t1				
2	t2	.440			
3	t3	.507	.283		
4	t4	-.048	.227	.486	

Analysis of Summary Statistics

The simplest method is to use a single value to summarize each individual's scores.

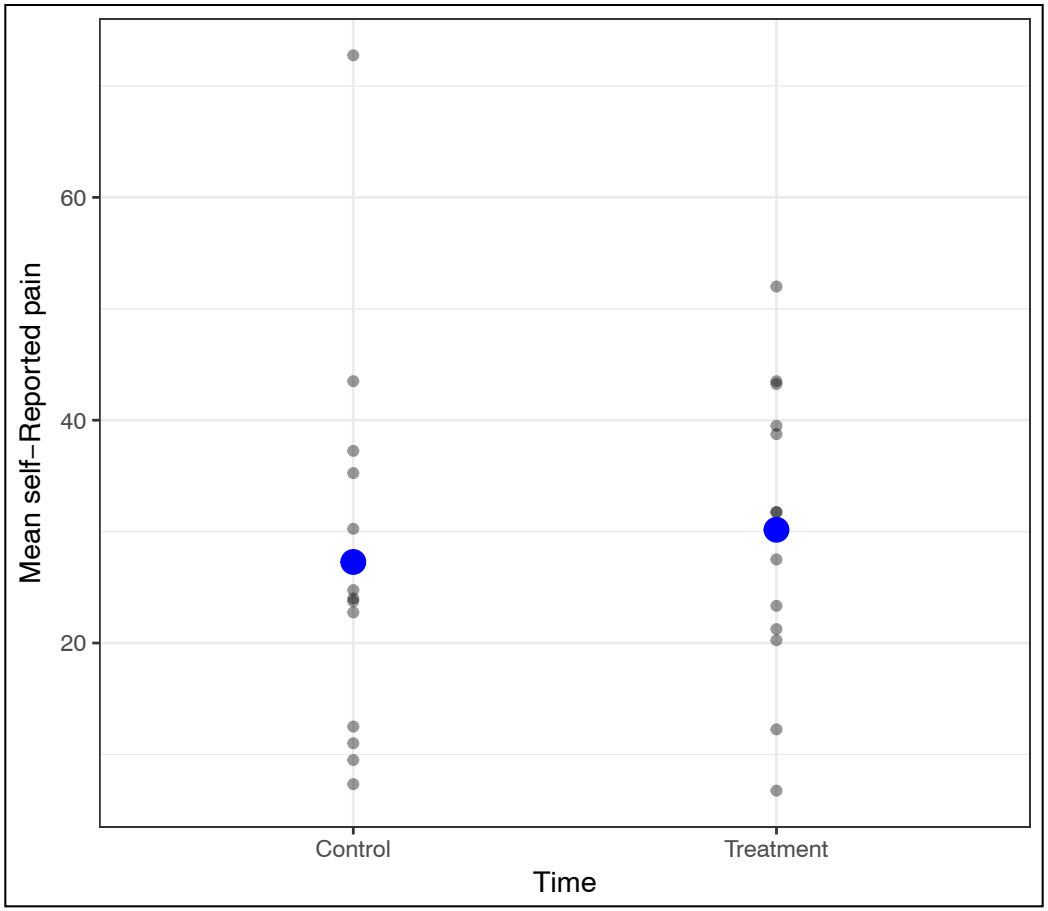
Several different summaries have been suggested:

- The mean score
- The maximum score
- The difference between the largest and smallest scores
- The difference between the first and last scores
- The score at the final wave
- The first principle component of scores

After distilling individuals' data to a single score, one then conducts group comparisons (e.g., *t*-test, ANOVA). Note: It is essential to pre-specify which summary you intend to use *and* which comparisons will be made before examining the data!!!

Example 1: Say we decide to use the *mean* pain score across measurement waves to summarize each individual's pain. Then, we will compare the average control pain scores to the average treatment pain scores.

id	mean_pain	treatment
1	23.75	Control
2	9.50	Control
3	35.25	Control
4	12.50	Control
5	24.00	Control
6	43.50	Control
:	:	:
:	:	:
22	43.25	Treatment
23	31.75	Treatment
24	31.75	Treatment
25	27.50	Treatment
26	38.75	Treatment
27	12.25	Treatment



The sample data suggests that there are mean differences in the the average treatment scores and the average control scores.

The primary hypothesis to be tested is:

$$H_0 : \mu_{\text{Treatment}} = \mu_{\text{Control}}$$

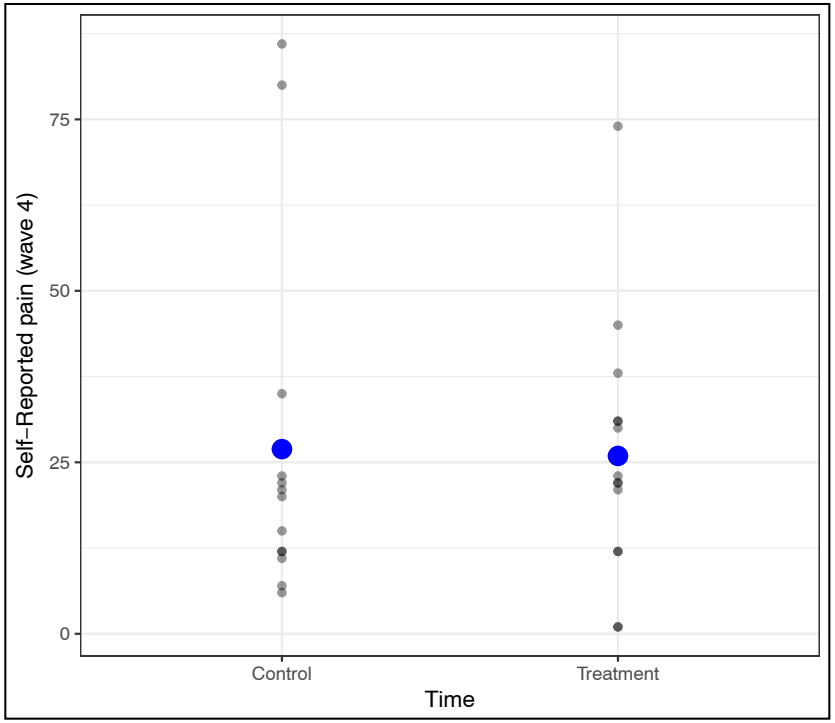
Assuming equal variance:
$t(25) = -0.491$, $p = 0.628$

Assuming unequal variance:
$t(21.652) = -0.485$, $p = 0.633$

The results of the *t*-test are non-significant. It is likely that there is NOT a treatment effect on pain. NOTE: Only one of these *t*-tests would be used, and you would need to make that decision prior to conducting the analysis.

Example 2: Say we decide to use the pain score from the last wave to summarize each individual's pain. Again, we will compare the average control pain scores to the average treatment pain scores.

id	t4	treatment
1	11	Control
2	6	Control
3	86	Control
4	7	Control
5	23	Control
6	20	Control
:	:	:
:	:	:
22	31	Treatment
23	31	Treatment
24	21	Treatment
25	23	Treatment
26	74	Treatment
27	12	Treatment



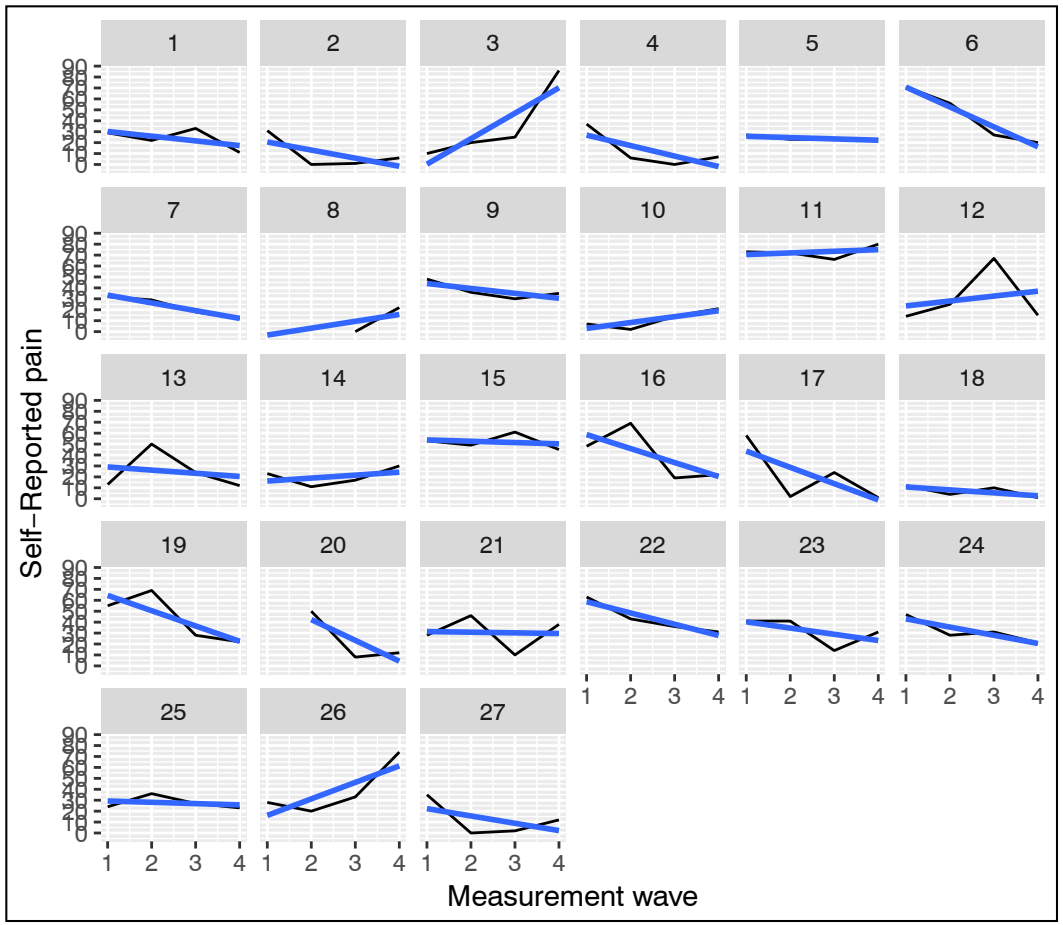
$H_0 : \mu_{\text{Treatment}} = \mu_{\text{Control}}$

Assuming unequal variance: t(21.685) = 0.113, p = 0.911

The results of the *t*-test are non-significant. It is likely that there is NOT a treatment effect on pain.

Although the analyses addressed the primary RQ (is there a treatment effect), these analyses do no address the effect of time. One summary measure to examine effect of time is the **regression slope**. computed on scores for each individual.

Example 2: Say we decide to fit a regress pain score on wave for each individual. Then we can summarize each individual's average rate-of-change in pain over time by using the slope. We can then compare the whether the slopes are the same/different for the control and treatment groups. (*Note:* This pre-supposes that it is reasonable to use the regression slope as a summary measure to describe the pattern of change for individuals.)



In order to fit the regression, we will need to convert time to an numeric value.

Does a linear model seem appropriate?

We are going to group by ID, and then fit a regression using wave to predict pain. We will use **dplyr**'s do() function to "do something" for each individual. What we will do is use **broom**'s tidy() function to output the regression for each individual in a tidy manner.

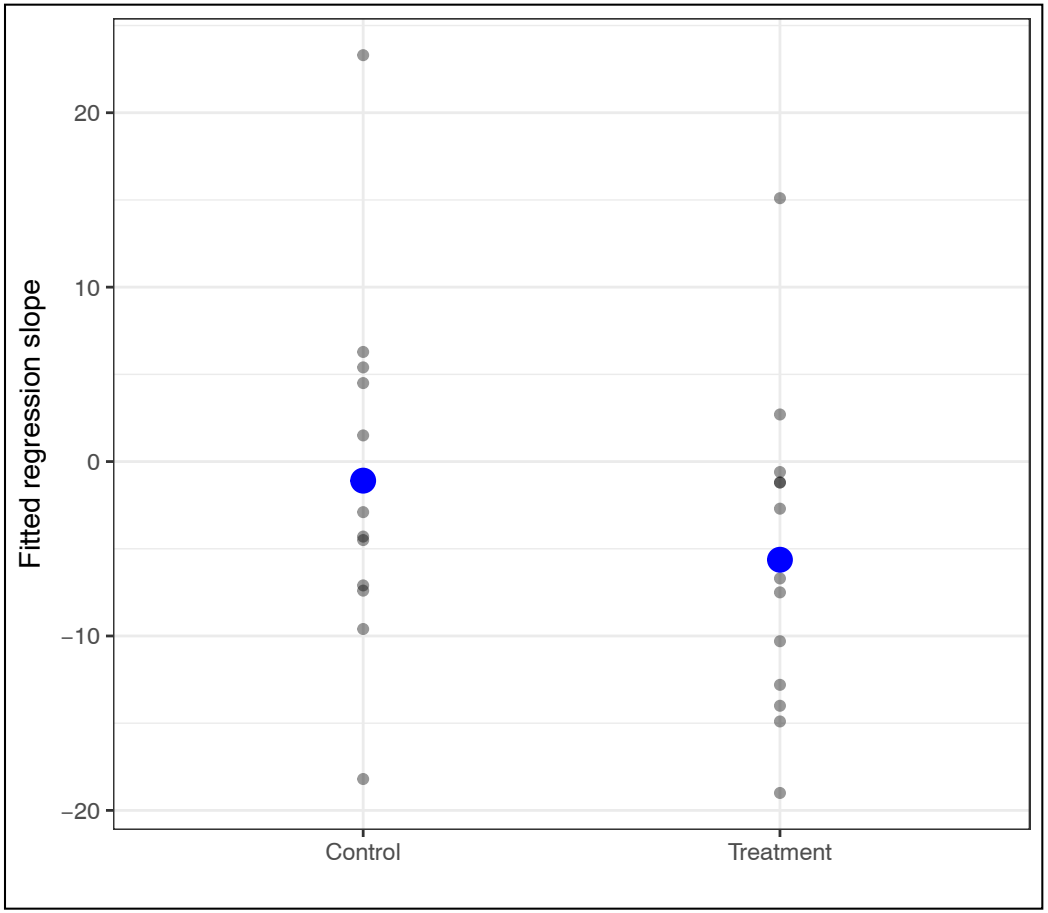
```
# Use the do() function along with the tidy() function
> backpain_4 %>%
    group_by(id) %>%
    do(tidy(lm(pain ~ 1 + wave, data=.)))
```

	id	term	estimate	std.error	statistic	p.value
	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	(Intercept)	34.5	11.8205330	2.9186501	0.10007825
2	1	wave	-4.3	4.3162484	-0.9962355	0.42410140
3	2	(Intercept)	28.0	16.5045448	1.6965024	0.23188151
4	2	wave	-7.4	6.0266077	-1.2278881	0.34438783
5	3	(Intercept)	-23.0	25.0444605	-0.9183668	0.45537526
6	3	wave	23.3	9.1449440	2.5478560	0.12565894
7	4	(Intercept)	36.5	16.6177616	2.1964450	0.15920711
8	4	wave	-9.6	6.0679486	-1.5820833	0.25444623
9	5	(Intercept)	27.0	1.8973666	14.2302495	0.00490199
10	5	wave	-1.2	0.6928203	-1.7320508	0.22540333

Note that all we care about is the summary of the slope…not the *p*-value or the intercept. So, we can filter this data frame to keep only the slope rows, and select only the ID and slope estimate.

```
> backpain_5 = backpain_4 %>%
    group_by(id) %>%
    do(tidy(lm(pain ~ 1 + wave, data=.)))  %>%
  filter(term == "wave") %>%
  select(id, Slope = estimate)

backpain_5 = data.frame(backpain_5) %>%
  mutate(treatment = factor(backpain$treatment, levels = c(0, 1), labels =
c("Control", "Treatment")))

backpain_5

    id      Slope treatment
1    1  -4.300000   Control
2    2  -7.400000   Control
3    3  23.300000   Control
4    4  -9.600000   Control
5    5  -1.200000   Control
6    6 -18.200000   Control
:    :       :          :
22 22 -10.300000 Treatment
23 23  -5.700000 Treatment
24 24  -7.500000 Treatment
25 25  -1.200000 Treatment
26 26  15.100000 Treatment
27 27  -6.700000 Treatment
```

We need to coerce this back to a regular data frame (it is a grouped data frame) before we can mutate on the treatment



The primary hypothesis to be tested is:

$$H_0 : \mu_{\text{Treatment}} = \mu_{\text{Control}}$$

Assuming equal variance:
$t(25) = 1.26$, $p = 0.219$

The results of the *t*-test are non-significant. It is likely that there is NOT a treatment effect on rate-of-change.

Here the research question we focused on was whether the improvement in pain was greater under one treatment than the other. There are other RQs one could test as well. For example, is the population mean slope *negative* (pain improves) under one or both treatments.

Advantages to analyzing summary measures:

- It is a simple method
- Although the procedure is crude and misses a lot of interesting detail, the analysis of summary measures certainly captures the big picture

It is important to note that I am not advocating that one avoid more sophisticated analyses, or suggesting that this method is satisfactory. There are many reasons to pursue more advanced methods of analyzing longitudinal data.

Repeated Measures ANOVA

One more sophisticated method of analyzing longitudinal data is to use RM-ANOVA.

Before we talk about RM-ANOVA, let's do a quick review of ANOVA analyses.

ANOVA focuses on explaining variation. It is an analysis typically undertaken to examine group differences (the *t*-test is the two-group simplification of an ANOVA). In ANOVA analyses, all predictors need to be categorical.

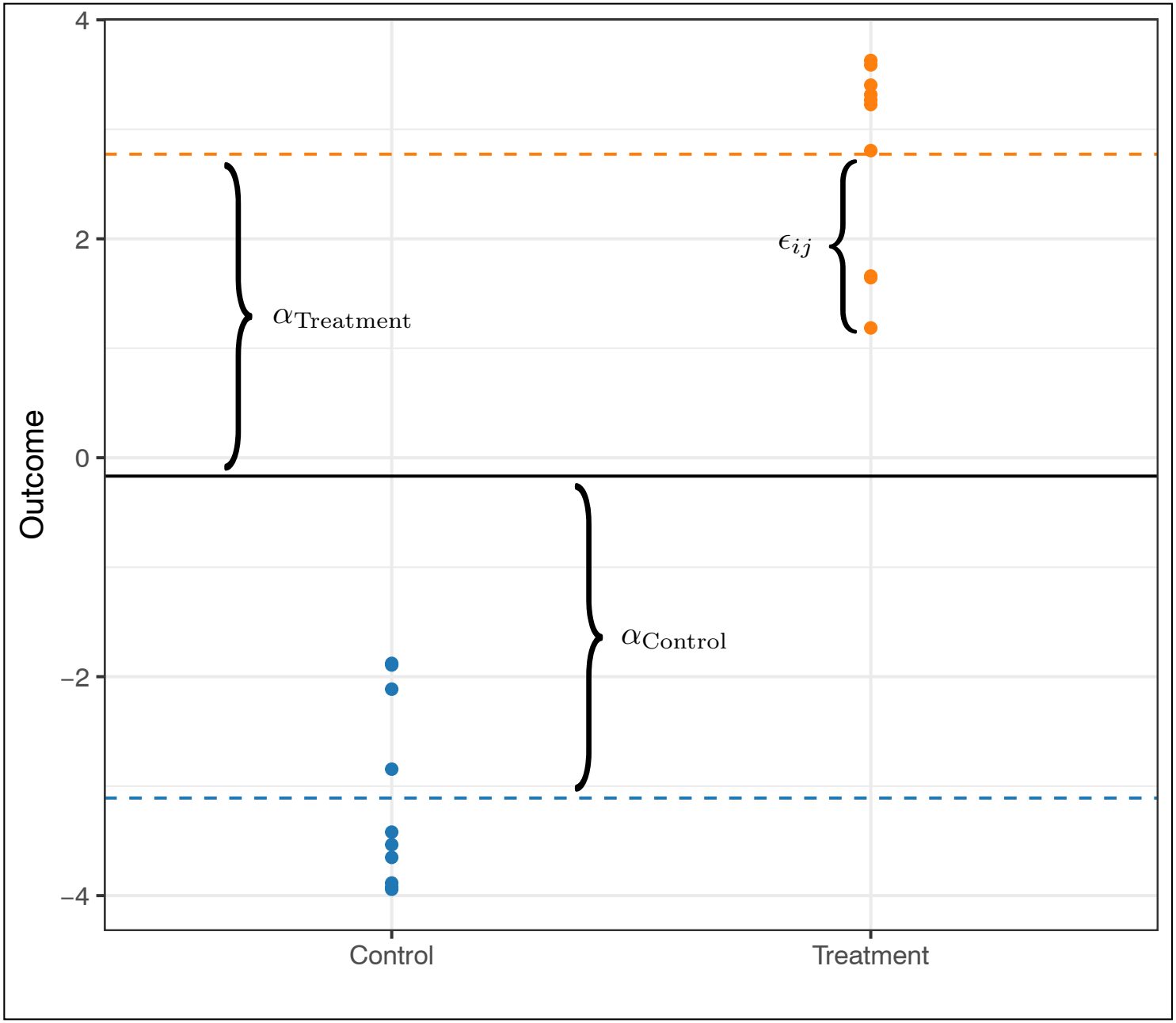
If we have a single categorical predictor (say treatment), the model underlying the ANOVA analysis is:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

where

- Y_{ij} is the outcome for individual i in condition j
- μ is the average effect (baseline) across all individuals in all groups
- α_j is the average effect (deviation from the baseline) for condition j
- ε_{ij} is the residual for individual i in condition j

Here is a graphical representation of this model.



In general we use this model to test whether there is a treatment effect,

$$H_0 : \alpha_{\text{Control}} = \alpha_{\text{Treatment}} = 0$$

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Fixed-effect

Random-effect

The assumptions for the model are on the random-effects (in this model, the residuals)

In longitudinal data we are interested in the effect of time (τ) and the effect of treatment (γ), and their interaction ($\tau\gamma$). This gives a model of:

$$Y_{ijk} = \mu + \tau_j + \gamma_k + (\tau\gamma)_{jk} + \epsilon_{ijk}$$

This would be an appropriate model, except that the ANOVA model has an assumption that the residuals are independent. This is untrue in longitudinal data; residuals are correlated within an individual. The RM-ANOVA adds an additional within-subjects (random) factor to account for correlation in the repeated measures.

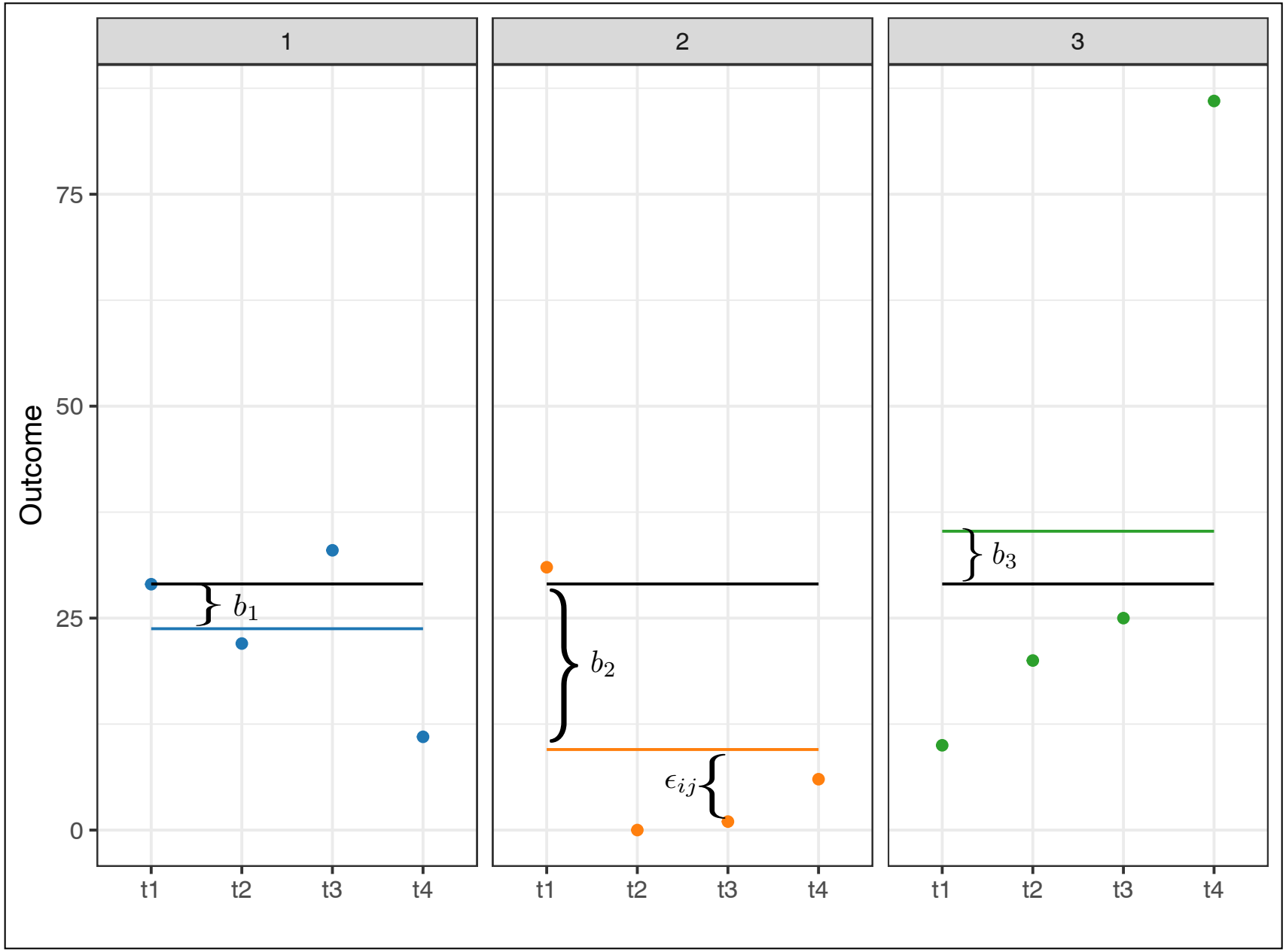
If we have a single categorical predictor (say treatment), the model underlying the ANOVA analysis is:

$$Y_{ijk} = \mu + \tau_j + \gamma_k + (\tau\gamma)_{jk} + b_i + \epsilon_{ijk}$$

where

- Y_{ijk} is the outcome for individual i at time j in condition k
- μ is the average effect (baseline) across all individuals at all time points in all groups
- τ_j is the main effect at time j
- γ_k is the main effect for condition k
- $(\tau\gamma)_{jk}$ is the interaction effect (time j and condition k)
- b_i is a random individual-specific effect
- ε_{ij} is the residual for individual i at time j in condition k

Graphically, the b_i term produces a different line for each subject.



The RM-ANOVA partitions variation as follows:

$$Y_{ijk} = \mu + \tau_j + \gamma_k + (\tau\gamma)_{jk} + b_i + \epsilon_{ijk}$$

Fixed-effects

Random-effects

The assumptions for the model are still on the random-effects (in this model, the individual-specific effects and the residuals).

$$b_i \sim \mathcal{N}(0, \sigma_b^2)$$

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

The individual-specific random effects and the residuals are assumed to be independent of one another.

The two sets of random effects allow a further partitioning of the variation into between-subjects variation (σ_b^2) and within-subject variation (σ_ϵ^2).

To fit the RM-ANOVA, we need to have data in the long format, BUT if an ID is missing data on ANY time point that ID needs to be deleted.

```
# Eliminate cases with missing data and make long format
> bp = backpain %>%
  na.omit() %>%
  gather(time, pain, t1:t4) %>%
  mutate(
    treatment = factor(treatment,
    levels = c(0, 1),
    labels = c("Control", "Treatment"))
  )
```

Then we will use the `ezANOVA()` function from the **ez** package.

```
# Fit RM-ANOVA
> ezANOVA(data = bp, dv = pain, wid = factor(id), within = time, return_aov = TRUE)
```

All predictors used in the model—including the time predictor (time), the ID predictor (id)—need to be factors.

There is a lot of output. The first part of the output ($ANOVA) gives the results of the *F*-test.

```
$ANOVA
  Effect DFn DFd        F         p p<.05       ges
2   time   3  72 2.066663 0.1121846       0.0392454
```

Remember, the hypothesis being tested here is that the mean pain value at all levels of time are equal.

$$H_0 : \mu_{\text{Time 1}} = \mu_{\text{Time 2}} = \mu_{\text{Time 3}} = \mu_{\text{Time 4}}$$

The results of this analysis suggest that we cannot reject this hypothesis.

$F(3, 72) = 2.07$

$p = 0.112$

The RM-ANOVA assumes the following structure on the variance–covariance matrix of the repeated measures:

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix}$$

The key here is that :(1) the variance at each time point is the same; and (2) the covariance between any two time points is equal to the covariance between any other two time points (the correlation is always the same).

This assumption is referred to as **compound symmetry**.

When subjects are randomly assigned to the within-subjects factor the homogenous correlation will be satisfied. In longitudinal data, this assumption is almost NEVER satisfied.

The assumption of constant variance is also almost never satisfied in longitudinal data.

"Sphericity is a less restrictive form of compound symmetry (in fact much of the early research into repeated measures ANOVA confused compound symmetry with sphericity). Sphericity refers to the equality of variances of the differences between treatment levels. Whereas compound symmetry concerns the covariation between treatments, sphericity is related to the variance of the differences between treatments. (Field, 1998, pp. 14–15)"

This less restrictive assumption is what we need to meet in order to have valid results from the RM-ANOVA. There primary method statisticians use to examine whether the assumption of sphericity is met is to statistically test the assumption via Mauchly's test. The null hypothesis for Machly's test is:

H_0: The sphericity assumption is reasonably satisfied.

The second part of the `ezANOVA()` output gives the results of Mauchly's test.

```
$`Mauchly's Test for Sphericity`
  Effect         W         p p<.05
2   time 0.7945441 0.3893358
```

The results of this analysis suggest that we cannot reject the null hypothesis ($p = 0.389$), indicating that the sphericity assumption may be reasonably satisfied.

If sphericity were violated, one option is to use adjusted results when interpreting output. Two common adjustment methods are:

- Greenhouse–Geiser epsilon adjustment
- Hyunh–Feldt epsilon adjustment

For our purposes we can consider epsilon to be a descriptive measure indicating the degree to which sphericity has been violated.

- If sphericity is met perfectly then epsilon will be exactly 1.
- If epsilon is below 1 then sphericity is violated. The further epsilon gets away from 1 the worse the violation.

The third part of the `ezANOVA()` output gives the adjusted results for the ANOVA.

```
$`Sphericity Corrections`
  Effect      GGe     p[GG] p[GG]<.05       HFe     p[HF] p[HF]<.05
2   time 0.855157 0.1227605           0.9663034 0.1145663
```

Analysis	Epsilon	p-Value
Unadjusted	1	0.112
Hyunh–Feldt	0.966	0.114
Greenhouse–Geiser	0.855	0.123

Note that as the epsilon value gets further from 1 (more sphericity violation), the *p*-value gets larger. With these data, the epsilon value for both adjustments are still relatively close to 1 (remember the lower-bound for epsilon for these data is 1/3).

Now let's also include the between-subjects treatment factor.

```
# Fit RM-ANOVA
> ezANOVA(data = bp, dv = pain, wid = factor(id),
          within = time, between = treatment, return_aov = TRUE)
```

Let's start by examining the results of Mauchly's test.

```
$`Mauchly's Test for Sphericity`
          Effect        W         p p<.05
3           time 0.791311 0.4060877
4 treatment:time 0.791311 0.4060877
```

The results of both tests suggest that we cannot reject the null hypothesis ($p = 0.406$, $p = 0.406$), indicating that the sphericity assumption may be reasonably satisfied.

Based on failing to reject Mauchly's test, we report and interpret the unadjusted *F*-test.

```
$ANOVA
          Effect DFn DFd           F         p p<.05         ges
2      treatment   1  23 0.08243136 0.7765999       0.001894772
3           time   3  69 2.02024347 0.1191028       0.039672179
4 treatment:time   3  69 0.46093318 0.7104713       0.009337413
```

Starting with the highest order term (the interaction) we find the data seem to support a time x treatment interaction, $F(3, 69) = 0.461$, $p = 0.710$. Similarly, both main-effects are also non-significant.

This suggests that there are no mean differences in pain over time, nor between the control and treatment subjects.

Comments regarding using RM-ANOVA to analyze longitudinal data:

- The RM-ANOVA model is really only applicable when the data are **complete** and the design is **balanced**.
- Because we are inputting time as a factor, the actual time when the repeated measures are taken does not appear explicitly in the model.
- Rarely is the assumptions of sphericity met in practice. It is only met in classroom examples.

Multivariate ANOVA (MANOVA)

MANOVA is the multivariate analog to ANOVA. The "multivariate" part of MANOVA indicates that there are more than one outcome variable.

MANOVA operates on the wide data where we have multiple columns of the repeated measures—multiple outcomes. In fact, each person will have a vector of outcomes (y_1, y_2, y_3, y_4).

In RM-ANOVA the way we accounted for the correlation in the repeated measures was to introduce a random-effect for subject. In MANOVA, since there are multiple outcome variables, we can actually model the variation in each Y and the correlation between the Y's. The advantage of MANOVA is that we don't assume an underlying structure to the covariation between the repeated measures (unstructured).

Fitting a MANOVA for repeated measures requires that we fit a multivariate intercept-only model. To do this, we first set up a matrix of the repeated measures using `cbind()`.

```
# Eliminate cases with missing data and make long format
> outcome = cbind(backpain$t1, backpain$t2, backpain$t3, backpain$t4)

# Fit the multivariate intercept-only model
> mod.mlm = lm(outcome ~ 1)
```

To actually produce the MANOVA, we will use the `Anova()` function from the **car** package. This function requires us to first set up a data frame that specifies the intra-subject design (names of the repeated measures).

```
# Set up data frame that specifies intra-subject design
> idata = data.frame(
    time = c("t1", "t2", "t3", "t4")
  )
```

Lastly, we specify this data frame using the `idata=` argument in `Anova()`. We also specify the `idesign=` argument by giving it a a one-sided model formula using the "data" in *idata* and specifying the intra-subject design.

```
# Fit the MANOVA
> manova1 = Anova(mod.mlm, idata = idata, idesign = ~time, type = 3)
```

We use the `summary()` function to output the results.

```
> summary(manova1)

Multivariate Tests: time
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.2057660 1.899882      3     22 0.15914
Wilks             1 0.7942340 1.899882      3     22 0.15914
Hotelling-Lawley  1 0.2590748 1.899882      3     22 0.15914
Roy               1 0.2590748 1.899882      3     22 0.15914
```

The part we care about is the multivariate tests for time. Each of the lines of output is a different multivariate test of whether there is an effect of time.

A multivariate analysis of variance was fitted to the data. The results of Wilk's test ($\lambda = 0.794$, $F(3, 22) = 1.90$, $p = 0.159$) was not statistically significant indicating no effect of time on pain.

We can also include between-subject factors. To do this we include treatment in the multivariate linear model.

```
# Fit the multivariate linear model
> mod.mlm2 = lm(cbind(t1, t2, t3, t4) ~ 1 + treatment, data = backpain)

# Fit the MANOVA
> manova2 = Anova(mod.mlm2, idata = idata, idesign = ~time, type = 3)
> summary(manova2)
```

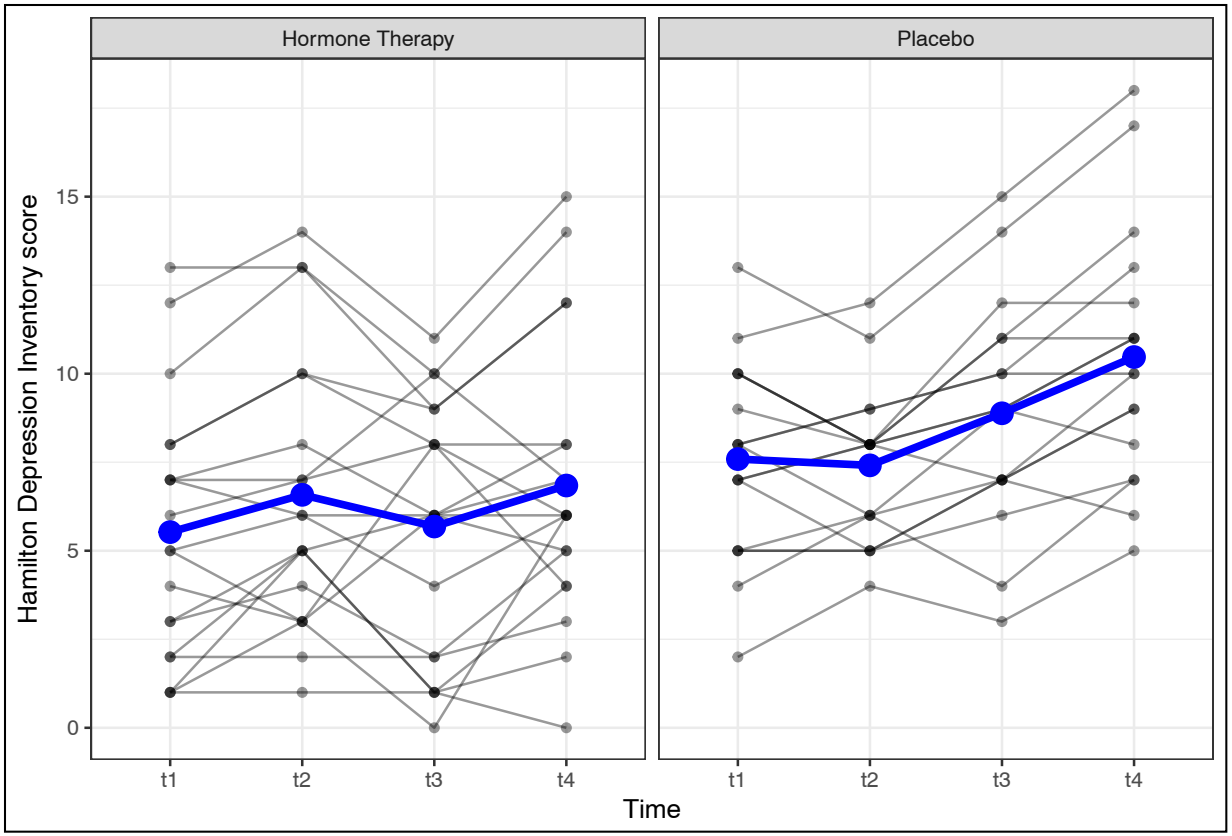
This will automatically output results for the main-effects of time and treatment, along with the interaction-effect of time x treatment.

```
Multivariate Tests: treatment:time
                 Df test stat  approx F num Df den Df  Pr(>F)
Wilks             1 0.9383759 0.4596973      3     21 0.71335

Multivariate Tests: time
                 Df test stat  approx F num Df den Df  Pr(>F)
Wilks             1 0.9636184 0.2642863      3     21 0.85033

Multivariate Tests: treatment
                 Df test stat   approx F num Df den Df Pr(>F)
Wilks             1 0.9964288 0.08243136      1     23 0.7766
```

The analysis finds no significant interaction nor main-effects.

This suggests that there are no mean differences in pain over time, nor between the control and treatment subjects.

Example: Forty women were enrolled in a study to investigate the effectiveness of hormone replacement therapy as treatment of depression. After randomization to hormone or placebo groups, a month baseline period was observed before the four month experiment began. The dependent variable is the Hamilton Depression Inventory, where low scores indicate less depression.



We will analyze these data using RM-ANOVA. Three major hypotheses are evaluated in the repeated measures ANOVA model are:

- No treatment-by-time interaction (non-parallel profiles in the population).
- No change over time (collapsing treatment).
- No treatment differences (collapsing time).

```
> ezANOVA(data = hormone_long, dv = hdi, wid = factor(id),
        within = time, between = condition, return_aov = TRUE)

$`Mauchly's Test for Sphericity`
          Effect         W         p p<.05
3           time 0.8398536 0.3355603
4 condition:time 0.8398536 0.3355603
```

Mauchly's test suggests that sphericity is tenable. Report and interpret the unadjusted tests.

```
$ANOVA
          Effect DFn DFd         F             p p<.05        ges
2      condition   1  34  5.124011 3.009365e-02     * 0.11710195
3           time   3 102 14.911287 4.038843e-08     * 0.04996494
4 condition:time   3 102  7.501274 1.380619e-04     * 0.02577536
```

Treatment-by-time interaction (non-parallel profiles in the population).
The first research question concerns whether there is a differential effect for the treatment group vs. control group over the four repeated measures of the experiment. The interaction null hypothesis is rejected $F(3, 102) = 7.50$, $p < 0.001$.

Change over time (collapsing treatment).
The second research question concerns whether collapsing across treatment, there is change over time. This hypothesis is also rejected $F(3, 102) = 14.91$, $p < 0.001$.

Treatment differences (collapsing time).
The third research question is the proposition of a treatment effect. This null hypothesis is also rejected, $F(1, 34) = 5.12$, $p = 0.03$.

Same dataset analyzed using MANOVA

```
> mod.mlm3 = lm(cbind(t1, t2, t3, t4) ~ 1 + condition, data = hormone)

> idata = data.frame(time = c("t1", "t2", "t3", "t4"))

> manova3 = Anova(mod.mlm3, idata = idata, idesign = ~time, type = 3)
> summary(manova3)

Multivariate Tests: condition:time
                 Df test stat approx F num Df den Df    Pr(>F)
Wilks             1 0.6132624  6.72665      3     32 0.0012011 **

Multivariate Tests: time
                 Df test stat approx F num Df den Df    Pr(>F)
Wilks             1 0.6791882 5.038357      3     32 0.0056805 **

Multivariate Tests: condition
                 Df test stat approx F num Df den Df   Pr(>F)
Wilks             1 0.8690316 5.124011      1     34 0.030094 *
```

Different analysis, similar results to the RM-ANOVA.

Comments regarding using MANOVA to analyze longitudinal data:

- The MANOVA model is really only applicable when the data are **complete** and the design is **balanced**.
- A practical issue with multivariate analysis is that many coefficients must be estimated in the covariance matrix. Because each parameter in the variance–covariance matrix for the repeated measures must be estimated, the hypothesis tests are somewhat less powerful than those for the univariate approach.
- Actual Measurements of time are not incorporated in the model.
- The focus of this analysis is on hypothesis tests. But, describing the way individuals change can be much more important than answering whether change has occurred. MANOVA does not easily address this.

References and Source Material

- Field, A. (1998). A bluffer's guide to ... sphericity. *The British Psychological Society: Mathematical, Statistical & Computing Section Newsletter, 6*, 13–22.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis.* New York: Wiley.
- Kohli, N. (2016). *Classical procedures for analyzing repeated measures data (course notes).* Minneapolis, MN: Author.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed). New York: Wiley.
- Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R.* Thousand Oaks, CA: Sage. (Minneapolis reading data)
- National Collegiate Athletic Association. (2016, December 12). *Graduation success rate.* Retrieved July 21, 2017, from http://www.ncaa.org/about/resources/research/graduation-success-rate
- Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.
- Graphics used from Open Clip Art Library (https://openclipart.org/)