EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

March 12, 2020

Last time

- Unbalanced ANOVA
- Restricted randomization and blocking to induce control and reduce confounding
- Repeated measures ANOVA

Today

- Repeated measures ANOVA
- Analysis of Covariance (ANCOVA)

Repeated measures

- When you have more than one observation on the *same* sample unit, the experiment is said to contain *repeated measures*.
- Ubiquitous in the health and social sciences.
- Classic example is measuring the effect of an intervention *pre* and *post* application. In this case, average treatment effect can be quantified with a (paired) t-statistic.
- But you may want to measure the effect of an intervention at *many* points in time over the *same* sample units. This suggests an ANOVA framework.
- A repeated measures design is a special case of a *nested* design.
- It is also a special case of a *blocked* design.

Assumptions of repeated measures ANOVA

The assumptions for a repeated measures ANOVA are a bit different:

- Independence of observations *between* subjects/factors only (obviously, observations *within* subjects are related).
- Equality of variances (homoskedasticity) over all levels of *between* subject factors.
- Normality assumption over all levels of *between* subject factors.
- Equality of variances and normality assumption *within* factors when *more* than two repeated measurements (time points): variances of the *differences* between all adjacent pairs of repeated measurements must be the same over all adjacent time points, and variances of the *differences* between all other possible pairs of repeated measurements must be the same over all possible pairs of time points, in addition to multivariate normality. This assumption is called *sphericity*.

Repeated measures ANOVA: example

- Assess student confidence in math abilities after participating in two weekend workshops.
- Students complete a questionaire to assess their math confidence levels before the workshops, after the first workshop, and after the second workshop. Confidence is measured on a 20-point scale, derived from a composite score from the questionaire.
- 8 students have not taken a math course in the past 5 years (L group), 8 students have taken a math course within the past 5 years, but not within the last year (M group), and 8 students have taken a math course within the last year (H group).

Repeated measures ANOVA: example

- RM-ANOVA shows evidence of an overall (marginal) treatment effect, and for a differential effect of treatment with Group.
- Note: marginal Group effect just reflects baseline differences between L/M/H Groups.

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	51.162	2	25.581	35.031	<.001
Time * Group	18.475	4	4.619	6.325	<.001
Residual	30.670	42	0.730		

Note. Type 3 Sums of Squares

[3]

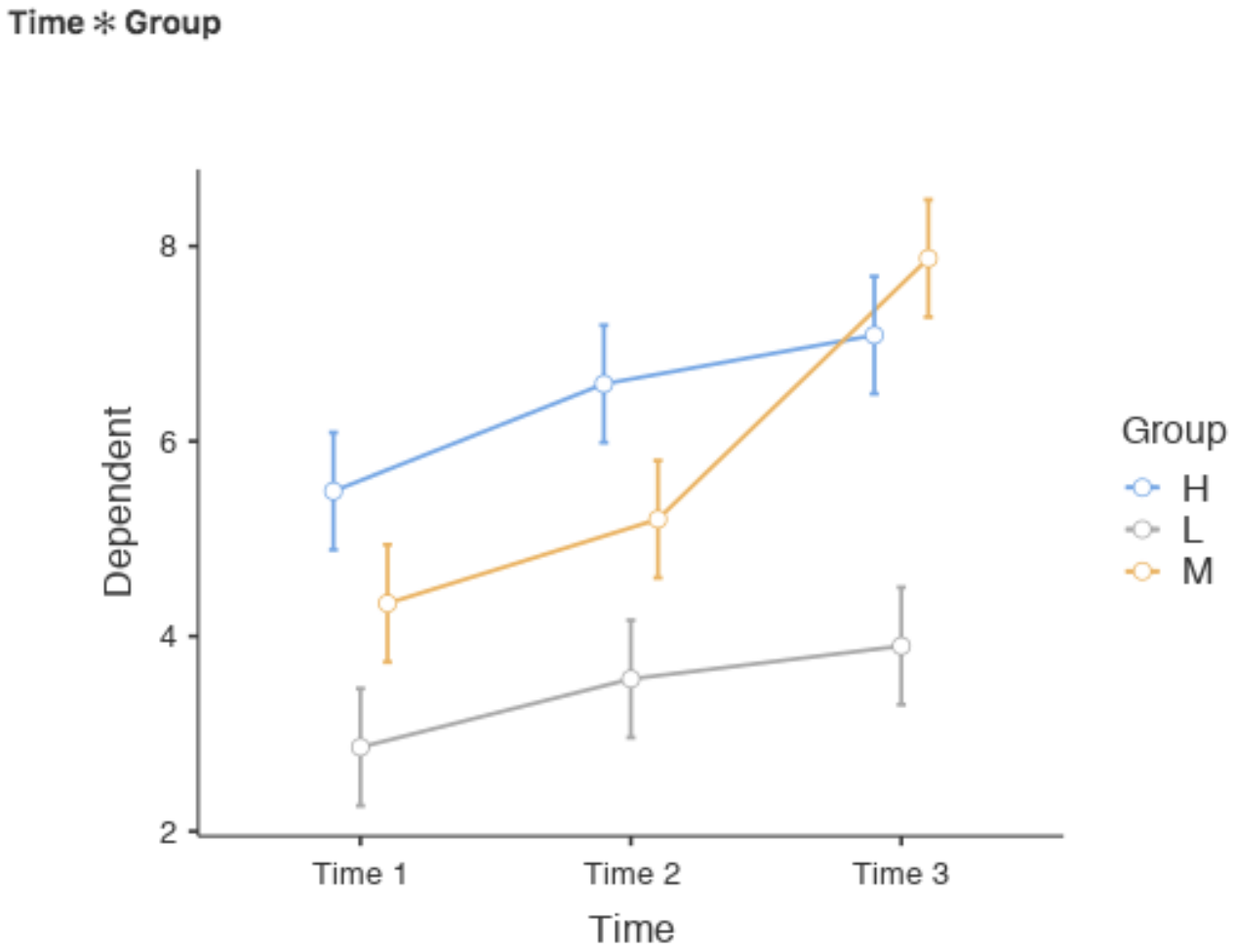
Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	116.797	2	58.398	81.826	<.001
Residual	14.987	21	0.714		

Note. Type 3 Sums of Squares

Repeated measures ANOVA: example

- Examining interaction plot shows where differential effect of treatment is present. Possible explanations for differential effect?



Repeated measures ANOVA: example

- Post-hocs on marginal effect of treatment show strong evidence for overall intervention effect and for effect of second workshop, but only moderate evidence for effect of first workshop.
- No evidence of violation of sphericity assumption.

Post Hoc Comparisons - Time

Comparison			Mean Difference	SE	df	t	p_{tukey}
Time		Time					
Time 1	-	Time 2	–0.887	0.247	42.000	–3.598	0.002
	-	Time 3	–2.058	0.247	42.000	–8.344	<.001
Time 2	-	Time 3	–1.171	0.247	42.000	–4.746	<.001

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Time	0.933	0.501	0.937	1.000

Fundamental problems with repeated measures ANOVA

Repeated measures ANOVA has been around for a long time (100+ years); thus, the methodology is ingrained in many fields. However, it suffers from several critical flaws:

- Repeated measures designs do not account for sequence or carryover effects.
- Repeated measures designs do not allow for patient drop-out.
- Repeated measures designs require the *sphericity* assumptions which is often extremely suspect in practice; moreover, RM-ANOVA is highly sensitive to violations of sphericity.

A few words on mixed effects models

While we do not have the time to treat these models properly, there is one important idea that we should note now.

- Recall we have only talked about "fixed effects" ANOVAs.
- An explanatory factor is called a *fixed effect* if its levels are either (1) fixed by the experimenter, or (2) exhausted by the experimental design.
- Alternatively, an explanatory factor is called a *random effect* if its levels are not fixed by the experimenter, but rather are *drawn randomly from a population of all possible factor levels*.
- *Mixed effects models* are simply statistical models (ANOVA or otherwise) that consider both fixed and random effects simultaneously.

A few words on mixed effects models

- The classic example of a *random effect* is a *randomly sampled* subject in a repeated measures design. Each sample's response at *baseline* can be considered a random effect.
- In this way, mixed effects modelling allows one to study the time effect *relative to each individual baseline*, which is assumed random.
- It turns out that this is a much more reasonable way to model repeated measures data: more flexible and more robust.
- Treating effects as random in an ANOVA framework *changes the F-statistic one should use to test for the presence of a nonzero effect* on the random term (no easy way to do this in Jamovi, but SPSS will handle such a model).

A few words on mixed effects models

- Mixed effects modelling can fix all the problems with RM-ANOVA (by proposing a different model and set of assumptions altogether).
- Mixed effects models allow you to explicitly study, quantify, and model *dependent or confounded data* in many different ways, e.g.
  - Accumulation effects of treatment in time or space.
  - Dispersion effects of treatment in time or space.
  - Other kinds of non-stationary treatment effects in time or space.
  - Drop-out effects.
  - Nonresponse bias.
  - Measurement error.
  - Preferential sampling.
  - And much, much more!

Practical repeated measures

- So if RM-ANOVA should be avoided, and we aren't learning about mixed effects modelling, then what should you do when you want to analyze repeated measures data?
- The problems with RM-ANOVA only really appear when we have *more than two time points* in our dataset.
- My advice: If you have more than two time points, just run *multiple* RM-ANOVAs on every *pair of time points* that you care about.
- Typical setup:
  - Measurements at time points 1, 2, and 3.
  - Care about possible changes in response from time point 1 to 2, and then from 2 to 3 (might also care about 1 to 3).
  - So perform two RM-ANOVAs on the two pairs of time points (1 to 2, and 2 to 3) and then adjust for the inflated Type I error rate (e.g. Bonferroni).

Repeated measures ANOVA: example

- Performing two RM-ANOVAs on each pair of time points yields same information as original analysis, without having to rely on the validity of the sphericity assumption.
- However, p-values not all the same (less power here).
- A bit of a multiple testing issue is present (though dependency of outcomes mitigates this concern somewhat).

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
RM Factor 1	9.452	1	9.452	17.231	<.001
RM Factor 1 ✻ Group	0.324	2	0.162	0.295	0.747
Residual	11.519	21	0.549		

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	64.065	2	32.033	63.856	<.001
Residual	10.534	21	0.502		

Note. Type 3 Sums of Squares

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
RM Factor 1	16.450	1	16.450	19.016	<.001
RM Factor 1 ✻ Group	13.628	2	6.814	7.877	0.003
Residual	18.167	21	0.865		

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Group	93.940	2	46.970	54.664	<.001
Residual	18.044	21	0.859		

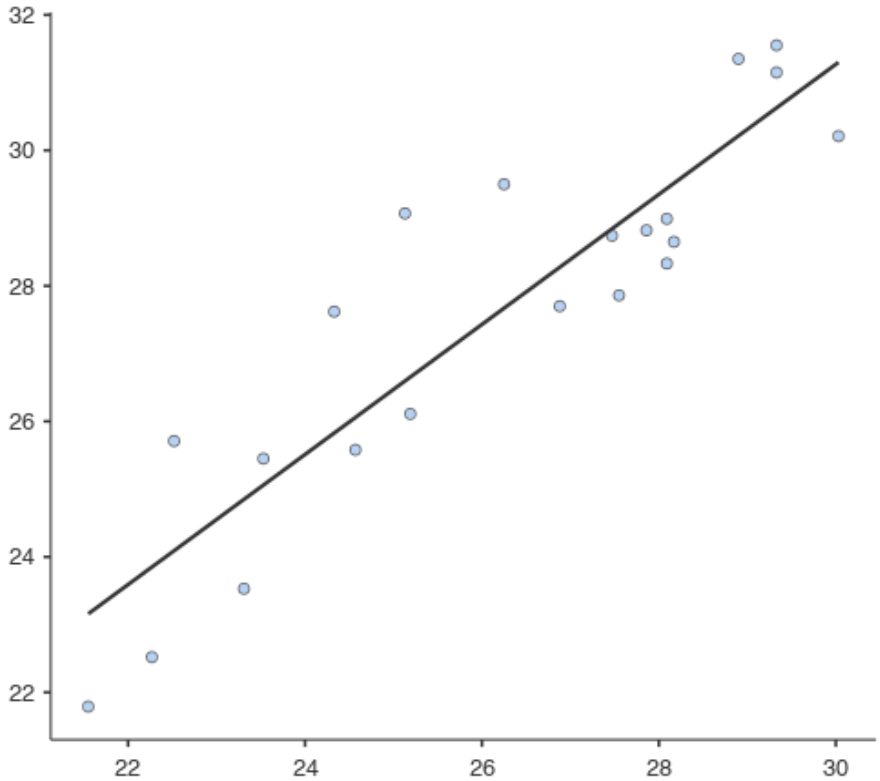
Note. Type 3 Sums of Squares

Analysis of Covariance (ANCOVA)

- ANOVA relates a *continuous response* of interest to a set of *categorical* explanatory variables.
- Analysis of Covariance (ANCOVA) extends the ANOVA framework to allow control for *continuous* explanatory variables as well.
- This is *NOT* the same thing as regression. In particular, ANCOVA does *not* allow you to estimate the *effect* of a continuous explanatory variable on a continuous response; it only *removes* the variation explained by the continuous explanatory variable, thus:
  - reducing residual error.
  - allowing better estimates of the categorical marginal and interaction effects of interest.
- In an ANCOVA, the continuous explanatory variable is *never* of interest. It is merely a *nuisance* variable to be eliminated.

Analysis of Covariance (ANCOVA) rationale

- Let Y_i be the response of interest for sample unit i. Let X_i be the covariate (continuous explanatory variable) for sample unit i
- First, find the "best fitting" line through the points (X_i, Y_i):



Analysis of Covariance (ANCOVA) rationale

- There are many ways to define "best fitting," but here we take the classical definition; i.e. the *ordinary least squares* (OLS) fitted line obtained by *minimizing the sum of the squared errors*.
- That is, if we write

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

for some random error $\varepsilon \sim N(0, \sigma^2)$, we can find numbers $\widehat{\beta}_0$ for β_0 and $\widehat{\beta}_1$ for β_1 that minimize

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- This is a simple calculus exercise and yields the OLS estimators:

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}, \qquad \widehat{\beta}_1 = \frac{S_{XY}}{S_X^2}$$

Analysis of Covariance (ANCOVA) rationale

- Now, with the "best fitting" (OLS regression) line estimated, we can plug in the OLS estimators and rearrange the equation:

$$\begin{aligned} Y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \varepsilon_i \\ &= \bar{Y} - \widehat{\beta}_1 \bar{X} + \widehat{\beta}_1 X_i + \varepsilon_i \\ &= \bar{Y} + \widehat{\beta}_1 (X_i - \bar{X}) + \varepsilon_i \end{aligned}$$

  Thus,

$$Y_i - \widehat{\beta}_1 (X_i - \bar{X}) = \bar{Y} + \varepsilon_i$$

- Denote the lefthand side of this equation by

$$Y_i^{adj} := Y_i - \widehat{\beta}_1 (X_i - \bar{X})$$

  This is our response of interest, Y, adjusted for the effect of the covariate X.

Analysis of Covariance (ANCOVA) rationale

- So, we now have a *transformed* version of Y that we can fit ANOVA models to. For example, if W is some categorical explanatory factor of interest for Y, we can now estimate the ANOVA model:

$$Y^{adj} = \mu + \tau_W + \delta$$

- This will give us information about the *effect of W on Y adjusted for the effect of X*.
- The classic (and most common) application: estimating the effect of some intervention Y *adjusting for baseline X* over groups of W.
- Note: we can adjust for *multiple covariates* by using the same "best fit" adjustment procedure for each covariate.

RM-ANOVA vs. ANCOVA

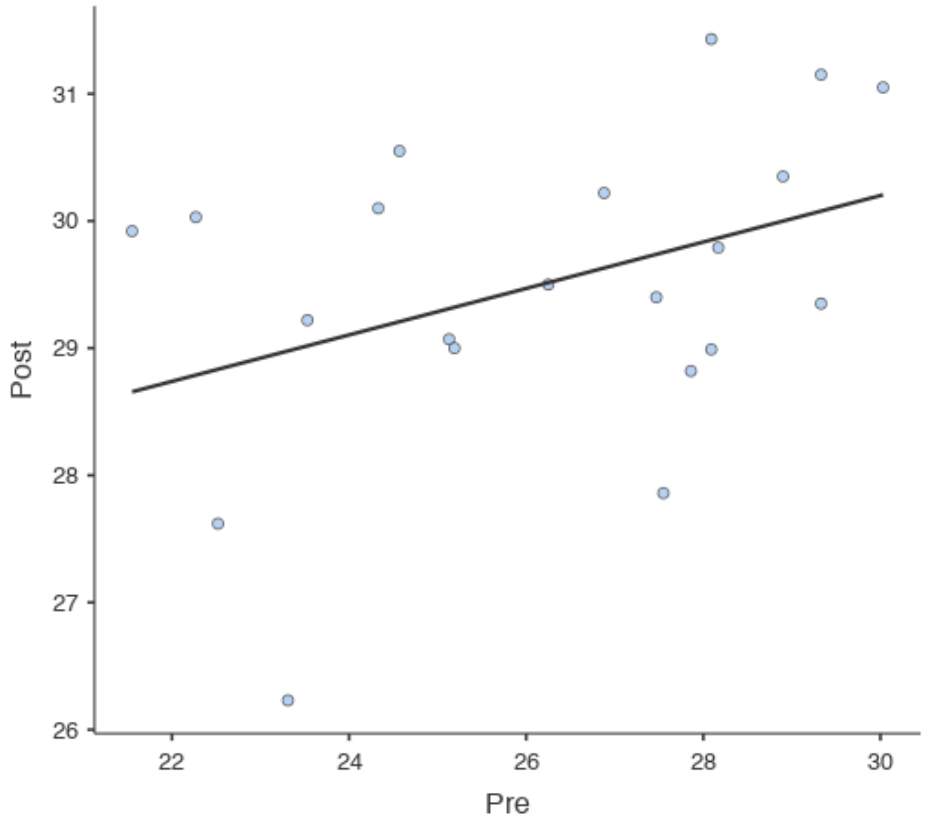
- Suppose we have a pre-test and post-test measurement on 21 people subjected to one of three experimental treatments (a *nested* design).
- Performing a RM-ANOVA, we could address the question of whether or not the average change in pre and post-test measurement differs among the three experimental groups.
- Or, treating the pre-test measurement as a nuisance variable, we can perform an ANCOVA to address the question of whether or not the average post-test measurement, adjusted for baseline differences in pre-test measurements, differs among the three experimental groups.
- ANCOVA quantifies *differences of post-test means* between groups (adjusted for baseline); RM-ANOVA quantifies *change from pre-test to post-test* between groups.

Assumptions of ANCOVA

- The usual ANOVA assumptions (independence, homoskedasticity, normality of residuals)
- Relationship between response and covariate is *linear*.
- All regression slopes between the covariate and the response are *equal* across each level of the explanatory factor(s).
- In an RM-ANCOVA framework, the regression slopes are also *equal* over each repeated measurement (virtually *never* satisfied in practice).
- Independence of the covariate and the other explanatory factors (often suspect).

ANCOVA Example 1 (covariate adjusting for baseline)

- Suppose we have a pre-test and post-test measurement on 21 people subjected to one of three experimental treatments (a *nested* design).
- We check if the pre-test baseline is linearly related to the post-test measurement:



ANCOVA Example 1

- There's somewhat of a linear relationship between our response of interest (post-test measurement) and nuisance covariate (pre-test measurement), so an ANCOVA approach may be reasonable.
- We estimate the improper ANCOVA model:

$$Y_{post} = \mu + \tau_{groups} + \beta \cdot Y_{pre} + \alpha \cdot \tau_{groups} \cdot Y_{Pre} + \delta$$

- Note: one of the assumptions of the ANCOVA rationale is that $\alpha = 0$. That is, all regression slopes between the covariate and the response are *equal* across all levels of the explanatory factor.
- By specifying the above model, we can explicitly *test* this assumption.
- **However, the improper model is NOT the model you should use to report your ANCOVA results.**

ANCOVA Example 1

In Jamovi:

- First create a column of data for each of: response of interest (Y_{post}), nuisance covariate (Y_{pre}), explanatory factor(s) (groups).
- Then select the 'ANCOVA' option from the 'ANOVA' analysis tab.
- Assign your dependent variable (response), fixed factors (explanatory factors), and covariates.
- In the 'Model' dialogue box, make sure a full two-way model is specified (with interaction).

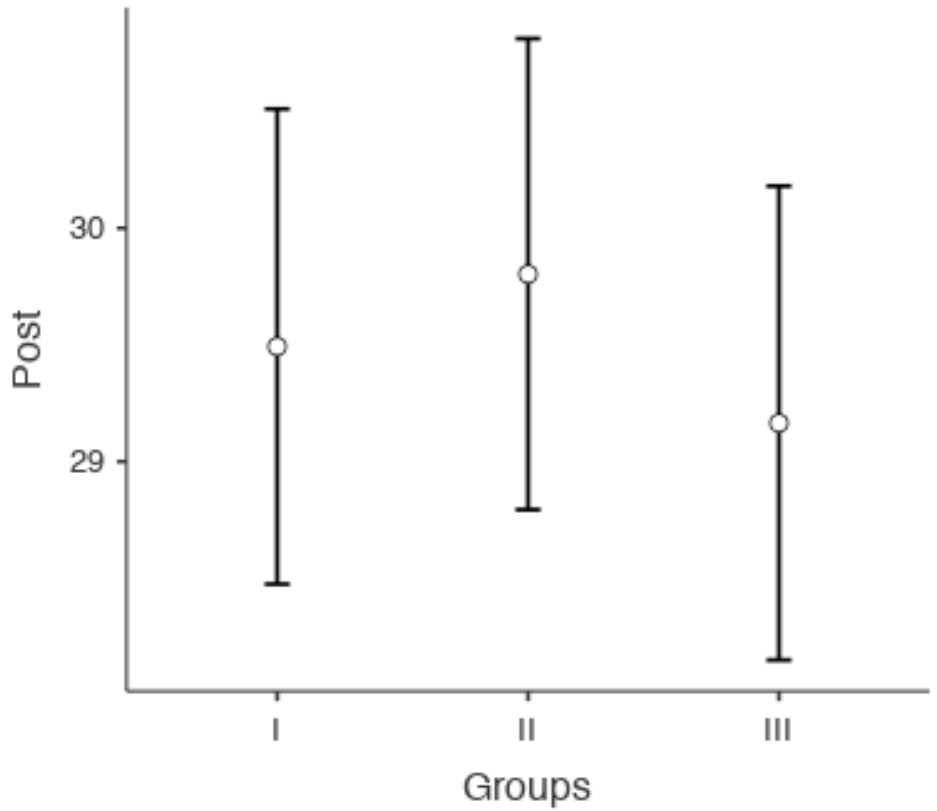
ANCOVA Example 1

	Sum of Squares	df	Mean Square	F	p
Pre	3.168	1	3.168	2.017	0.176
Groups	0.846	2	0.423	0.269	0.768
Groups ✻ Pre	0.935	2	0.467	0.298	0.747
Residuals	23.558	15	1.571		

- Notice: no significant effect of 'Groups $\times$ Pre-test'; so no evidence against ANCOVA assumption of equal regression slopes ($\alpha = 0$).
- Not much variation explained by baseline differences ('Pre' sum of squares).
- No evidence of a group effect on the post-test measurements (this is our main effect of interest).

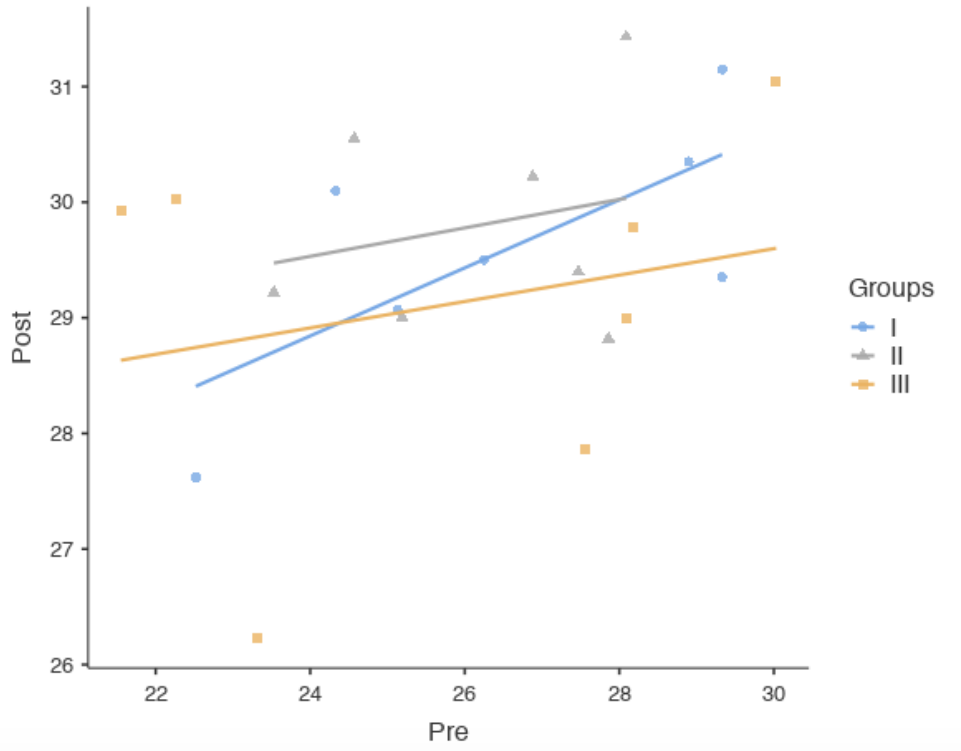
ANCOVA Example 1

- Estimates of the average post-treatment measurement between experimental groups: Group I average = 29.493 Group II average = 29.803, Group III average = 29.165.



ANCOVA Example 1

- Can plot regression lines by group easily with Jamovi's 'Exploration' → 'Scatterplot' option:



- Note: just because "best fit" lines cross, does not mean that we have evidence that they are different: there is a lot of uncertainty in the "best fit" estimates!

ANCOVA Example 1

- The $\alpha = 0$ assumption seems reasonable for our data.
- Thus, we can estimate the proper ANCOVA model:

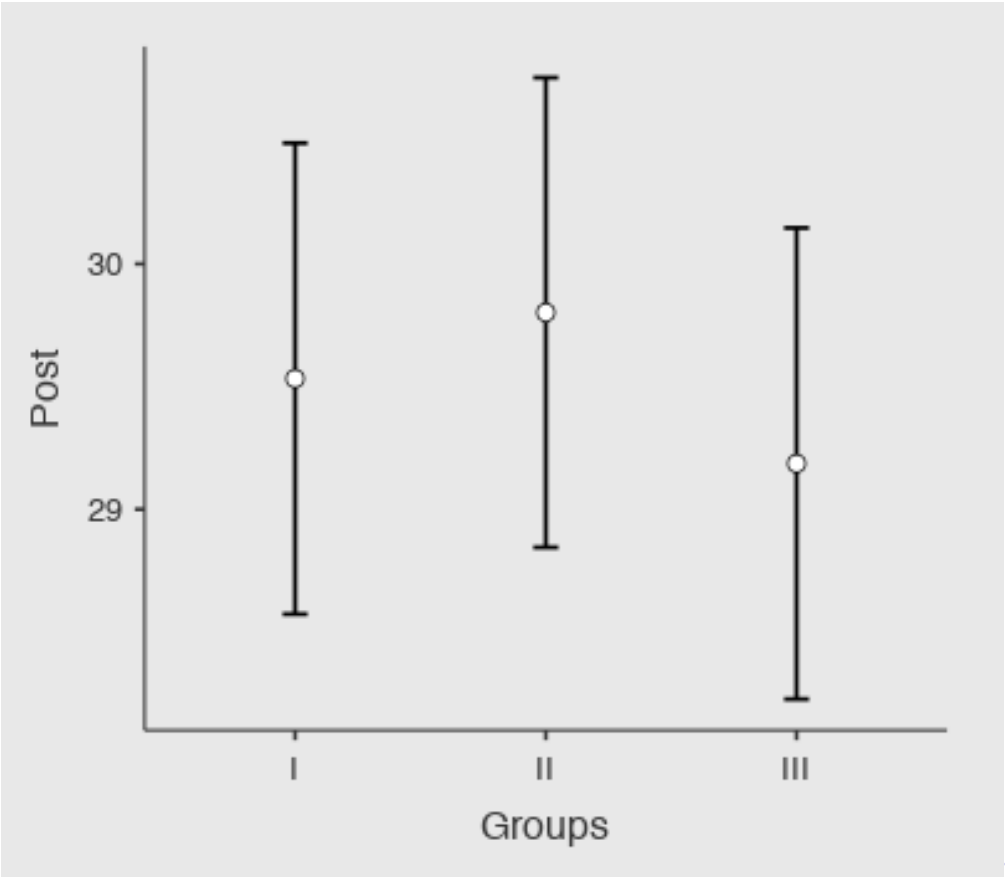
$$Y_{post} = \mu + \tau_{groups} + \beta \cdot Y_{pre} + \delta$$

ANCOVA

	Sum of Squares	df	Mean Square	F	p
Groups	1.327	2	0.664	0.461	0.639
Pre	4.087	1	4.087	2.836	0.110
Residuals	24.492	17	1.441		

ANCOVA Example 1

- Estimates of the average post-treatment measurement between experimental groups: Group I average = 29.533, Group II average = 29.802, Group III average = 29.187. **These are the effect sizes we should report.**



ANCOVA Example 1

- Now suppose we ran a RM-ANOVA on these data instead:

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Pre.vs.Post	114.345	1	114.345	36.074	<.001
Pre.vs.Post * Groups	0.493	2	0.247	0.078	0.925
Residual	57.055	18	3.170		

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Groups	2.870	2	1.435	0.250	0.782
Residual	103.503	18	5.750		

- Definite evidence for a change in time.
- No significant group effect, marginally or in time.
- Note: Post-hoc test on interaction would provide same info as the ANCOVA.

ANCOVA Example 2 (covariate adjusting for baseline)

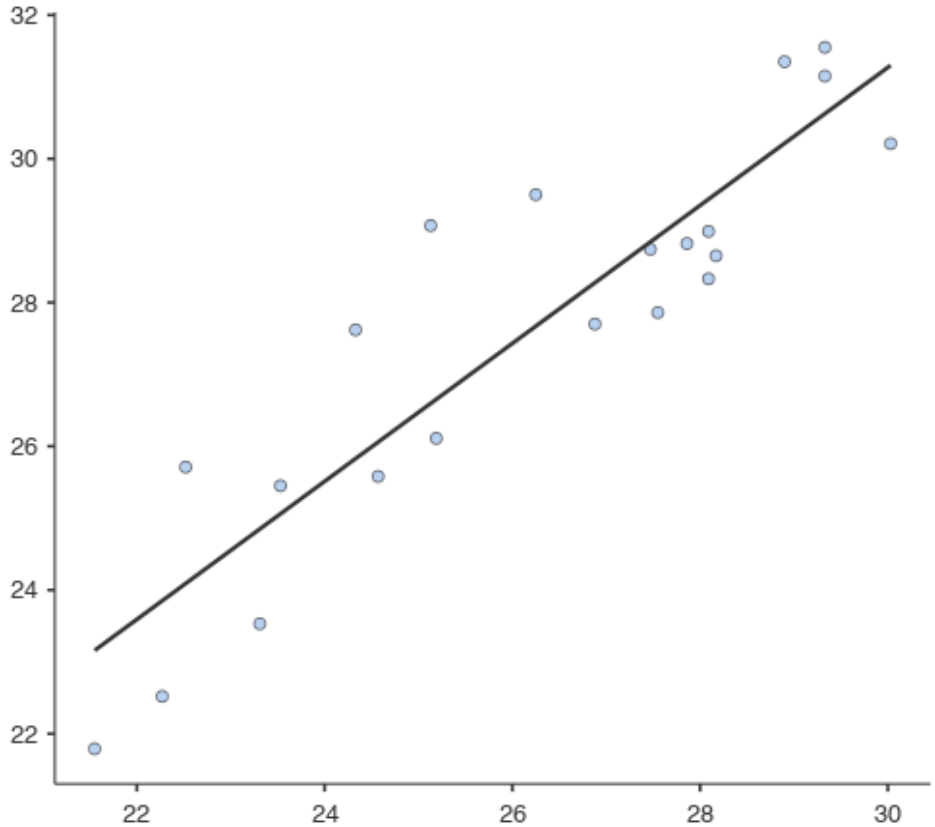
- Examine the difference between two exercise regimens vs. a control (no special training) on 21 people equally and randomly assigned to one of the three experimental groups. Our data look like:

Group	Subject	Measurement	Response
I	1	Pre	26.25
I	1	Post	29.50
I	2	Pre	24.33
I	2	Post	27.62
⋮	⋮	⋮	

- Will use ANCOVA to see if there are differences in the post-treatment measurements, controlling for baseline differences.

ANCOVA Example 2

- We will treat the pre-test measurement as our baseline measurement of physical fitness for each individual.
- In this case, baseline should be strongly correlated with the post-test measurement, which we can see explicitly if we graph 'Pre' vs. 'Post':



ANCOVA Example 2

- Due to this strong linear relationship between our response of interest (post-test measurement) and the nuisance covariate (pre-test measurement), an ANCOVA approach may be reasonable.
- We estimate the improper ANCOVA model:

$$Y_{post} = \mu + \tau_{groups} + \beta \cdot Y_{pre} + \alpha \cdot \tau_{groups} \cdot Y_{Pre} + \delta$$

- Note: we will again test if the ANCOVA assumption $\alpha = 0$ is reasonable.

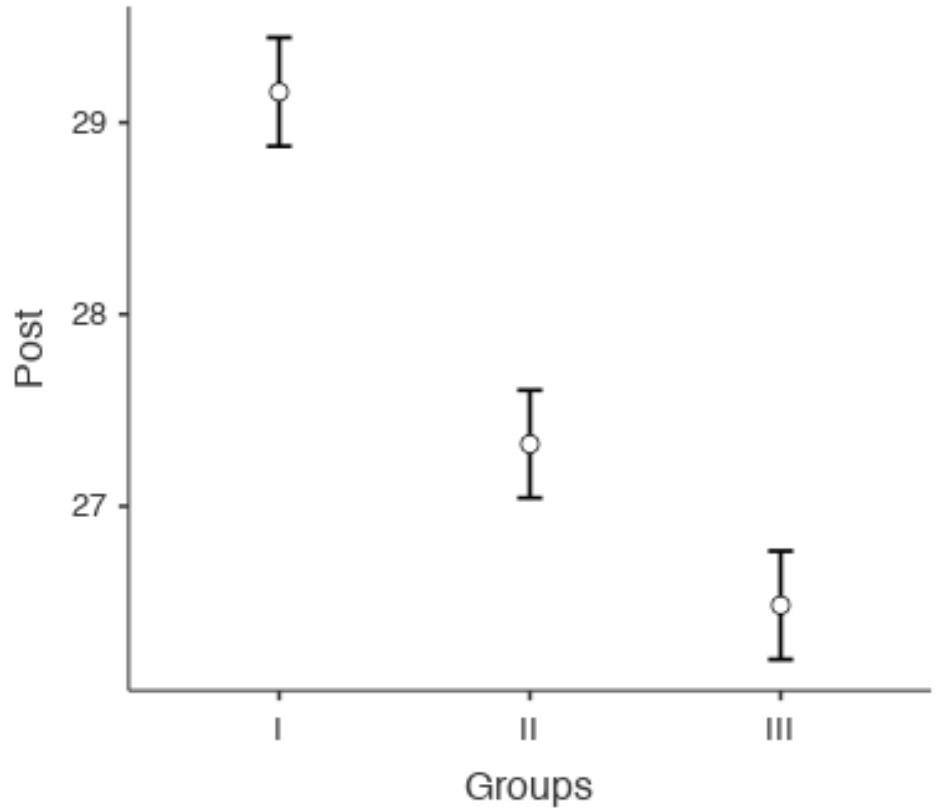
ANCOVA Example 2

	Sum of Squares	df	Mean Square	F	p
Pre	79.428	1	79.428	653.911	<.001
Groups	2.819	2	1.409	11.603	<.001
Groups ✻ Pre	1.368	2	0.684	5.631	0.015
Residuals	1.822	15	0.121		

- Lots of variation explained by baseline differences ('Pre' sum of squares).
- Also have evidence of a group effect on the post-test measurements (this is our main effect of interest).
- Also have weak evidence of a significant effect of 'Groups $\times$ Pre-test'; so the ANCOVA assumption of equal regression slopes ($\alpha = 0$) may be untenable.

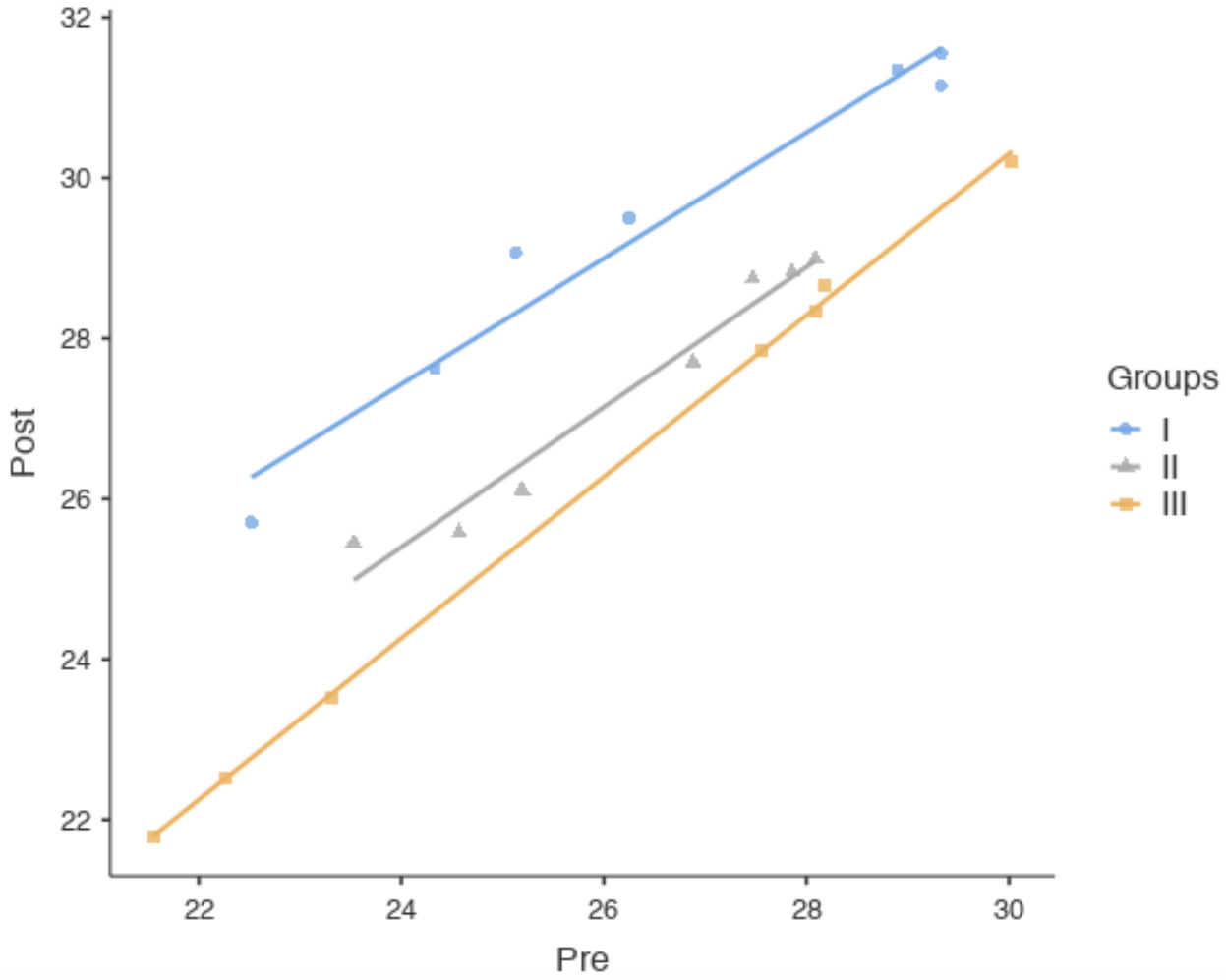
ANCOVA Example 2

- Estimates of the average post-treatment measurement between experimental groups: Group I average = 29.159, Group II average = 27.324, Group III average = 26.484.



ANCOVA Example 2

- "Best fit" lines for post-test (response) vs. pre-test (covariate) between groups:



ANCOVA Example 2

- We had evidence of possible heterogeneity of regression slopes for post-test (response) vs. pre-test (covariate) between groups: $F(2, 15) = 5.631$, p-value $= 0.015$.
- Notice in the plot: these lines look very close to parallel! But because the (Pre,Post) data (by group) fall so close to each line, we have *little residual variability*. This is reflected in the very small SS(residuals) in the ANCOVA.
- Thus, we have *high power* to detect small differences between the slopes. The question now is are these obviously small differences meaningful enough for us to distrust the ANCOVA?
- This is a judgment call in general, but here, the slopes are so close that the results of the ANCOVA should not be greatly affected by assuming $\alpha = 0$.

ANCOVA Example 2

- We can verify this by comparing our results with the proper ANCOVA model:

$$Y_{post} = \mu + \tau_{groups} + \beta \cdot Y_{pre} + \delta$$

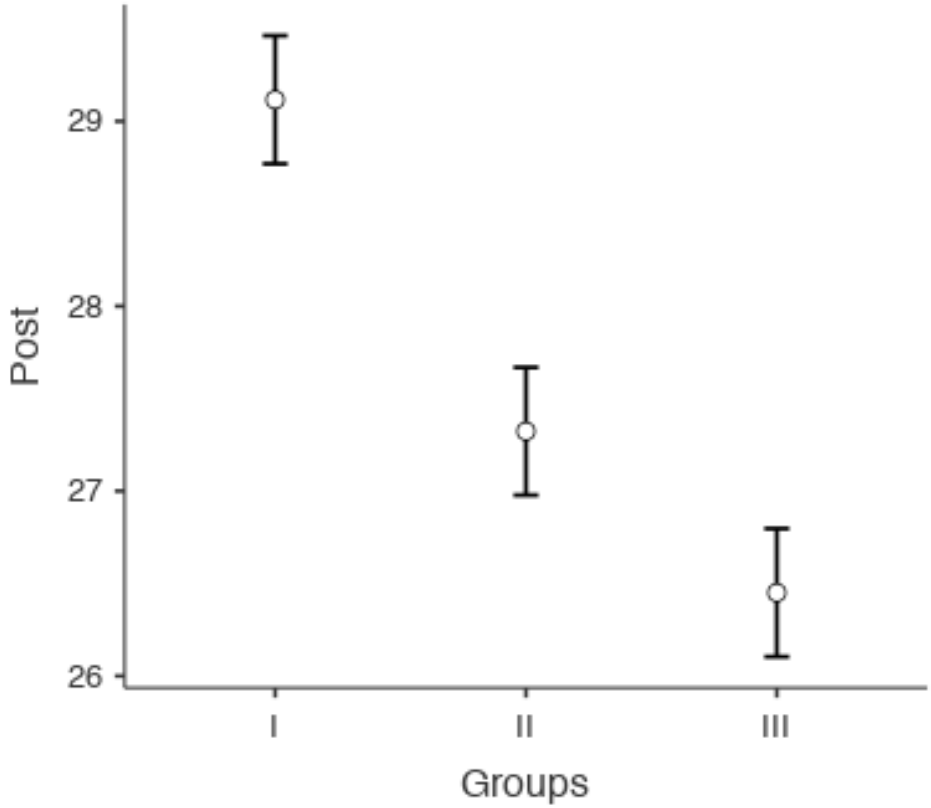
ANCOVA

	Sum of Squares	df	Mean Square	F	p
Groups	25.564	2	12.782	68.117	<.001
Pre	109.840	1	109.840	585.351	<.001
Residuals	3.190	17	0.188		

- Still have significant Groups and Baseline effects.

ANCOVA Example 2

- Estimates of the average post-treatment measurement between experimental groups: Group I average = 29.116, Group II average = 27.323, Group III average = 26.450. **Again, these are the results one should report.**



ANCOVA Example 2

- Compare to a RM-ANOVA:

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Pre.vs.Post	21.257	1	21.257	181.967	<.001
Pre.vs.Post ✻ Groups	12.382	2	6.191	52.996	<.001
Residual	2.103	18	0.117		

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Groups	28.139	2	14.070	1.043	0.373
Residual	242.906	18	13.495		

- Definite evidence for a marginal change in time.
- Significant group interaction with time, but no marginal group effect.
- Again, note that post hoc tests on interaction term would provide same info as ANCOVA.