Multiple Regression 2 – Enhancing the Multiple Regression Model

1. Polynomial terms ← review 511
2. Interaction terms ← stay for rest of life
3. Analysis of Covariance (ANCOVA) – 1 continuous predictor
   – 1 categorical (2 levels)
4. Misc R Topics

<u>Examples:</u>

1. Steel Example ← polynomial
2. Process Example ← interaction
3. Tool Example – might skip
4. Meadow Foam ← ANCOVA
5. Glue Strength

1. Polynomial Terms

Lack of linearity in the scatter plot may be addressed by expanding the model to allow for curvature. One method to do this is to add polynomial terms:

quadratic term

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

seen in 511

Steel Example: An experiment was conducted to examine the relationship between Strength (Y) and coating Thickness (X) in steel. Scatter plot shows strong curvature and plot of residuals vs fitted values also shows curvature. Hence simple linear regression is not appropriate. A quadratic regression seems to fit the data well.

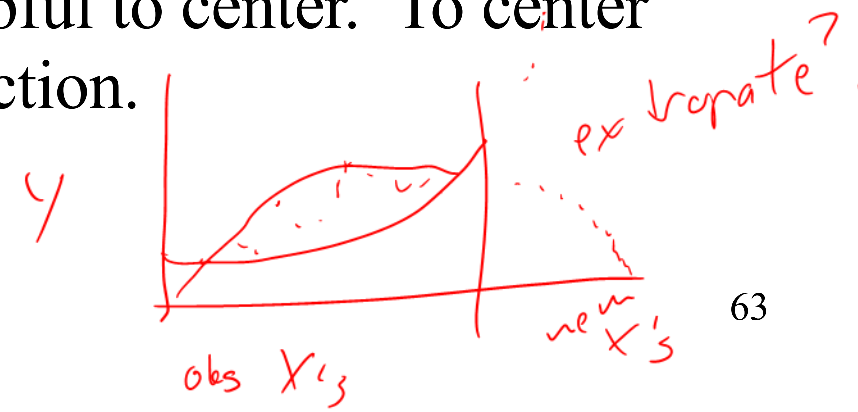
Comments about polynomial models:

1. Selected to approximate the data. Seldom are the parameters of any physical significance.

2. Extrapolation beyond the range of the collected X values is even more risky than for linear regression.

3. X^3, X^4, X^5, etc. terms can also be added, up to the number of unique levels of X (minus 1). In practice, it is unusual to see more than X and X^2.

4. It sometimes helps to center X (by subtracting off the mean) to reduce correlation between X and X^2. Also if the values of X are very large, it can be helpful to center. To center data in R, use the `scale()` function.



5. In general, polynomial regression models should obey the "**Principle of Hierarchy**", which says that if your model includes X^2 you should keep X in the model even if it is no longer statistically significant.

This idea can be extended to higher order polynomial regression. Specifically, if X^p is statistically significant and included in your model, you should also include X^j for $j < p$, whether or not coefficients for these lower order terms are statistically significant.

2. Interaction Terms

Process Example: The response is the yield (Y) of a certain process, the predictors are temperature (X_1 = 130 or 150) and concentration (X_2 = 1 or 2). This is a 2x2 factorial. Two observations per treatment for a total of n = 8 obs.

2-way ANOVA

Consider the "**no interaction**" model:

$$\text{Model3}: \hat{y} = 8.75 + 13.5 * \text{conc} + 0.30 * \text{temp}$$

If conc = 1, then

$$\hat{y} = 8.75 + 13.5 * 1 + 0.30 * \text{temp}$$

$$\hat{y} = 22.25 + 0.30 * \text{temp}$$

If conc = 2, then

$$\hat{y} = 8.75 + 13.5 * 2 + 0.30 * \text{temp}$$

$$\hat{y} = 35.75 + 0.30 * \text{temp}$$

Interpretation of the "no interaction" model:

Changing conc from 1 to 2 changes the height of the line, but the slope stays the same. (Changing conc from 1 to 2 increases the process yield, but the response to temp when conc is 1 is the same as the response to temp when conc is 2.)

What if changing conc may change the response to temp?
Solution: Add an "interaction" term:

$$\text{Model5: } \hat{y} = 82.25 - 35.5 * \text{conc} - 0.225 * \text{temp} + 0.35 * \text{temp} * \text{conc}$$

Interpretation of the Process "interaction" model:

If conc = 1, then:

$$\hat{y} = 82.25 - 35.5*1 - 0.225*\text{temp} + 0.35*\text{temp}*1$$

$$\hat{y} = 46.75 + (-0.225 + 0.35)*\text{temp}$$

$$\hat{y} = 46.75 + 0.125*\text{temp}$$

If conc = 2, then:

$$\hat{y} = 82.25 - 35.5*2 - 0.225*\text{temp} + 0.35*\text{temp}*2$$

$$\hat{y} = 11.25 + (-0.225 + 0.7)*\text{temp}$$

$$\hat{y} = 11.25 + 0.475*\text{temp}$$

If conc = 1.5, then:

$$\hat{y} = 82.25 - 35.5*1.5 - 0.225*\text{temp} + 0.35*\text{temp}*1.5$$

$$\hat{y} = 29.0 + (-0.225 + 0.525)*\text{temp}$$

$$\hat{y} = 29.0 + 0.30*\text{temp}$$

The slope with respect to temp depends on the value of conc.

or response

Definition of Interaction: An interaction between two variables indicates that the effect of one variable depends on the level of the other variable.

For an interaction between **two continuous predictors** (say X_1 and X_2), this means that the slope corresponding to X_1 depends on the value of X_2 (or vice versa). – this

Hence an interaction allows for non-parallel lines.

Interaction is different from correlation between two variables. For the Process data, there is a significant interaction between temp and conc. Because this is a designed experiment, the correlation between temp and conc is zero!

Notes about the model with interaction:

1. For any value of conc, we can get a line, representing the response to temp, at that conc.

2. For any value of temp, we can get a line representing the response to conc, at that temp.

3. A good way to interpret the interaction is by graphing. The multiple regression model may be presented as plot Y by X1, with multiple lines for various values of X2, or vice versa. We could also use a contour plot (see Tool example).

4. Since there are 4 parameters (β_0, β_1, β_2, β_3) and 4 unique combinations of the X variables (2 temps x 2 concs), the predicted values match the four means exactly. The model is "**saturated**".

imagine 3 cont predictors

5. In general, models with interactions should obey the "**Principle of Hierarchy**", which says that if your model includes the interaction between X1*X2, should keep X1 and X2 in the model even if they are no longer statistically significant. JMP won't allow interaction w/o main effects

6. All of the estimated responses for intermediate values of temp and conc depend on the assumption that the model (with interaction) is correct. Another approach would be to treat conc and temp as categorical and run a 2way ANOVA.

7. Without more data we cannot check the adequacy of the model (e.g., with only two values of temp, there is no way to check for a quadratic term). These designs are "exploratory". Can be used for planning new studies.

Tool Example: An example with quadratic and interaction terms

skip

Y = tool life (coded)

X_1 = cutting Angle (15, 20, and 25)
X_2 = cutting Speed (125, 150 and 175)

Nine treatments with two observations per treatment for a total of n=18 obs
Later we will call this a "three by three" or 3^2, factorial.

$$\text{Full model}: \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3 x_2 + \hat{\beta}_4 x_2^2 + \hat{\beta}_5 x_1 x_2 + \hat{\beta}_6 x_1 x_2^2 + \hat{\beta}_7 x_1^2 x_2 + \hat{\beta}_8 x_1^2 x_2^2$$

"Full model" has linear and quadratic terms for both variables plus all interactions.

We use the full model because the highest order interaction terms are significant.

$$\hat{y} = -1068 + 136.3x_1 - 4.08x_1^2 + 14.48x_2 - 0.0496x_2^2$$
$$-1.864x_1x_2 + 0.0064x_1x_2^2 + 0.056x_1^2x_2 - 0.00192x_1^2x_2^2$$

We plot the "response surface" and response surface contours.

Notes:

1. The full model is saturated because it has **nine** β parameters and there are **nine** unique trt combinations. Hence, at the "design points" the predicted value equals the sample mean.

2. Some model selection methods might simplify the model slightly.

3. Not a good idea to extrapolate beyond the range of the X-values. But this can be a concern even for simpler models.

start here 2/19

3. Analysis of Covariance (ANCOVA)

Some perspective:

- ANCOVA is just another example of the general linear model.
- Continuous Response with 1 Continuous Predictor → Simple Linear Regression (STAT511)

  Goal: Inference about slope

- Continuous Response with 1 Categorical Predictor → one-way ANOVA or 2-sample t-test (STAT511)

  Goal: Compare means

- Continuous Response with 1 Continuous and 1 Categorical Predictor → ANCOVA (Analysis of Covariance)

  Multiple goals depending on research questions!

not of primary interest

A "**covariate**" is a "secondary" predictor that may be related to the response variable. A covariate can be categorical or continuous.

Candidates for covariates in a given study depend on the particular experiment. The covariate should not be affected by the treatment. In most cases, the covariate is measured on the experimental unit before the treatment is applied to the unit. Examples include: initial soil fertility before applying fertilizer or age of a human (or animal) subject.

The ANCOVA analysis can be used whether the continuous predictor is of direct interest or just a covariate.

Builds on Process example

Meadowfoam Example (Ramsey and Schafer)

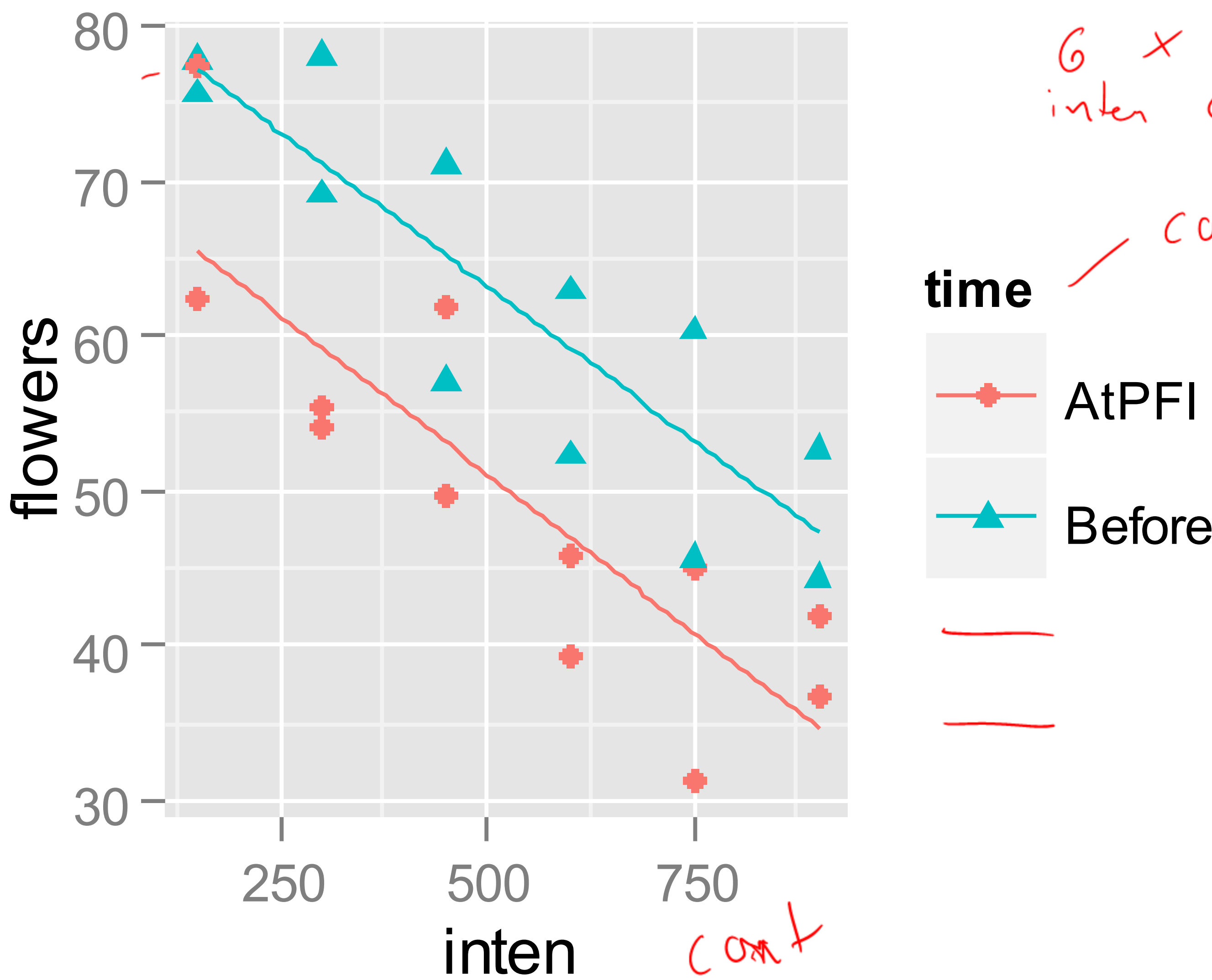
Meadowfoam is a small plant that grows in moist areas of the Pacific Northwest. Its seed oil has desirable qualities and potential commercial value. Six light intensity levels (150, 300, 450, 600, 750 and 900) and two timings of the onset of light treatment (at Photoperiodic Floral Induction-PFI, or 24 days before PFI) were considered in an experiment that had 2 replications of each intensity and time combination, for a total of n=24 observations.

response

Continuous Response: Y = **flowers** per plant cont

Continuous Predictor: X = Light **inten**sity level

Categorical Predictor: **time** (BeforePFI or AtPFI) cat



6 × 2 × 2
inten cat 24 obj

cat

cont

Some ANCOVA Analysis Goals:

Remember, the specific analysis will depend on individual research questions!

maybe not

- Estimate intercept for each trt/group.
- Estimate slope for each trt/group.
- Test for differences between intercepts - probably
- Test for differences between slopes (With Interaction only)
- Test whether common slope (No Interaction) or individual slopes (With Interaction) are different from zero.
- Test for difference between mean response at fixed value of the continuous predictor. In other words, test for a difference between means (for the categorical predictor) controlling for the continuous predictor.

Recall the Definition of Interaction: An interaction between two variables indicates that the effect of one variable depends on the level of the other variable.

In the ANCOVA setting we are interested in an interaction between **a continuous and categorical predictor**, the interaction means that the slope depends on the group.

A model WITH interaction allows the slopes to be different for each group. In other words, the interaction allows for non-parallel lines.

A model WITHOUT interaction forces the slopes for all the groups to be the same. In other words, the no interaction model forces parallel lines.

Our plan for the analysis of the Meadowfoam example:

We will compare **two possible models**:

1. ANCOVA NO interaction: No interaction forces the slopes for both groups to be the same.
2. ANCOVA WITH interaction: Including the interaction allows the slopes to be different for each group.

For the Interaction model, we will also consider **two different parameterizations**. We will see that the standard parameterization makes some testing easier. While the alternate parameterization makes some estimation easier.

Important Note:

Except for β_0 (intercept) the subscripting of the β's is arbitrary.

The order of the coefficients/parameter estimates matches the order of the variables in the model statement (which you write)!

In the following discussion, pay attention to which parameters correspond to categorical, continuous and interaction variables.

Recall that when a factor variable (categorical predictor) is included in the model, R creates indicator variables omitting the indicator for the first level of the factor. The ordering of the levels can be checked using `levels()`.

ANCOVA NO Interaction

ANCOVA with No Interaction allows <u>different intercepts</u> but <u>same slope</u> for trt groups.

indicator

cont

$$y_i = \beta_0 + \beta_1 \mathrm{I(time.AtPFI)}_i + \beta_2 \mathrm{x}_i + \varepsilon_i$$

y_i = #flowers/plant for the i^{th} observation

x_i = light intensity for the i^{th} observation

$\mathrm{I(time.AtPFI)}_i$ is an indicator variable:

$\mathrm{I(time.AtPFI)}_i = 1$ when time = "AtPFI"

$\mathrm{I(time.AtPFI)}_i = 0$ when time = "Before"

ordering

Fitting ANCOVA No Interaction

```
Model1 <- lm(flowers ~ time + inten, data
   = Meadow)
summary(Model1)
```

$\hat{\beta}$'s

cat

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.464167   3.273772  25.495  < 2e-16
timeAtPFI   -12.158333   2.629557  -4.624 0.000146
inten        -0.040471   0.005132  -7.886 1.04e-07
```

conf

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 83.46 \\ -12.16 \\ -0.0405 \end{pmatrix}$$

"common" slope

Parameter Interpretation ANCOVA No Interaction Model

Hint: write an equation for a predicted value in each group.

For data in the "AtPFI" group ($I(\text{time.AtPFI})_i = 1$):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 1 + \hat{\beta}_2 * x_i$$

intercept

this slope for AtPFI

$$\hat{y}_i = 83.46 - 12.16 - 0.0405 * x_i = 71.3 - 0.0405 * x_i$$

For data in the "Before" group ($I(\text{time.AtPFI})_i = 0$): Before

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 * x_i$$

slope for Before

$$\hat{y}_i = \underline{83.46} - 0.0405 * x_i$$

intercept

Interpret the parameters (simplest eqn. first):

$\hat{\beta}_0 = 83.46 =$ intercept for Before group

$\hat{\beta}_0 + \hat{\beta}_1 = 83.46 - 12.16 = 71.3$

$=$ intercept for AtPFI group

$\hat{\beta}_1 = -12.16 =$ difference between intercepts

$\hat{\beta}_2 = -0.0405 =$ common slope for both groups

In this case, with just 2 groups, the tests from the `summary()` output are equivalent to the `Anova()` tests, with $F = t^2$. The F-tests extend to more than two groups.

```
summary(Model1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.464167   3.273772  25.495  < 2e-16
timeAtPFI   -12.158333   2.629557  -4.624 0.000146
inten        -0.040471   0.005132  -7.886 1.04e-07

Anova(Model1, type = 3)

            Sum Sq Df F value    Pr(>F)
time         887.0  1  21.379 0.0001464 ***
inten       2579.8  1  62.181 1.037e-07 ***
Residuals    871.2 21
```

cat

cont

how close is close for parallel lines

cat

cont

or

H_0: common slope $= 0$

Hypothesis Testing ANCOVA No Interaction Model

cont

1. **Is there an effect of intensity? Or, Is the common <u>slope</u> different from <u>zero</u>?** Use the "inten" line in the ANOVA table (or test the "inten" slope parameter).

   H_0: $\beta_2 = 0$ vs H_A: $\beta_2 \neq 0$
   $F = 62.18$ (or $t=-7.89$), $p<0.0001$
   Reject H0, strong evidence that the slope is not zero.

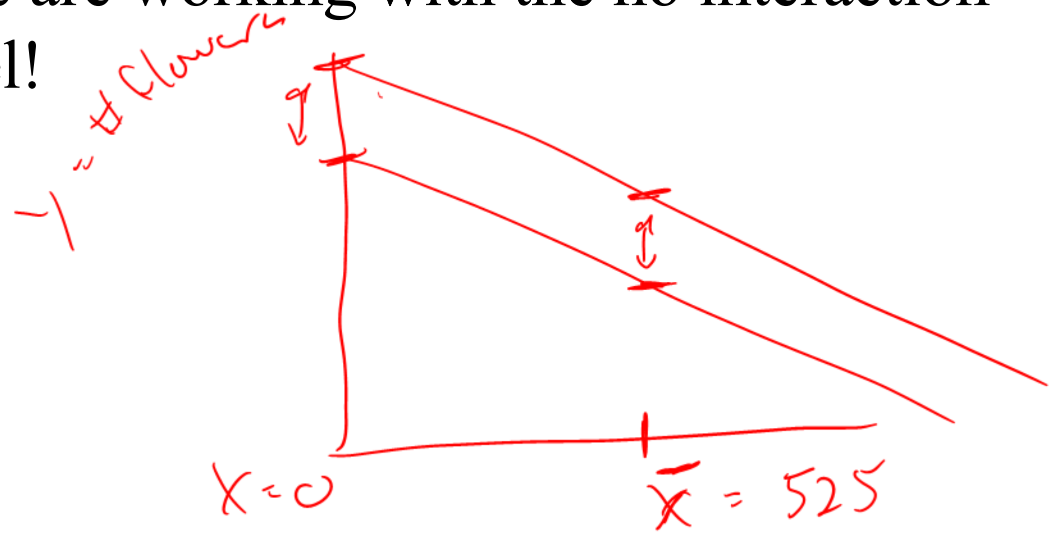
2. **Is there an effect of timing? Or, is there a <u>difference</u> between the <u>intercepts</u>?** Use the "time" line in the ANOVA table (or test the "time.ATPFI" indicator variable).

   H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$
   $F = 21.38$ (or $t=-4.62$), $p<0.0001$
   Reject H0, strong evidence that the intercepts are different.

only two cats

3. Is there an effect of timing? Is there a <u>difference</u> between <u>mean response</u> at fixed value of the inten?
Use `emmeans()` to test for a difference between the mean response for Before vs AtPFI groups when inten = 525 (average intensity value).
$t = 4.62$, $p<0.0001$
Reject H0, conclude there is a difference.

Note that the test statistic is the same for the emmeans pairwise comparisons as for the test of intercepts (inten = 0). That is because we are working with the no interaction (common slope) model!



C M

Expected Marginal Means (emmeans aka lsmeans)

Sometimes called "adjusted" means, because the means are "adjusted" to as if the two groups had the same mean on the X variable.

Predicted response (mean) at the average value of the continuous predictor. Calculated by putting parameter estimates into the model with X equal to its overall mean value. In this case, overall average of inten=525.

cat

In R: `emmeans(Model1, pairwise ~ time)`

$$\text{emmean(AtPFI)} = \hat{\beta}_0 + \hat{\beta}_1 * I(\text{time.AtPFI}) + \hat{\beta}_2(\bar{\text{x}})$$

$$= 83.46\text{-}12.15\text{-}0.0405(525)$$

$$= 50.06$$

$$\text{emmean(Before)} = \hat{\beta}_0 + \hat{\beta}_2(\bar{\text{x}})$$

$$= 83.46\text{-}0.0405(525)$$

$$= 62.21$$

62.21 - 50.06
= 12.16

ANCOVA WITH Interaction

for illustration

ANCOVA With Interaction allows <u>different intercepts</u> and <u>different slopes</u> for trt groups.

indicator cont indicator

$$y_i = \beta_0 + \beta_1 \mathrm{I(time.AtPFI)}_i + \beta_2 x_i + \beta_3 x_i \mathrm{I(time.AtPFI)}_i + \varepsilon_i$$

y_i = #flowers/plant for the i$^{\text{th}}$ observation

x_i = light intensity for the i$^{\text{th}}$ observation

$\mathrm{I(time.AtPFI)}_i = 1$ when time = "AtPFI"

$\mathrm{I(time.AtPFI)}_i = 0$ when time = "Before"

$x_i\mathrm{I(time.AtPFI)}_i = x_i$ when time = "AtPFI"

$x_i\mathrm{I(time.AtPFI)}_i = 0$ when time = "Before"

Fitting ANCOVA With Interaction

```
Model2 <- lm(flowers ~ time*inten, data =
  Meadow)

summary(Model2)
```

short-hand for "full" model + interaction

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     83.146667   4.343305  19.144 2.49e-14
timeAtPFI      -11.523333   6.142360  -1.876   0.0753
inten           -0.039867   0.007435  -5.362 3.01e-05
timeAtPFI:inten -0.001210   0.010515  -0.115   0.9096
```

cat

cont

interact

partial regression coeff for int

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 83.15 \\ -11.52 \\ -0.039 \\ -0.0012 \end{pmatrix}$$

Parameter/Coef Interpretation ANCOVA With Interaction

Hint: write an equation for a predicted value in each group.

For the "AtPFI" group ($I(\text{time.AtPFI})_i = 1$):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 1 + \hat{\beta}_2 * \text{Inten}_i + \hat{\beta}_3 * \text{Inten}_i * 1$$

$$\hat{y}_i = \underbrace{83.15 - 11.52}_{} - \underbrace{0.0399}_{} * \text{Inten}_i \underbrace{-0.0012}_{} * \text{Inten}_i$$

$$\hat{y}_i = \underbrace{71.63}_{\text{intercept of AtPFI}} - \underbrace{0.0411}_{\text{slope of AtPFI}} * \text{Inten}_i$$

For the "Before" group ($I(\text{time.AtPFI})_i = 0$):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 * \text{Inten}_i + \hat{\beta}_3 * \text{Inten}_i * 0$$

$$\hat{y}_i = \underbrace{83.15}_{\text{intercept for Before}} - \underbrace{0.0399}_{\text{slope for Before}} * \text{Inten}_i$$

different slopes (but not that different)

Parameter/Coef Interpretation ANCOVA With Interaction

R label

intercept $\hat{\beta}_0$ = intercept for Before group — "reference" category

$\hat{\beta}_0 + \hat{\beta}_1$ = intercept for AtPFI group

timeAtPFI $\hat{\beta}_1$ = difference between intercepts

inten $\hat{\beta}_2$ = slope for Before group → "generally" slope of first of the groups

$\hat{\beta}_2 + \hat{\beta}_3$ = slope for AtPFI group

$\hat{\beta}_3$ = difference between slopes

timeAtPFI x inten

In this case, with just 2 groups, the tests from the `summary()` output are equivalent to the `Anova()` tests, with $F = t^2$. The F-tests extend to more than two groups.

```
summary(Model2)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      83.146667   4.343305  19.144 2.49e-14
timeAtPFI       -11.523333   6.142360  -1.876   0.0753
inten            -0.039867   0.007435  -5.362 3.01e-05
timeAtPFI:inten  -0.001210   0.010515  -0.115   0.9096
```

cat · cont · int

H0: intercept for "ref" = 0

not of interest

H0: intercept diff = 0

```
Anova(Model2, type = 3)
             Sum Sq Df F value    Pr(>F)
time          153.2  1  3.5195   0.07532 .
inten        1251.6  1 28.7509 3.008e-05 ***
time:inten      0.6  1  0.0132   0.90957
```

H0: slope ref = 0

H0: diff in slope = 0

Hypothesis Testing ANCOVA With Interaction Model

1. **Does the effect of intensity (X) depend on time?**
   Or, Is there a <u>difference</u> between the <u>slopes</u>? Use the "time:inten" line in the ANOVA table.
   H_0: $\beta_3 = 0$ vs H_A: $\beta_3 \neq 0$
   $F = 0.013$ or $t = -0.12$, $p=0.9096$
   Fail to Reject H0, no evidence that the slopes differ.

2. **Does the intercept depend on time? Or, Is there a <u>difference</u> between the <u>intercepts</u>?** Use the "time" line in the ANOVA table.
   H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$ FTR
   $F = 3.52$ or $t=-1.88$, $p=0.0753$
   Fail to Reject H0, weak evidence that intercepts differ.

3. **Is there a difference between mean responses for the two timing groups at fixed value of the inten?** Use `emmeans()` to test for a difference between the mean response for Before vs AtPFI groups when inten = 525 (average intensity value).
   $t = 4.51$, $p=0.0002$
   Reject H0, conclude there is a difference.

default in R at $\bar{x} = 525$

* Note that when the ANCOVA model includes interaction (allowing different slopes), the result of this test depends on the value of x where we test!

By default, `emmeans()` uses the average x value. To test for a difference between groups at a different value of x, use the "`at`" option. Ex: `emmeans(Model2, pairwise ~ time, at = list(inten = 700))` not often

“Alternate” Parameterization for ANCOVA With Interaction (to calculate intercepts and slopes directly)

For "easy" equations of line(s)

```
Model3 <- lm(flowers ~ time + time:inten - 1,
data = Meadow)
```

- The “-1” omits the intercept, this forces R to estimate a separate intercept for each level of the categorical predictor (time).
- Omitting the main effect for the continuous predictor (inten) forces R to estimate a separate slope for each level of the categorical predictor (or factor).
- Often helpful to run the model both ways. Standard good for testing, “alternate” approach for estimation.
- Do NOT use the Type 3 tests from Alternate Parameterization!

Comparing estimates from “Alternate” and “Standard” Parameterizations

easier for eq of lines model 3

omit intercept -1

model 2

Directly from			"By Hand" from
Alternate Parameterization			Standard Parameterization
timeBefore	Int Before	83.1467	83.1467
timeAtPFI	Int AtPFI	71.6233	83.1467 - 11.5233
timeBefore:inten	Slope Before	-0.03987	-0.03987
timeAtPFI:inten	Slope AtPFI	-0.04108	-0.03987 - 0.00121

End here 2/19

Return to ANCOVA Analysis Goals:

1. • Estimate slopes and intercepts. Use parameter estimate information ("Coefficients" table). Consider using "alternate" parameterization.
2. • Estimate mean response at fixed value of the continuous predictor and/or test for differences. In other words, test for a difference between means (for the categorical predictor) controlling for the continuous predictor. Use `emmeans()`.

- Testing for NO Interaction model. Use Type3 ANOVA tests.
  - Test corresponding to categorical predictor tests for differences between the intercepts.
  - Test corresponding to continuous predictor tests whether the common slope is different from zero.

- Testing WITH Interaction model. Use Type3 tests from the "standard" parameterization.
  - Test corresponding to categorical predictor tests for differences between the intercepts.
  - Test corresponding to interaction tests for differences between the slopes.
  - To get pairwise comparisons of slopes or intercepts
    Easy way: use `emmeans()` and/or `emtrends()`
    Or use: `lht()` function (from `car` package) with the "alternate" parameterization.
    See the Glue Strength Example.

like emmeans but w/ slopes

lm(y ~ . -1)

Example write-up for Meadow Foam Data:

Option 1 (With Interaction Model, Focusing on Effect of light Intensity separately for the two groups):
Analysis was done using R. A linear model was fit to the data with number of flowers per plant as the response. Predictor variables included timing (At or Before PFI) and light intensity (in Watts) plus interaction. For the AtPFI group, the estimated slope corresponding to light intensity was found to be -0.041 (SE = 0.007, p-value < 0.001). For the BeforePFI group, the estimated slope corresponding to light intensity was found to be -0.039 (SE = 0.007, p-value < 0.001). There was not a statistically significant difference between the slopes for the two groups (p = 0.9096)

interaction not significant

Option 2 (No Interaction Model):
Analysis was done using R. A linear model was fit to the data with number of flowers per plant as the response. Predictor variables included timing (At or Before PFI) and light intensity (in Watts). The estimated slope corresponding to light intensity was found to be -0.040 (SE = 0.005, p-value < 0.001). Adjusting for light intensity, there was a significant difference between mean flowers for the AtPFI vs BeforePFI groups (estimated difference = -12.1583, SE = 2.63, p-value = 0.0001)

emmeans

gap b/w lines

Summary Graph: Color coded scatterplot shown earlier.

Glue Strength Example: Four formulations of glue are tested for strength in a one-way design (completely randomized). Because strength of a test sample is related to the thickness of glue application, which is hard to control precisely, the glue thickness was measured on each sample.

The objective is the compare average strength of glue formulations, <u>adjusted to as if they were applied at the same thickness.</u>

Y = measured strength (continuous response)
X = thickness (continuous predictor=covariate)
Glue =A, B, C and D (categorical predictor) with 5 obs/trt.

We consider three models :

Model1: One-way ANOVA

Model2: ANCOVA No Interaction

Model3: ANCOVA With Interaction

Conclusions about the Glue Strength Example:

1. Without any adjustment for application thickness (one-way ANOVA model, Model1), the formulations are not significantly different (p = 0.0983).

2. Based on the ANCOVA with interaction model (Model3), there is no evidence of differences between slopes (F-test for thick:glue p = 0.5312). Remember we want to use the "standard parametrization" for most testing.

3. Since interaction is not significant but thick is significant, it is reasonable to use the ANCOVA no interaction model (Model2) for further analysis.

4. The research goals were focused on comparing strengths of the glue formulations (no research questions about slopes or intercepts!), we will focus on the Model2 emmeans output. After Tukey adjustment, A is significantly ($p=0.009$) weaker than C. No other comparisons are statistically significant.

Important note: In practice, when the continuous predictor is really considered a "covariate" (of secondary interest), the no interaction model is usually used unless there is some compelling reason to include interaction. That said, it never hurts to check for interaction using a summary graphic.

Choosing a model for Glue Strength:

We will discuss model selection in more detail later (including defining AIC). But here are 2 approaches to model selection for Glue Strength:

1. **Backward Elimination:** Start with full model: ANCOVA With Interaction (Model3). Based on the Type3 F-test for "thick:glue" with p-value > 0.05, we would drop that term from the model. That reduces the model to ANCOVA with NO Interaction (Model2). All Type3 F-tests are significant, so we would use that model.

2. **AIC**: Choose model with lowest AIC.

   Model1: One Way ANOVA (AIC=35.86)

   Model2: ANCOVA No Interaction (AIC=10.94)

   Model3: ANCOVA With Interaction (AIC = 13.41)

Example write-up for Glue Strength data:

Analysis was done using R and the emmeans package (REF). A linear model (or ANCOVA model) was fit to the data. The response variable is strength and predictor variable is glue (A, B, C or D). Application thickness (measured in mm) was included in the model as a covariate. Tukey adjusted pairwise comparisons were considered. Glue C was found to be significantly stronger that Glue A (Tukey adjusted p-value = 0.0095). No other significant differences were found.

Yay!! glue C

Glue	n	mean	SE
A	5	46.08	0.54
B	5	47.64	0.54
C	5	48.91	0.53
D	5	47.43	0.53

Notes: (1) emmeans and model based SE used here. Simple means and std dev (or SE) also reasonable. (2) Boxplots or bar charts would be reasonable graphics.

Referencing R and packages:
Remember that R packages are created by individual authors. Packages that were critical to your analysis should be mentioned in the write up and included in your references.

You can get the citation information within R.

To get the citation for R:
`citation()`

To get the citation for a specific package (ex: emmeans):
`citation("emmeans")`

4. Misc R Topics

If you want to regress a response on ALL other variables in the data.frame you can use this short cut:

```
lm(y ~ . , data = InData)
```

The `update()` function can be used to create a new model object by changing one or more arguments. Here are some example with the Glue Strength data:

- Remove the "thick" term from the ANCOVA (reducing to the one-way ANOVA):

```
newmodel <- update(model2, ~ . - thick)
```

- Remove the first observation from the dataset:

```
newmodel  <- update(model2, subset = -1)
```

FYI

Standardized Regression Coefficients are obtained when the predictors and response are standardized to have sample means of 0 and standard deviations of 1. This can be done using the `scale()` function in R.

It does NOT make sense to scale factors, indicator variables or interactions!

Comparing standardized vs "raw" regression, the test statistics, p-values and R2 will be the same. All that changes is the interpretation of the regression coefficients.

* Specifically instead "a one-unit" increase in X, we can think of a "one standard deviation" increase in X.