Linear Regression with Continuous and Categorical Predictors

PSY 612

Linear Regression with Continuous & Categorical Predictors

- Linear regression models that included both a continuous and a categorical predictor were originally developed in the context of (quasi-)experimental studies
  - Traditionally called an **ANCOVA** (Analysis of Covariance)
  - Researchers were mainly interested in the effect of a categorical predictor (the experimental manipulation) on the outcome variable while controlling for continuous variable(s), called **covariates**

- We can think of these models more generally as we've thought of previous models with multiple predictors
  - Interested in the relationships between a continuous and a categorical predictor with an outcome variable while controlling for the other predictor in the model

Example

- Let's say a research developed a new educational intervention that college students could sign up for in which they work collaboratively with a group of peers on their weekly course assignments. **The researcher wants to test the effect of this educational intervention on students' final grades in their most difficult course**. However, the researcher is aware that other variables could also systematically differ between the types of students who sign up versus do not sign up for this intervention. Thus, the researcher also decides to control for the number of hours spent studying by each student when predicting their final grades.

intervention	hours_study	final_grade
Control	34	77
Control	50	89
Control	30	72
Control	48	74
Intervention	44	88
Intervention	56	93
Intervention	58	94
Intervention	32	90

Example

- Education Intervention (Categorical predictor)
  - Control Group
  - Intervention Group
- Hours spent studying (Continuous predictor)
- Final grade (Outcome variable)

intervention	hours_study	final_grade
Control	34	77
Control	50	89
Control	30	72
Control	48	74
Intervention	44	88
Intervention	56	93
Intervention	58	94
Intervention	32	90

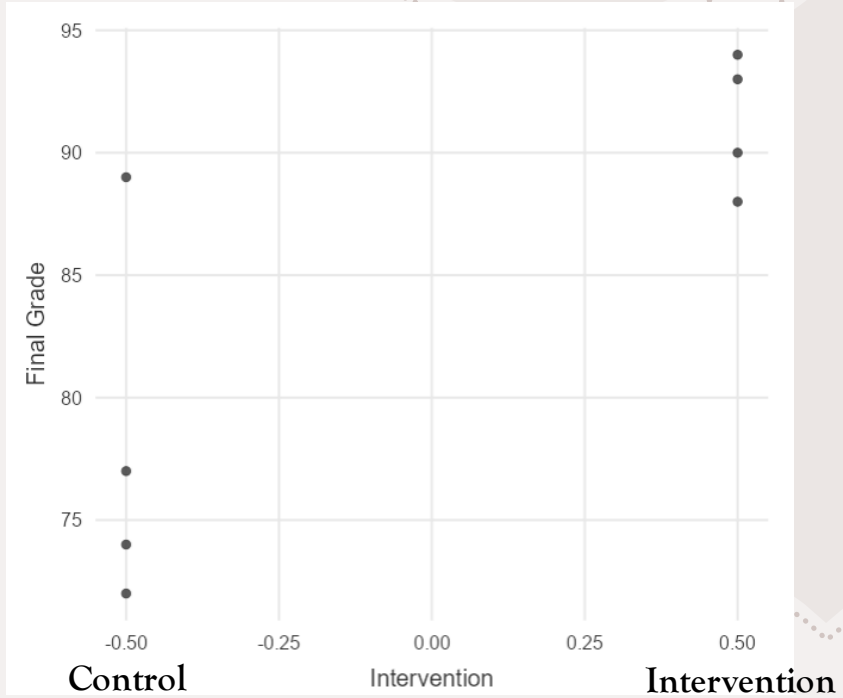
Code the Categorical Predictor

Recall the Rules of Contrast Coding:

1. Each set of codes must sum to 0.
2. The sum of the products of codes in corresponding positions must equal 0.
3. *Recommended:* Put each contrast code on a scale of "1", meaning that the span between the contrast codes is equal to 1.

For our example:

	Control	Intervention
InterventionCC	-1/2	+1/2



Center the Continuous Predictor

The advantages of mean-centering the continuous predictor include:

1) A more meaningful y-intercept value

2) Less multicollinearity between continuous predictors with a continuous X continuous interaction term composed from them

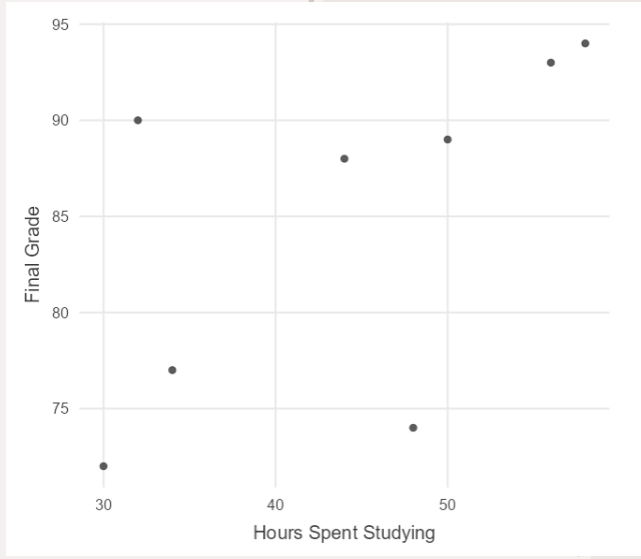
The second advantage doesn't apply here, but we'll center our continuous predictor for the first benefit.

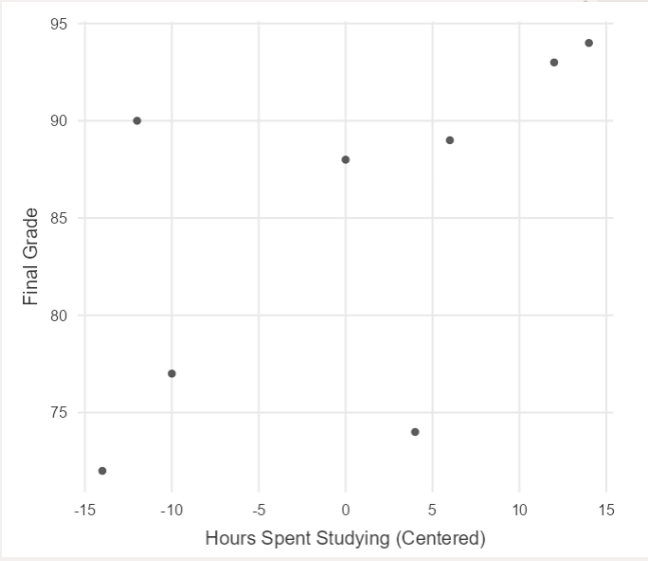
hours_study	hours_study_c
34	34 – 44 = -10
50	6
30	-14
48	4
44	0
56	12
58	14
32	-12

$M_{HoursStudy} = 44$

Center the Continuous Predictor

hours_study	hours_study_c
34	-10
50	6
30	-14
48	4
44	0
56	12
58	14
32	-12





Model Comparison

Model Comparison:

Model A: $Y_i = \beta_0 + \beta_1 InterventionCC_i + \beta_2 Hours\_Study\_C_i + \varepsilon_i$ PA = 3

Model C: $Y_i = \beta_0 + \beta_2 Hours\_Study\_C_i + \varepsilon_i$ PC = 2

Null & Alternative Hypotheses:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Fit the Model in R

- Import (or set up) data

```
intervention <- c(rep("Control",4),rep("Intervention",4))
hours_study <- c(34, 50, 30, 48, 44, 56, 58, 32)
final_grade <- c(77, 89, 72, 74, 88, 93, 94, 90)

data <- cbind.data.frame(intervention,hours_study,final_grade)
```

- Check measure types & convert if needed

```
> str(data)
'data.frame':   8 obs. of  3 variables:
 $ intervention: chr  "Control" "Control" "Control" "Control"
 $ hours_study : num  30 34 50 48 44 56 58 32
 $ final_grade : num  77 89 72 74 88 93 94 90
```

```
data <- data %>%
  mutate(intervention = factor(intervention, levels = c("Control","Intervention")))
```

Descriptive Statistics

- Overall mean and SD on Hours Studied and Final Grade

Table: Descriptive Statistics

n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
8.00	44.00	10.90	84.62	8.85

- Table of the mean and SD separately for each group

Table: Descriptive Statistics

Intervention	n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
Control	4.00	40.50	9.98	78.00	7.62
Intervention	4.00	47.50	12.04	91.25	2.75

Fit the Model in R

- Contrast code the categorical predictor

```
InterventionCC <- c(-1/2, 1/2)
contrasts(data$intervention) <- InterventionCC
```

Check your codes!

```
> contrasts(data$intervention)
             [,1]
Control      -0.5
Intervention  0.5
```

- Center the continuous predictor

```
data$hours_study_c <- c(scale(hours_study, center = TRUE, scale = FALSE))
```

- Fit the model

```
model <- lm(final_grade ~ intervention + hours_study_c, data = data)
```

Assumption of Homogeneity of Regression

- Before we interpret the model's output, let's discuss the test's assumptions. We are still making the familiar assumptions:
  - Normally distributed errors
  - Independence of errors
  - Homogeneity of variance

- Remember that when an interaction effect between predictors is not included in the model, the researcher is making the implicit assumption that there is no interaction between the predictors.
  - The traditional ANCOVA analysis makes this assumption explicitly, and it's called the **Assumption of Homogeneity of Regression**

Assumption of Homogeneity of Regression

- The **assumption of homogeneity of regression** assumes that:

  - The relationship between the continuous predictor and the outcome variable is the same across both levels of the categorical predictor

- We can test this assumption by simply fitting a model that includes the continuous by categorical predictor interaction effect and making sure it is non-significant

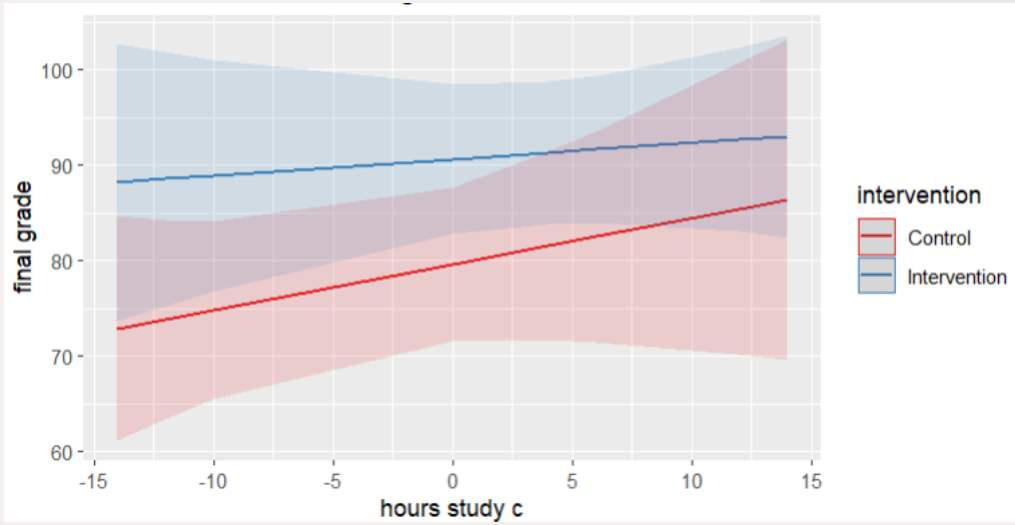
Have we met the assumption of homogeneity of regression?

```
model_int <- lm(final_grade ~ intervention*hours_study_c, data = data)
```

```
Anova Table (Type III tests)

Response: final_grade
                          Sum Sq Df   F value    Pr(>F)
(Intercept)                50980  1 1778.7994 1.889e-06
intervention                 211  1    7.3710   0.05326
hours_study_c                 76  1    2.6354   0.17983
intervention:hours_study_c    17  1    0.5955   0.48336
Residuals                    115  4
```

```
plot_model(model_int,
           type = "pred",
           terms = c("hours_study_c", "intervention"))
```



Model Output

We can examine our model's results using the following functions:

- **summary(model)**
  - For parameter estimates & the significance of each individual predictor
- **confint(model)**
  - For confidence intervals around each parameter estimate
- **Anova(model, type = 3)**
  - For the ANOVA summary table
- **etaSquared(model, type = 3)**
  - For effect sizes

Summary Output

summary( ) output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   84.6250     1.8146  46.636 8.56e-08 ***
intervention1 11.1662     3.8638   2.890   0.0342 *
hours_study_c  0.2977     0.1894   1.571   0.1769
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.132 on 5 degrees of freedom
Multiple R-squared:  0.7596,    Adjusted R-squared:  0.6634
F-statistic:   7.9 on 2 and 5 DF,  p-value: 0.02833
```

confint( ) output:

```
                  2.5 %     97.5 %
(Intercept)   79.960508 89.2894917
intervention1  1.233955 21.0984696
hours_study_c -0.189285  0.7846529
```

Parameter Estimates:

- b_0 = 84.63, 95%CI[79.96, 89.29]
- b_1 = 11.17, 95%CI[1.23, 21.10]
- b_2 = 0.30, 95%CI[-0.19, 0.78]

Full Model Estimate Equation:

- final_grade’ = 84.63 + 11.17*intervention + 0.30*hours_study_c

Interpreting the Parameter Estimates

Full Model Estimate Equation:

final_grade' = 84.63 + 11.17*intervention + 0.30*hours_study_c

What does each parameter estimate mean?

Interpreting the Parameter Estimates:

- $b_0 = 84.63$ The mean of the groups' means on final grades because we contrast coded intervention and centered hours spent studying
- $b_1 = 11.17$ The mean final grade of the intervention condition minus the mean final grade of the control condition *after adjusting the means for hours spent studying*
- $b_2 = 0.30$ The predicted change in final grades per 1-unit change in hours spent studying *when controlling for intervention condition*

Table: Descriptive Statistics

n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
8.00	44.00	10.90	84.62	8.85

Table: Descriptive Statistics

Intervention	n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
Control	4.00	40.50	9.98	78.00	7.62
Intervention	4.00	47.50	12.04	91.25	2.75

Adjusted Means

- Formula for Calculating the Adjusted Mean for each condition of the categorical predictor:

$$\bar{Y}'_k = \bar{Y}_k - b_z(\bar{Z}_k - \bar{Z})$$

- $\bar{Y}'_K$ = adjusted mean on the outcome variable for condition k of the categorical predictor
- $\bar{Y}_K$ = original mean on the outcome variable for condition k of the categorical predictor
- b_Z = the parameter estimate corresponding to the continuous covariate from the model output
- $\bar{Z}_K$ = the mean of the continuous predictor for condition k of the categorical predictor
- $\bar{Z}$ = the overall mean on the continuous predictor

Adjusted Means

Table: Descriptive Statistics

n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
8.00	44.00	10.90	84.62	8.85

Table: Descriptive Statistics

Intervention	n	Mean Hours Studied	SD Hours Studied	Mean Final Grade	SD Final Grade
Control	4.00	40.50	9.98	78.00	7.62
Intervention	4.00	47.50	12.04	91.25	2.75

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   84.6250     1.8146  46.636 8.56e-08 ***
intervention1 11.1662     3.8638   2.890   0.0342 *
hours_study_c  0.2977     0.1894   1.571   0.1769
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.132 on 5 degrees of freedom
Multiple R-squared:  0.7596,    Adjusted R-squared:  0.6634
F-statistic:   7.9 on 2 and 5 DF,  p-value: 0.02833
```

For our example:

$$\bar{Y}'_k = \bar{Y}_k - b_z(\bar{Z}_k - \bar{Z})$$

- The adjusted mean for the **Control** condition:
  - $M_{Adj} = 78 - (0.30)*(40.50 - 44) = 79.05$

- The adjusted mean for the **Intervention** condition:
  - $M_{Adj} = 91.25 - (0.30)*(47.5 - 44) = 90.20$

Getting adjusted means in R using the emmeans( ) function:

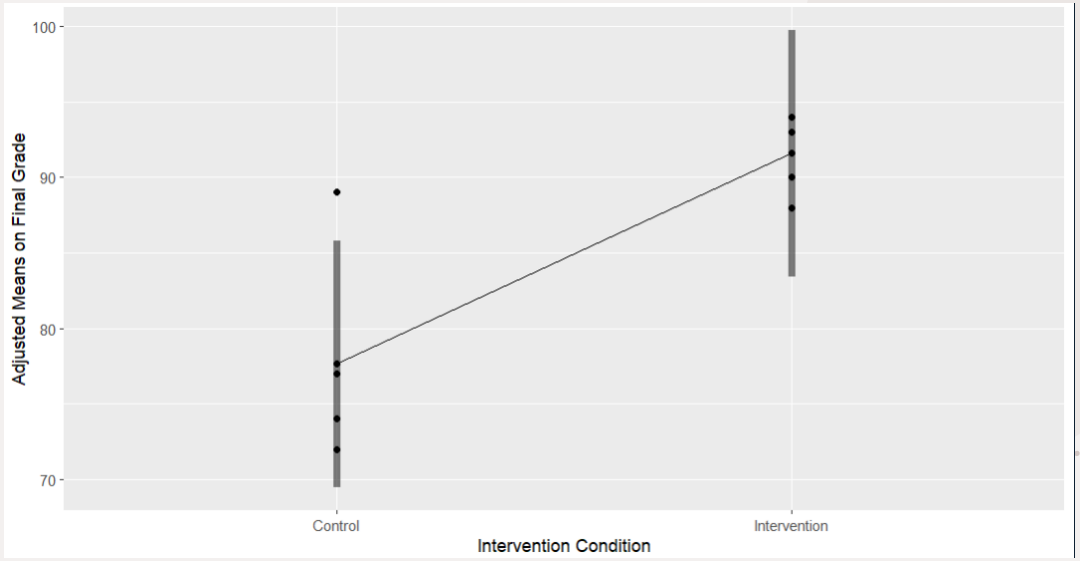
```
emmeans(model, specs = c("intervention","hours_study_c"))
```

```
 intervention hours_study_c emmean   SE df lower.CL upper.CL
 Control                  0   79.0 2.65  5     72.2     85.9
 Intervention             0   90.2 2.65  5     83.4     97.0
```

Adjusted Means

- Graphing the adjusted means for each condition of the categorical predictor using emmip( ) function:

```
emmip(model, ~ intervention, CIs = TRUE, xlab = "Intervention
Condition", ylab = "Adjusted Means on Final Grade") +
geom_point(data = data, aes(x = intervention, y = final_grade))
```



Anova( ) Output

Anova(model, type = 3) output:

```
Anova Table (Type III tests)

Response: final_grade
             Sum Sq Df   F value    Pr(>F)
(Intercept)   57291  1 2174.9616 8.561e-08 ***
intervention    220  1    8.3518   0.03419 *
hours_study_c    65  1    2.4693   0.17689
Residuals       132  5
```

Model Comparison:

Model A: $Y_i = \beta_0 + \beta_1 \text{InterventionCC}_i + \beta_2 \text{Hours\_Study\_C}_i + \varepsilon_i$

Model C: $Y_i = \beta_0 + \beta_2 \text{Hours\_Study\_C}_i + \varepsilon_i$

Source	*SS*	*df*	*MS*	*F*	*p*
intervention	SSR = 220	PA-PC = 1	MSR = 220	$F = 8.33$	$p = .034$
hours_study_c					
Model A	SSE(A) = 132	n-PA = 5	MSE = 26.4		

Effect Sizes

etaSquared(model, type = 3) output:

```
                 eta.sq eta.sq.part
intervention  0.4015439   0.6255179
hours_study_c 0.1187204   0.3305918
```

SS_{Total} on Y (final grades):

```
> var(data$final_grade)*(nrow(data)-1)
[1] 547.875
```

- To find the **total SS on our outcome variable**, recall that the variance of a variable is equal to its SS divided by its df
  - Variance = $\frac{SS}{df}$, where df = n-1 for a single variable
- SS_{Total} = variance*(n-1)

Eta-Squared: $\eta^2 = \frac{SSR}{SS_{Total}}$

For intervention in our example:

- $\eta^2 = \frac{220}{547.875} = 0.40$

Partial Eta-Squared: $\eta^2_{Partial} = \frac{SSR}{SSE(C)}$

For intervention in our example:

- $\eta^2_{Partial} = \frac{220}{352} = 0.63$
- Also called *PRE*

Summary of the Findings

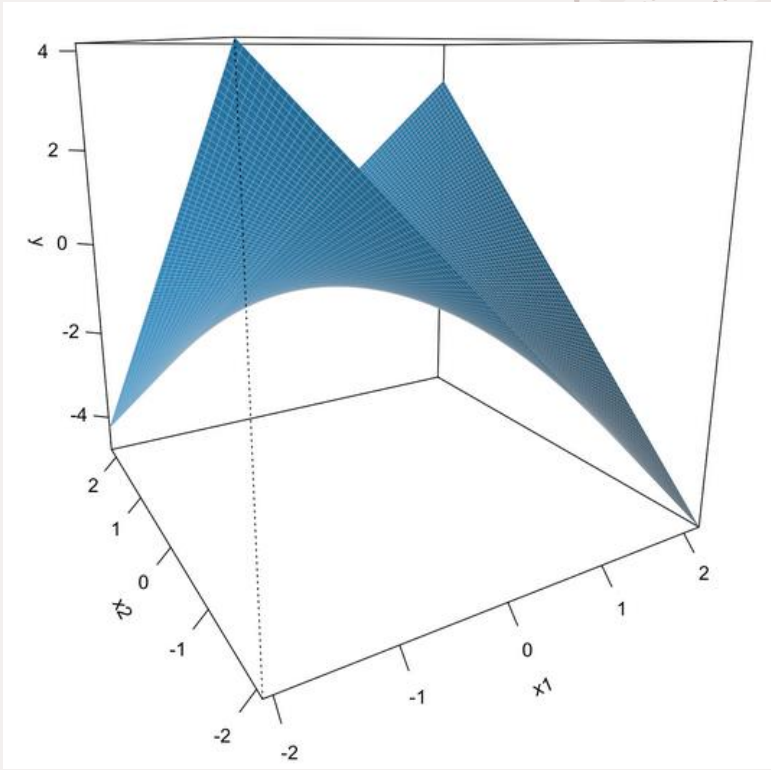
Final grades for students who underwent the educational intervention (M = 91.25, SD = 2.75, $M_{Adjusted}$ = 90.20) were significantly higher compared to students in the control condition (M = 78.00, SD = 7.62, $M_{Adjusted}$ = 79.05) even when controlling for number of hours that student spent studying, b_1 = 11.17, 95%CI[1.23, 21.10], $t(5) = 2.89$ (or $F(1,5) = 8.35$), $p = .034$, $\eta^2 = 0.40$, $\eta_p^2 = 0.63$.

Continuous by Categorical Interactions

Continuous by Categorical Interactions

- Although the traditional ANCOVA assumes that there is no interaction between the continuous and categorical predictor in the model, we may be interested in the nature of a continuous by categorical predictor interaction effect.

- Remember that **an interaction effect is when the relationship (or effect) of one predictor with the outcome variable <u>varies depending on</u> the level of another predictor.**



Example

- In the previous example, we saw that there was no interaction effect between hours that students spent studying and which intervention condition they were in.
- Let's say the researcher also measured students' satisfaction with their college experiences during the study. The researcher wants to examine whether there is an interaction effect between student satisfaction and intervention condition.
- In other words, the researcher wants to examine whether the relationship between student satisfaction and final grades *differs depending on* which intervention condition students were in.

intervention	satisfaction	final_grade
Control	7	77
Control	4	89
Control	9	72
Control	8	74
Intervention	6	88
Intervention	8	93
Intervention	10	94
Intervention	7	90

Example

- Before fitting the model, we'll…
  - Contrast code the categorical predictor
    - Control = -1/2
    - Intervention = +1/2
  - Center the continuous predictor

intervention	satisfaction_c	final_grade
-1/2	-0.375	77
-1/2	-3.375	89
-1/2	1.625	72
-1/2	0.625	74
1/2	-1.375	88
1/2	0.625	93
1/2	2.625	94
1/2	-0.375	90

Model Comparison

Model Comparison:

Model A: $Y_i = \beta_0 + \beta_1 InterventionCC_i + \beta_2 Satisfaction\_C_i + \beta_3 (InterventionCC*Satisfaction\_C)_i + \varepsilon_i$

Model C: $Y_i = \beta_0 + \beta_1 InterventionCC_i + \beta_2 Satisfaction\_C_i + \varepsilon_i$

Null & Alternative Hypotheses:

H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

Fit the Model in R

- Import (or set up) data

```r
intervention <- c(rep("Control",4),rep("Intervention",4))
hours_study <- c(34, 50, 30, 48, 44, 56, 58, 32)
final_grade <- c(77, 89, 72, 74, 88, 93, 94, 90)
satisfaction <- c(7, 4, 9, 8, 6, 8, 10, 7)

data <- cbind.data.frame(intervention,hours_study,satisfaction,final_grade)
```

- Contrast code the categorical predictor

```r
InterventionCC <- c(-1/2, 1/2)
contrasts(data$intervention) <- InterventionCC
```

- Center the continuous predictor

```r
data$satisfaction_c <- c(scale(data$satisfaction, center = TRUE, scale = FALSE))
```

- Fit the model with the interaction effect

```r
model <- lm(final_grade ~ intervention*satisfaction_c, data = data)
```

Anova Output

Anova(model, type = 3) output:

```
Anova Table (Type III tests)

Response: final_grade
                         Sum Sq Df   F value    Pr(>F)
(Intercept)               53244  1 41069.859 3.557e-09 ***
intervention                372  1   287.140 7.111e-05 ***
satisfaction_c               21  1    16.377 0.0155157 *
intervention:satisfaction_c  135  1   104.430 0.0005167 ***
Residuals                    5  4
```

Model Comparison:

Model A: $Y_i = \beta_0 + \beta_1 InterventionCC_i + \beta_2 Satisfaction\_C_i + \beta_3 (InterventionCC*Satisfaction\_C)_i + \varepsilon_i$

Model C: $Y_i = \beta_0 + \beta_1 InterventionCC_i + \beta_2 Satisfaction\_C_i + \varepsilon_i$

Source	*SS*	*df*	*MS*	*F*	*p*
intervention					
satisfaction_c					
intervention*satisfaction_c	SSR = 135	PA-PC = 1	MSR = 135	$F = 108$ (rounding error)	$p < .001$
Model A	SSE(A) = 5	n-PA = 4	MSE = 1.25		

Summary Output

- summary( ) output

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                83.6848     0.4129 202.657 3.56e-09 ***
intervention1              13.9946     0.8259  16.945 7.11e-05 ***
satisfaction_c             -0.9929     0.2453  -4.047 0.015516 *
intervention1:satisfaction_c  5.0143     0.4907  10.219 0.000517 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.139 on 4 degrees of freedom
Multiple R-squared:  0.9905,    Adjusted R-squared:  0.9834
F-statistic: 139.5 on 3 and 4 DF,  p-value: 0.0001674
```

- For this example, we’re particularly interested in the interaction effect.
- To unpack the nature of the interaction effect, let’s examine how the slope of the relationship between `satisfaction_c` and `final_grade` varies between the two intervention conditions.

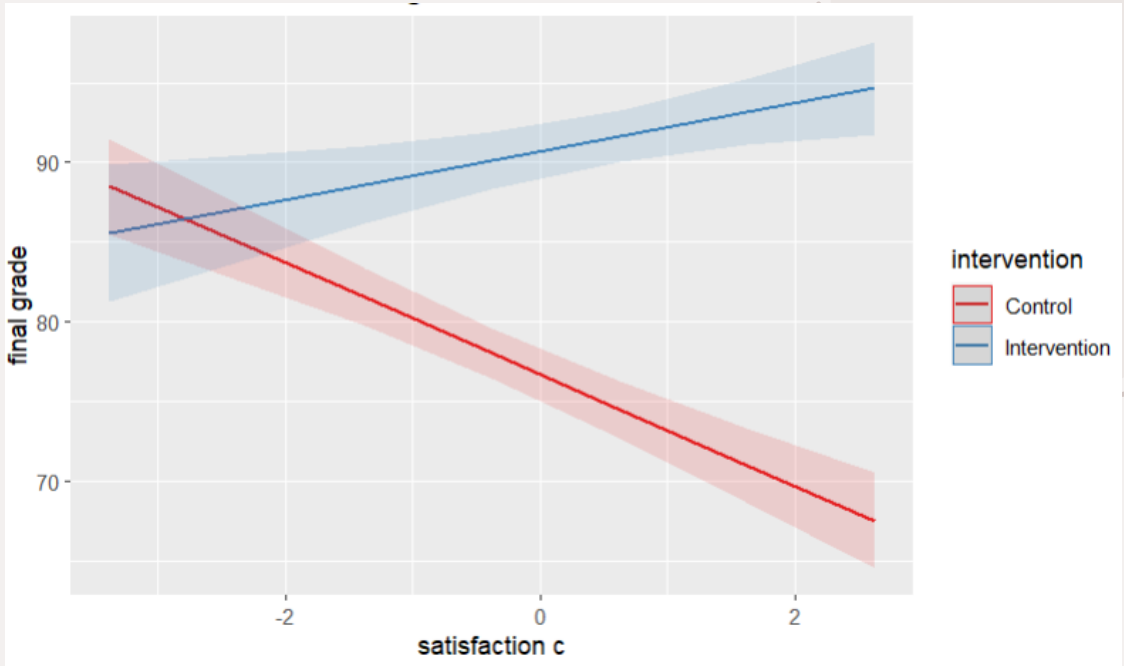
Full Model Estimate Equation:

final_grade’ = 83.68 + 13.99*Intervention – 0.99*Satisfaction_C + 5.01*(Intervention*Satisfaction_C)

Visualizing Simple Slopes

- We can visualize the simple slopes using the `plot_model()` function in R
  - Question: How would you interpret the nature of the interaction effect? In other words, how does the relationship between satisfaction and final grade appear to differ *depending on* which level of the intervention students were in?

```
plot_model(model,
           type = "pred",
           terms = c("satisfaction_c", "intervention"))
```



Simple Slopes Analyses

Full Model Estimate Equation:

final_grade' = 83.68 + 13.99*Intervention – 0.99*Satisfaction_C + 5.01*(Intervention*Satisfaction_C)

- We can solve for the **simple slope** for the relationship between `satisfaction_c` and `final_grade` for each level of the `intervention` by plugging in the value that we coded each level of the intervention to solve for its simple slope equation
  - Control = -1/2
  - Intervention = +1/2

Simple Slope for the Control Condition

Full Model Estimate Equation:

final_grade' = 83.68 + 13.99*Intervention – 0.99*Satisfaction_C + 5.01*(Intervention*Satisfaction_C)

- Solving for the **simple slope equation for the control condition:**
- final_grade' = 83.68 + 13.99*(-1/2) – 0.99*Satisfaction_C + 5.01*(-1/2*Satisfaction_C)
- final_grade' = 83.68 – 6.995 – 0.99*Satisfaction_C – 2.505*Satisfaction_C
- **final_grade' = 76.69 – 3.50*Satisfaction_C**

Simple Slope for the Intervention Condition

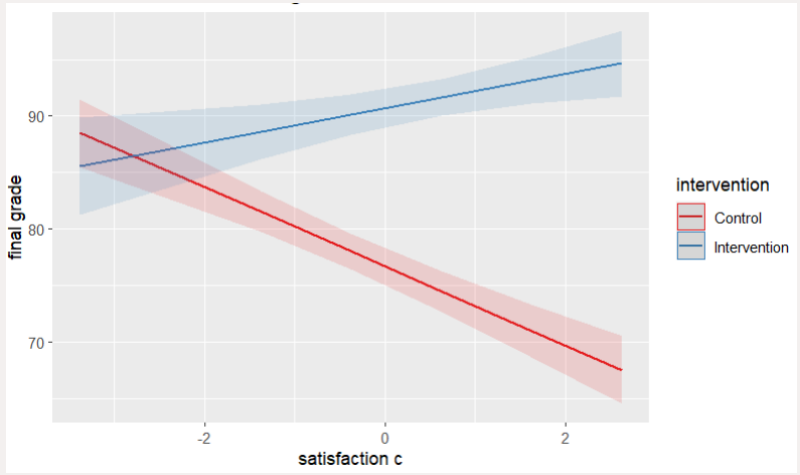
Full Model Estimate Equation:

final_grade' = 83.68 + 13.99*Intervention – 0.99*Satisfaction_C + 5.01*(Intervention*Satisfaction_C)

- Solving for the **simple slope equation for the intervention condition:**
- final_grade' = 83.68 + 13.99*(1/2) – 0.99*Satisfaction_C + 5.01*(1/2*Satisfaction_C)
- final_grade' = 83.68 + 6.995 – 0.99*Satisfaction_C + 2.505*Satisfaction_C
- **final_grade' = 90.68 + 1.52*Satisfaction_C**

Simple Slopes in R

- We can perform simple slopes analyses in R to unpack the nature of a continuous by categorical interaction using the `emtrends( )` function




```
emtrends(model, ~intervention, var = "satisfaction_c") # confidence intervals
```

```
 intervention satisfaction_c.trend    SE df lower.CL upper.CL
 Control                    -3.50 0.304  4   -4.345    -2.66
 Intervention                1.51 0.385  4    0.446     2.58

Confidence level used: 0.95
```

```
test(emtrends(model, ~intervention, var = "satisfaction_c")) # p-values
```

```
 intervention satisfaction_c.trend    SE df t.ratio p.value
 Control                    -3.50 0.304  4 -11.502  0.0003
 Intervention                1.51 0.385  4   3.934  0.0170
```