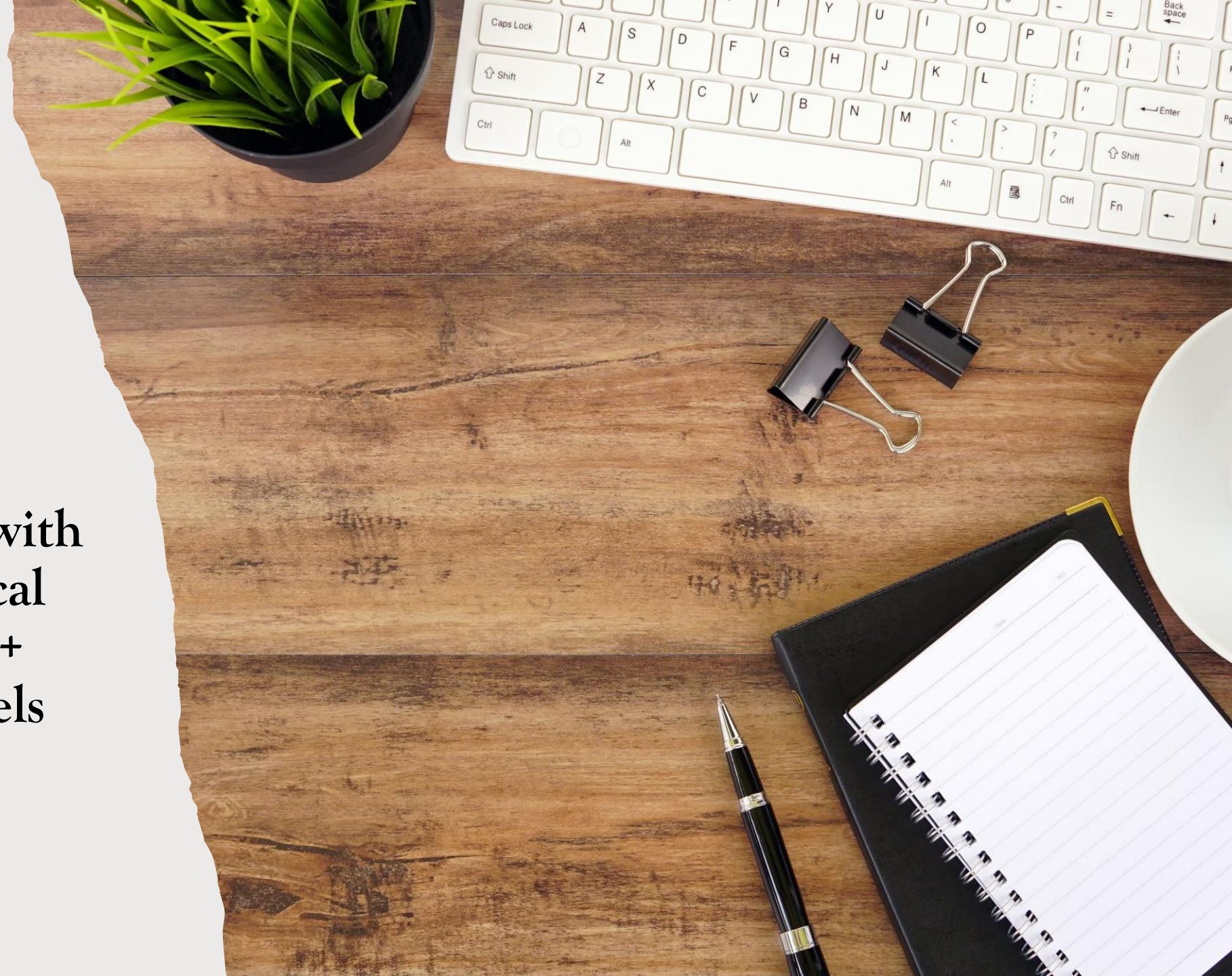
Linear Regression with a Single Categorical Predictor with 2+ Independent Levels

aka,
One-Way Between-Subjects ANOVA



RECAP

- Last time, we talked about adding a single categorical predictor to the model with **two independent levels**

- Next, we'll discuss adding a single categorical predictor to the model with **more than two independent levels**

EXAMPLE

- **Example:** A developmental psychologist is interested in whether the type of feedback children receive has an effect on their subsequent motivation to do a task. The researcher randomly assigns 24 children to either receive 1) no feedback, 2) feedback that they were successful on an initial trial, or 3) feedback that they failed on an initial trial. Then, the researcher gives the children a list of brain teasers and tells them that they can attempt to complete as many of them as they wish (i.e., "What has a face and two hands but no arms or legs?").

IV: Feedback Type
- None
- Success
- Failed

DV: Number of brain teasers attempted

No Feedback	Success	Failure
4	4	2
3	6	2
4	5	2
5	4	3
5	6	4
2	4	4
4	3	3
3	3	4
$M_1 = 3.75$	$M_2 = 4.38$	$M_3 = 3.00$

CODING THE CATEGORICAL PREDICTOR

- Recall that whenever a categorical predictor is used in a model, its levels must be coded. We covered two types of coding schemes last week:
  - Dummy coding
  - Contrast coding

- The number of codes that need to be included in the model to capture a categorical predictor are:
  - $m - 1$,
    - m = the number of levels of the categorical predictor

- For our example, we need $m - 1 = 3 - 1 = 2$ codes to represent feedback type

CONTRAST CODING

- One of the benefits of contrast codes is that they can be used to test a number of different specific mean comparisons that are of theoretical interest to the researcher
  (i.e., **planned comparisons** when they are specified prior to examining the data)

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.

Recommended for ease of interpretation:
Put each contrast code on a scale of "1", meaning that the span between the contrast codes is equal to 1 (for example [-1/3, -1/3, 2/3] would be preferable to [-1, -1, 2]).

CONTRAST CODING

- For our example, let's say the researcher was theoretically interested in testing whether children's motivation to continue attempting brain teasers in the **Success** condition is significantly different from the average of both the **No Feedback** and **Failure** conditions.

- Q: How could we code the levels of Feedback Type so that they test this mean comparison and follow the rules of contrast coding?

	None	Success	Failure
FeedbackCode1	**-1/3**	**2/3**	**-1/3**
FeedbackCode2			

CONTRAST CODING

- Once you have the first, theoretically-motivated contrast code, the second contrast code must be specified so that the sum of the cross products with the first code equals zero.

- The sum of the products of the codes that are in corresponding positions is calculated by multiplying codes in the same vertical position and then taking the sum of these products

	None	Success	Failure
FeedbackCode1	**-1/3**	**2/3**	**-1/3**
FeedbackCode2	**1/2**	**0**	**-1/2**
Sum of Products:	-1/6	0	+1/6 = 0

THE MODEL COMPARISON: TESTING A SINGLE PREDICTOR

- Let's specify the model comparison corresponding to a test of whether the mean of the success condition is significantly different from the mean of the no feedback and failure conditions.

- Model A is the full model including both contrast codes:

$$\textbf{Model A: } Y_i = \beta_0 + \beta_1 \text{FeedbackCode1}_i + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$$

- Model C corresponds to the null hypothesis. It leaves out the predictor that we are wanting to test the significance of:

$$\textbf{Model C: } Y_i = \beta_0 + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$$

THE MODEL COMPARISON: TESTING A SINGLE PREDICTOR

Model Comparison:

- **Model A:** $Y_i = \beta_0 + \beta_1 \text{FeedbackCode1}_i + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$
- **Model C:** $Y_i = \beta_0 + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$

Q: What does the null hypothesis state?

- $H_0: \beta_1 = 0$

Q: How many parameters are in each of our models?

- PA = 3
- PC = 2

THE MODEL COMPARISON: TESTING A SINGLE PREDICTOR

Model Comparison:

- **Model A:** $Y_i = \beta_0 + \beta_1 \text{FeedbackCode1}_i + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$
- **Model C:** $Y_i = \beta_0 + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$

Which model does better at reducing SSE?

- It's getting quite long to calculate the SSE(C) and SSE(A) for each model "by hand," so let's use R

PERFORMING THE ANALYSIS IN R

- Import or set up the data
  - The IV (feedback) and DV (brainteasers) should each make up their own columns

```
> head(data)
  feedback brainteasers
1     None            4
2     None            3
3     None            4
4     None            5
5     None            5
6     None            2
```

- Check the measures types (and convert if needed)
  - The categorical IV should be a factor
    - The `factor` function allows you to also assign labels to the levels of the categorical IV
  - The DV should be numeric

```
data$feedback <- factor(data$feedback,
                        labels = c("None","Success","Failied"))
```

PERFORMING THE ANALYSIS IN R

- There are two methods for coding the categorical predictor variable:
  - **Method 1: Making new variables**
    - Make the contrast codes each a variable in the original data set

```
data$FeedbackCode1 <- c(rep(-1/3,8),rep(2/3,8),rep(-1/3,8))
data$FeedbackCode2 <- c(rep(1/2,8),rep(0,8),rep(-1/2,8))
```

- Fit the model
  - With this method of constructing the contrast codes, *each of the feedback contrast codes created* need to be manually included as predictors in the model

```
model2 <- lm(brainteasers ~ FeedbackCode1 + FeedbackCode2,
data = data)
```

PERFORMING THE ANALYSIS IN R

- Examine the model output
  - First, let's use anova()
  - When Method 1 of constructing the codes is used, this produces the familiar ANOVA summary table

SSR = 5.33

SSE(A) = 23.375

SSE(C) = SSR + SSE(A) = 28.71

```
> anova(model2)
Analysis of Variance Table

Response: brainteasers
               Df  Sum Sq Mean Sq F value  Pr(>F)
FeedbackCode1   1  5.3333  5.3333  4.7914 0.04004 *
FeedbackCode2   1  2.2500  2.2500  2.0214 0.16978
Residuals      21 23.3750  1.1131
```

$F(1, 21) = \frac{MSR}{MSE} = 4.79$

$p = .040$

PRE = 5.33/28.71 = 0.19

- FeedbackCode1 accounts for 19% more variability than a model without this predictor

PERFORMING THE ANALYSIS IN R

- There are two methods for coding the categorical predictor variable:
  - **Method 2: Assigning codes to contrast()**
    - Construct each code
    - Assign them both to the `contrasts()` function using `cbind`

```
FeedbackCode1 <- c(-1/3, 2/3, -1/3)
FeedbackCode2 <- c(1/2, 0, -1/2)

contrasts(data$feedback) <- cbind(FeedbackCode1, FeedbackCode2)
```

- Fit the model
  - With this method, feedback is used as the predictor in the model

```
model <- lm(brainteasers ~ feedback, data = data)
```

PERFORMING THE ANALYSIS IN R

- Examine the model output
  - Next, let's use summary() to examine the output

- The full estimated model is:

$$Y_i = 3.7083 + 1.00*\text{FeedbackCode1} + 0.75*\text{FeedbackCode2}$$

- Our parameter estimates are:
  - $b_0 = 3.71$
  - $b_1 = 1.00$
  - $b_2 = 0.75$

- Let's interpret the meaning of each of these

```
> summary(model)

Call:
lm(formula = brainteasers ~ feedback, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7500 -0.8125  0.0000  1.0000  1.6250

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.7083     0.2154  17.219 7.31e-14 ***
feedbackFeedbackCode1 1.0000     0.4568   2.189     0.04 *
feedbackFeedbackCode2 0.7500     0.5275   1.422     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 21 degrees of freedom
Multiple R-squared:  0.245,     Adjusted R-squared:  0.173
F-statistic: 3.406 on 2 and 21 DF,  p-value: 0.05233
```

INTERPRETING THE PARAMETER ESTIMATES

- Remember that our interpretation of the parameter estimates all depends on how the categorical predictor variable was coded

	None	Success	Failure
FeedbackCode1	-1/3	2/3	-1/3
FeedbackCode2	1/2	0	-1/2

- Interpret the parameter estimates
  - $b_0 = 3.71$
    - The mean of the group means
  - $b_1 = 1.00$
    - The mean of the success condition minus the average across the no feedback and failed conditions
  - $b_2 = 0.75$
    - The mean of the no feedback condition minus the mean of the failed condition

No Feedback	Success	Failure
4	4	2
3	6	2
4	5	2
5	4	3
5	6	4
2	4	4
4	3	3
3	3	4
$M_1 = 3.75$	$M_2 = 4.38$	$M_3 = 3.00$

TESTING A SINGLE PREDICTOR

- The row corresponding to **FeedbackCode1** is a test of whether a model including this predictor makes a significant improvement compared to a model without this predictor

- Is the difference between the mean of the success condition and the average of the no feedback and failure conditions significant?
  - Yes, $b_1 = 1.00$, $t(21) = 2.19$, $p = .040$
  - Children in the success condition completed significantly more brain teasers than the average number of brain teasers completed by children in the no feedback and failure conditions

```
> summary(model)

Call:
lm(formula = brainteasers ~ feedback, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7500 -0.8125  0.0000  1.0000  1.6250

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.7083     0.2154  17.219 7.31e-14 ***
feedbackFeedbackCode1  1.0000     0.4568   2.189     0.04 *
feedbackFeedbackCode2  0.7500     0.5275   1.422     0.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 21 degrees of freedom
Multiple R-squared:  0.245,     Adjusted R-squared:  0.173
F-statistic: 3.406 on 2 and 21 DF,  p-value: 0.05233
```

THE MODEL COMPARISON: TESTING MULTIPLE PREDICTORS

- We could also be interested in testing whether type of feedback matters *overall* for predicting how many brain teasers children were motivated to complete.

Model Comparison

- **Model A:** $Y_i = \beta_0 + \beta_1 \text{FeedbackCode1}_i + \beta_2 \text{FeedbackCode2}_i + \varepsilon_i$
- **Model C:** $Y_i = \beta_0 + \varepsilon_i$

Null hypothesis:

- H_0: $\beta_1 = \beta_2 = 0$

Parameters in each model:

- PA = 3
- PC = 1

TESTING MULTIPLE PREDICTORS

- We can examine whether the **full model** is significant by looking at the bottom of the summary() output:
  - Multiple $R^2 = 0.24$, meaning feedback accounted for approximately 24% of the variation in brain teaser scores
  - This amount of variability accounted for by the full model is not significant, $F(2, 21) = 3.41$, $p = .052$.

```
> summary(model)

Call:
lm(formula = brainteasers ~ feedback, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7500 -0.8125  0.0000  1.0000  1.6250

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.7083     0.2154  17.219 7.31e-14 ***
feedbackFeedbackCode1  1.0000     0.4568   2.189     0.04 *
feedbackFeedbackCode2  0.7500     0.5275   1.422     0.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 21 degrees of freedom
Multiple R-squared:  0.245,     Adjusted R-squared:  0.173
F-statistic: 3.406 on 2 and 21 DF,  p-value: 0.05233
```

TESTING MULTIPLE PREDICTORS

- We could also pass the model we ran earlier to the `anova()` function to get a test of the full model's significance
  - So long as we used Method 2 for contrast coding the categorical predictor

- Was feedback, overall, a significant predictor of the number of brain teasers children completed?
  - No, $F(2, 21) = 3.41$, $p = .052$, $R^2 = 0.24$.
  - Although feedback accounted for 24% of the variation in brain teaser scores, it was not a significant overall predictor.

```
> anova(model)
Analysis of Variance Table

Response: brainteasers
          Df  Sum Sq Mean Sq F value  Pr(>F)
feedback   2  7.5833  3.7917  3.4064 0.05233
Residuals 21 23.3750  1.1131
```

```
> etaSquared(model)
             eta.sq eta.sq.part
feedback 0.2449529   0.2449529
```

- *Note*: Notice that the overall predictor does not need to be significant in order for specific comparisons between group means to be significant.

WHAT IF WE HAD USED DUMMY CODES?

- To use **dummy codes**, the group that you wish to treat as the reference group receives a 0 across all codes
  - Each other group receives a 1 in one of the codes
  - Let's treat the No Feedback condition as the reference group

	None	Success	Failure
DummyCode1	**0**	**1**	**0**
DummyCode2	**0**	**0**	**1**

PERFORMING THE ANALYSIS IN R

- Import or set up the data
  - The IV (feedback) and DV (brainteasers) should each make up their own columns

```
> head(data)
  feedback brainteasers
1     None            4
2     None            3
3     None            4
4     None            5
5     None            5
6     None            2
```

- Check the measures types (and convert if needed)
  - The categorical IV should be a factor
    - The `factor` function allows you to also assign labels to the levels of the categorical IV
  - The DV should be numeric

```
data$feedback <- factor(data$feedback,
                        labels = c("None","Success","Failied"))
```

PERFORMING THE ANALYSIS IN R

- There are two methods for coding the categorical predictor variable:
  - **Method 1: Making new variables**
    - Make the dummy codes each a variable in the original data set

```
data$DummyCode1 <- c(rep(0,8),rep(1,8),rep(0,8))
data$DummyCode2 <- c(rep(0,8),rep(0,8),rep(1,8))
```

- Fit the model
  - With this method of constructing the contrast codes, *each of the feedback contrast codes created* need to be manually included as predictors in the model

```
model2 <- lm(brainteasers ~ DummyCode1 + DummyCode2,
data = data)
```

PERFORMING THE ANALYSIS IN R

- Examine the model output
  - First, using anova()
  - Notice the ANOVA summary table has not changed as a result of the new coding scheme

```
> anova(model2)
Analysis of Variance Table

Response: brainteasers
           Df  Sum Sq Mean Sq F value  Pr(>F)
DummyCode1  1  5.3333  5.3333  4.7914 0.04004
DummyCode2  1  2.2500  2.2500  2.0214 0.16978
Residuals  21 23.3750  1.1131
```

PERFORMING THE ANALYSIS IN R

- There are two methods for coding the categorical predictor variable:
  - **Method 2: Assigning codes to contrast()**
    - Construct each code
    - Assign them both to the `contrasts()` function using `cbind`

```
DummyCode1 <- c(0,1,0)
DummyCode2 <- c(0,0,1)

contrasts(data$feedback) <- cbind(DummyCode1, DummyCode2)
```

- Fit the model
  - With this method, feedback is used as the predictor in the model

```
model <- lm(brainteasers ~ feedback, data = data)
```

PERFORMING THE ANALYSIS IN R

- Examine the model output
  - Next, let's use summary() to examine the output

```
> summary(model)

Call:
lm(formula = brainteasers ~ feedback, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7500 -0.8125  0.0000  1.0000  1.6250

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.7500     0.3730  10.053 1.76e-09 ***
feedbackDummyCode1 0.6250     0.5275   1.185    0.249
feedbackDummyCode2 -0.7500    0.5275  -1.422    0.170
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 21 degrees of freedom
Multiple R-squared:  0.245,     Adjusted R-squared:  0.173
F-statistic: 3.406 on 2 and 21 DF,  p-value: 0.05233
```

- The full estimated model has changed. Now, it's:

$Y_i = 3.75 + 0.625*\text{FeedbackCode1} - 0.75*\text{FeedbackCode2}$

- Our parameter estimates are:
  - $b_0 = 3.75$
  - $b_1 = 0.625$
  - $b_2 = -0.75$

- Let's interpret the meaning of each of these

INTERPRETING THE PARAMETER ESTIMATES

- Interpret the parameter estimates
  - $b_0 = 3.75$
    - The mean of the No Feedback condition
  - $b_1 = 0.625$
    - The mean of the Success minus the mean of the No Feedback condition
  - $b_2 = -0.75$
    - The mean of the Failure minus the mean of the No Feedback condition

	None	Success	Failure
FeedbackCode1	0	1	0
FeedbackCode2	0	0	1

No Feedback	Success	Failure
4	4	2
3	6	2
4	5	2
5	4	3
5	6	4
2	4	4
4	3	3
3	3	4
$M_1 = 3.75$	$M_2 = 4.38$	$M_3 = 3.00$

PERFORMING THE ANALYSIS IN R

- Testing a single predictor:
  - Examine the significance of an individual parameter estimate of interest in the model
    - It is very important to be aware of what each parameter estimate is representing because the *p*-value on that row corresponds to whether that comparison is significantly different from zero

```
> summary(model)

Call:
lm(formula = brainteasers ~ feedback, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7500 -0.8125  0.0000  1.0000  1.6250

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.7500     0.3730  10.053 1.76e-09 ***
feedbackDummyCode1  0.6250     0.5275   1.185    0.249
feedbackDummyCode2 -0.7500     0.5275  -1.422    0.170
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 21 degrees of freedom
Multiple R-squared:  0.245,     Adjusted R-squared:  0.173
F-statistic: 3.406 on 2 and 21 DF,  p-value: 0.05233
```

- Testing the full model (all of the predictors together):
  - Examine the Multiple R^2 value and its accompanying *F*-statistic
  - Recall that $R^2 = \frac{SSR}{SS_{Total}}$ which is equal to PRE when Model C is the one parameter model.