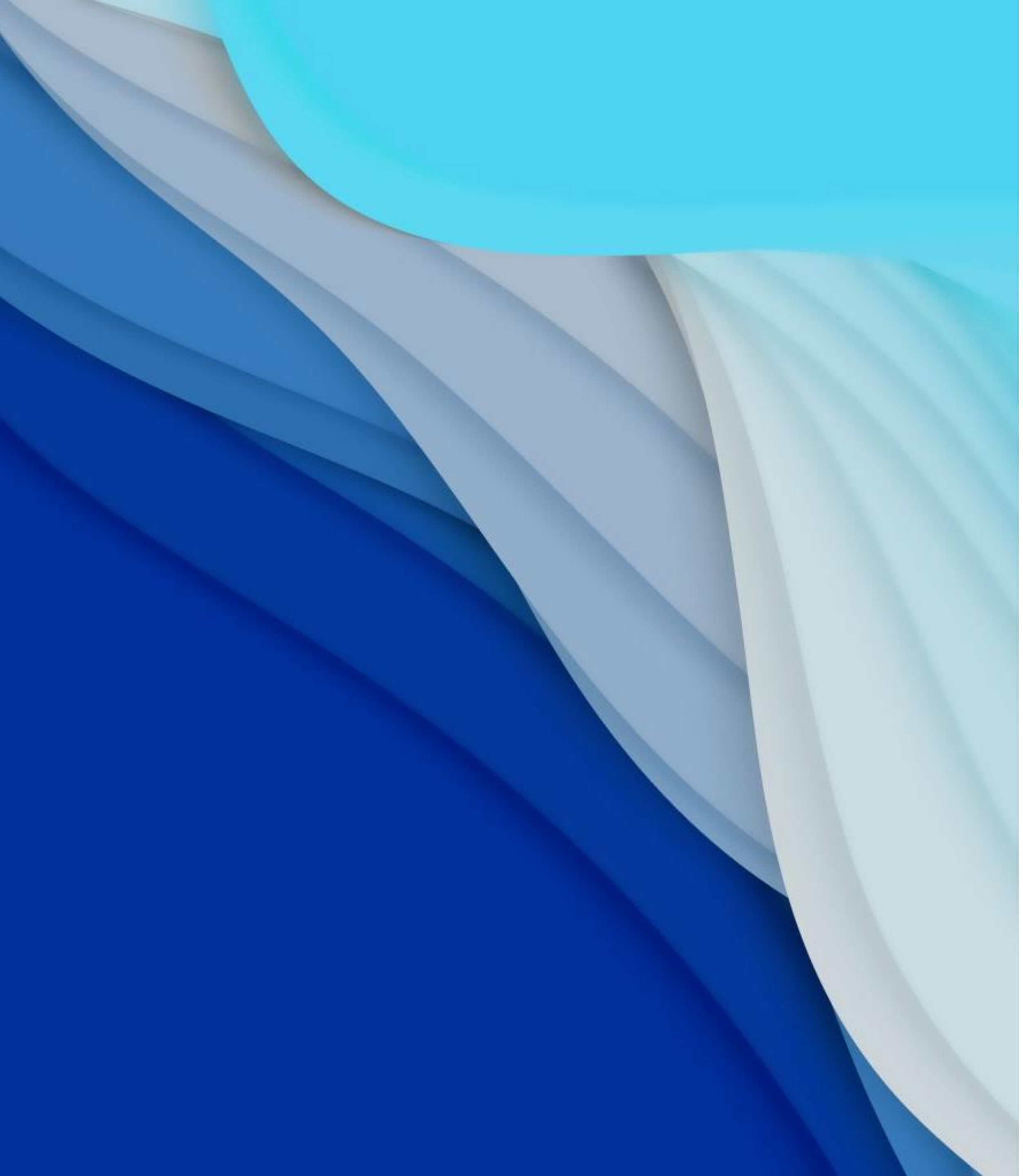


LINEAR REGRESSION WITH A SINGLE CATEGORICAL PREDICTOR WITH 2 INDEPENDENT LEVELS

aka, Independent Samples t-Test

Review of Models Covered

Zero Parameter Model

- $Y_i = B_0 + \varepsilon_i$
- B_0 is an a priori numerical value based on prior research

One Parameter Model

- $Y_i = \beta_0 + \varepsilon_i$
- β_0 is a single numerical value estimated from the data, b_0
- "Intercept-only model"

Now, let's add a predictor to the model.

Model with a Single Categorical Predictor

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

- Y_i is each participant's score on the outcome variable
- X_i is the predictor variable
- β_0 is the model intercept
- β_1 is the model's slope
- ε_i is the error (or residual) for each participant

Estimate of the Model from Sample Data: $Y_i = b_0 + b_1 X_i + e_i$

Example of a Model with a Single Categorical Predictor

A researcher is interested in whether the presence of music, versus no music, in the background has an effect on people's pain tolerance. The researcher randomly assigns eight people to either sit in silence or sit in a room playing music while they complete a pain tolerance task (the number of minutes participants can leave their hands in an ice bath). The participants' scores are shown below:

Silence
3
3
4
2

M = 3.00, *SD* = 0.82

Music
7
3
6
4

M = 5.00, *SD* = 1.83

The researcher wants to test whether the environment (silence vs music) has a significant effect on pain tolerance.

Testing the Hypothesis Using a Model Comparison

Model representing the null hypothesis (i.e., if environment **does not predict** pain tolerance):

$$\text{Model C: } Y_i = \beta_0 + \varepsilon_i$$

Model representing the alternative hypothesis (i.e., if environment **does** predict pain tolerance)

$$\text{Model A: } Y_i = \beta_0 + \beta_1 \text{Environment}_i + \varepsilon_i$$

How many parameters are in each model?

- PC = 1
- PA = 2

The Null & Alternative Hypotheses

Model Comparison:

Model C: $Y_i = \beta_0 + \varepsilon_i$ PC = 1

Model A: $Y_i = \beta_0 + \beta_1 \text{Environment}_i + \varepsilon_i$ PA = 2

- What is the null hypothesis that we are testing?

Null & Alternative Hypotheses:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Coding Categorical Predictors

- In order to be included in a regression model, categorical predictors must be **coded numerically**
  - The number of codes you must construct is equal to $m - 1$, where m is the number of levels of the categorical predictor

- The levels of our categorical predictor, Environment, are:
  - Silence
  - Music

- What numerical values should we choose to represent each level?
  - The way you choose to code a categorical predictor will affect how you should interpret the output of conducting your regression analysis

Option 1: Dummy Coding

Purpose: To compare each group to a reference (or control) group.

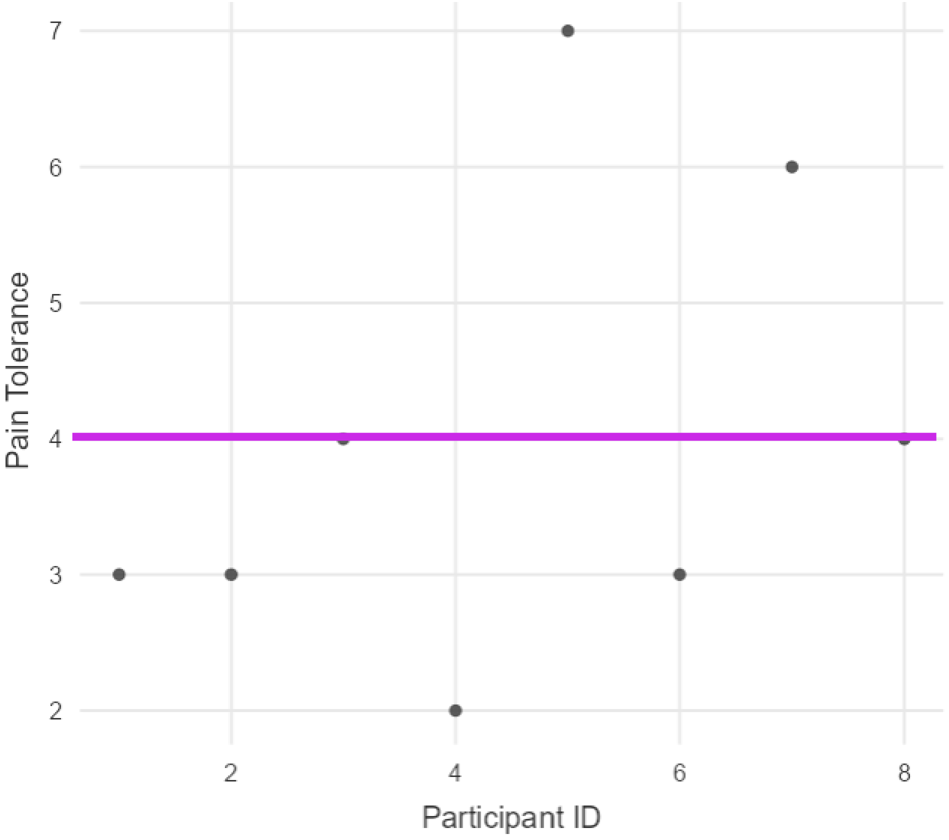
How to use: Assign a 0 to the group you want to treat as the "reference" group and a 1 to the other group.

- If there are more than two groups, then the reference group will be assigned a 0 in all of the codes, while each of the other groups will be assigned a 1 in only one of the codes (more to come on this)

Dummy coding for our example:

- Silence = 0, Music = 1

Fit Model C to the Data



$b_0 = 4$

Model C: $Y_i = \beta_0 + \varepsilon_i$

- Remember when predicting data from a single numerical value, the sample mean does the best at minimizing SSE
  - The mean for *all* of the participants' pain tolerance scores (also called the *grand mean*) is 4.00

- To evaluate the fit of this model, we need to calculate SSE for Model C
  - $\text{SSE} = \Sigma(Y_i - Y'_i)^2$

Evaluate Fit of Model C

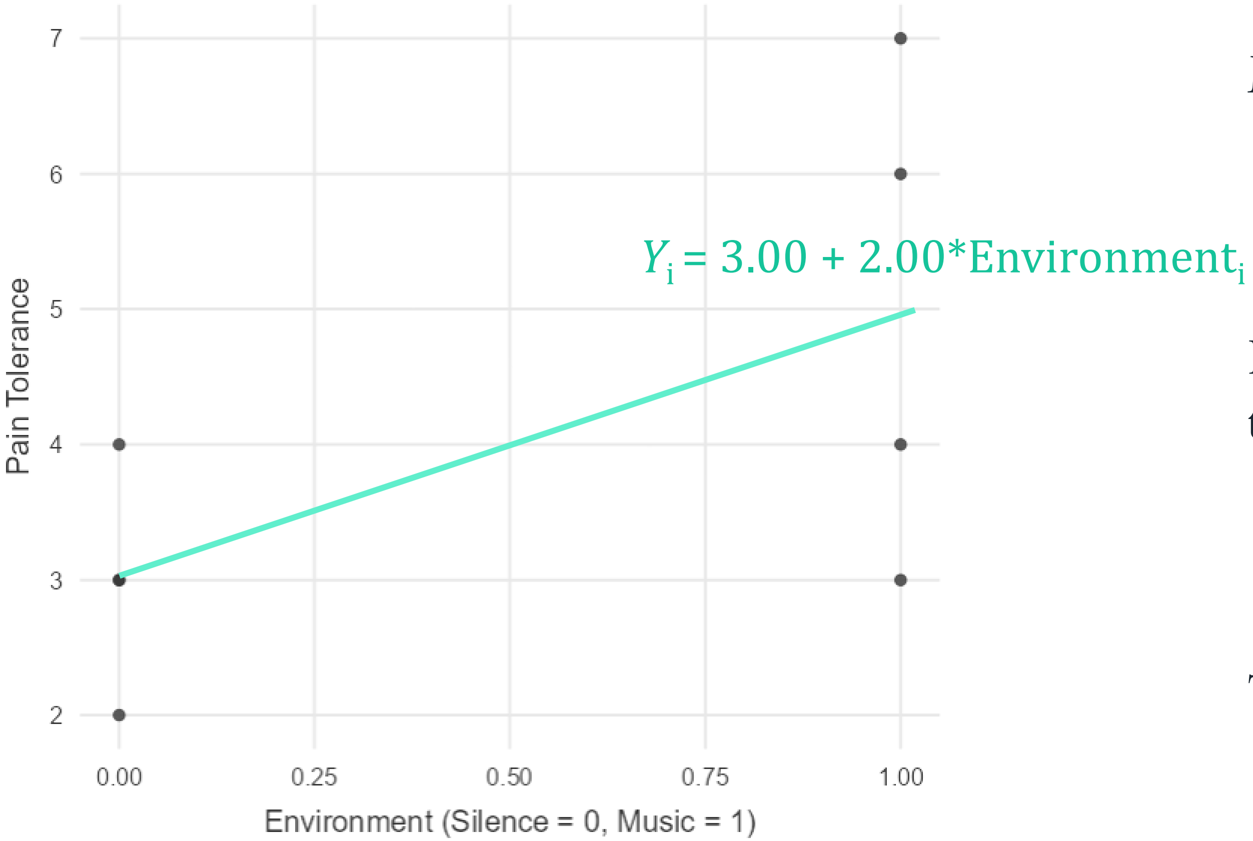
Participant ID	Pain Tolerance (Y_i)	Value Predicted by Model (Y’)	$e_i = Y_i - Y'_i$	$(e_i)^2 = (Y_i - Y'_i)^2$
1	3	4	-1	1
2	3	4	-1	1
3	4	4	0	0
4	2	4	-2	4
5	7	4	3	9
6	3	4	-1	1
7	6	4	2	4
8	4	4	0	0

$SSE = \Sigma(Y_i - Y'_i)^2$

SSE for Model C:

- SSE(C) = 20

Fit Model A to the Data



Model A: $Y_i = \beta_0 + \beta_1\text{Environment}_i + \varepsilon_i$

- We have to estimate this model from the data: $Y_i = b_0 + b_1\text{Environment}_i + e_i$

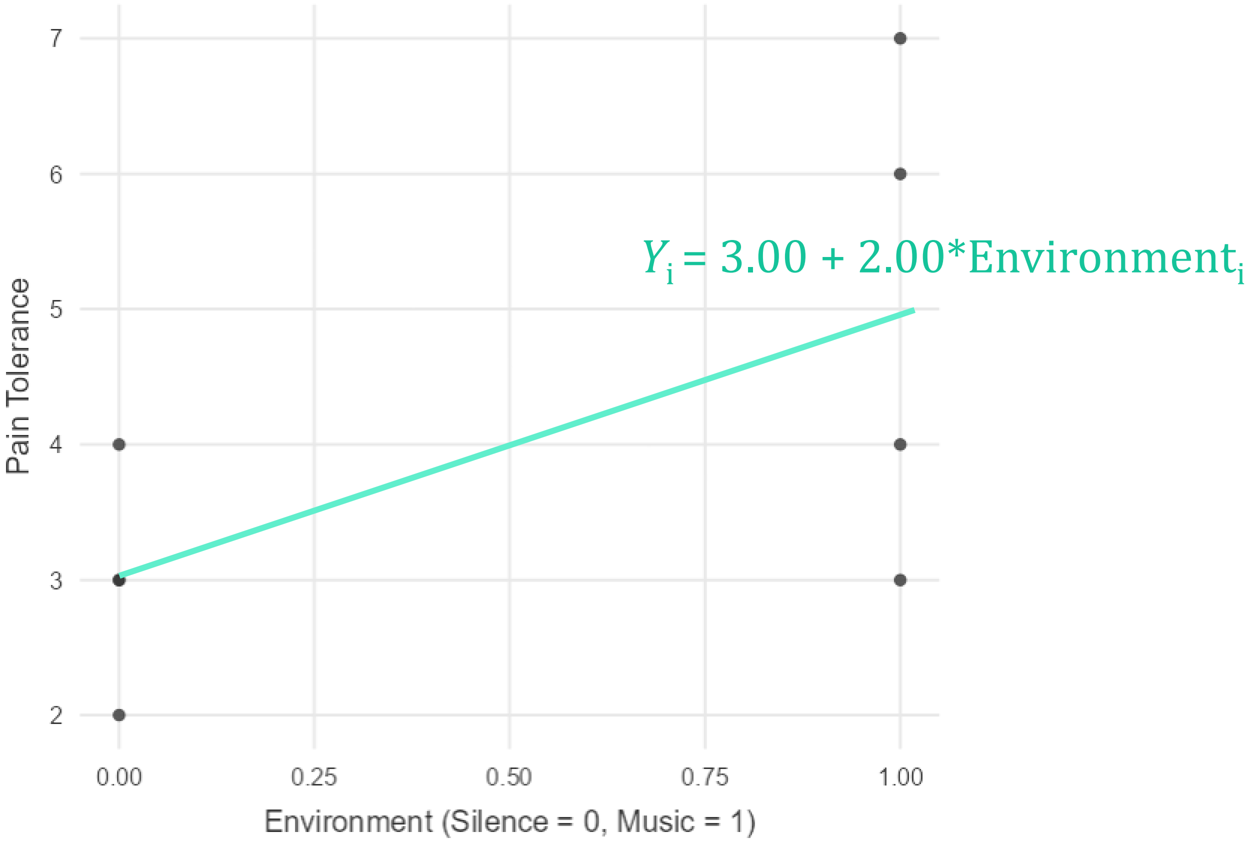
Find the best-fitting line, meaning the line that best minimizes the SSE

- The line that best minimizes SSE is the one that passes through the means of each group

The estimate of the model is:

$$Y_i = 3.00 + 2.00\text{*Environment}_i$$

Interpreting Parameter Estimates



The estimate of the model is:

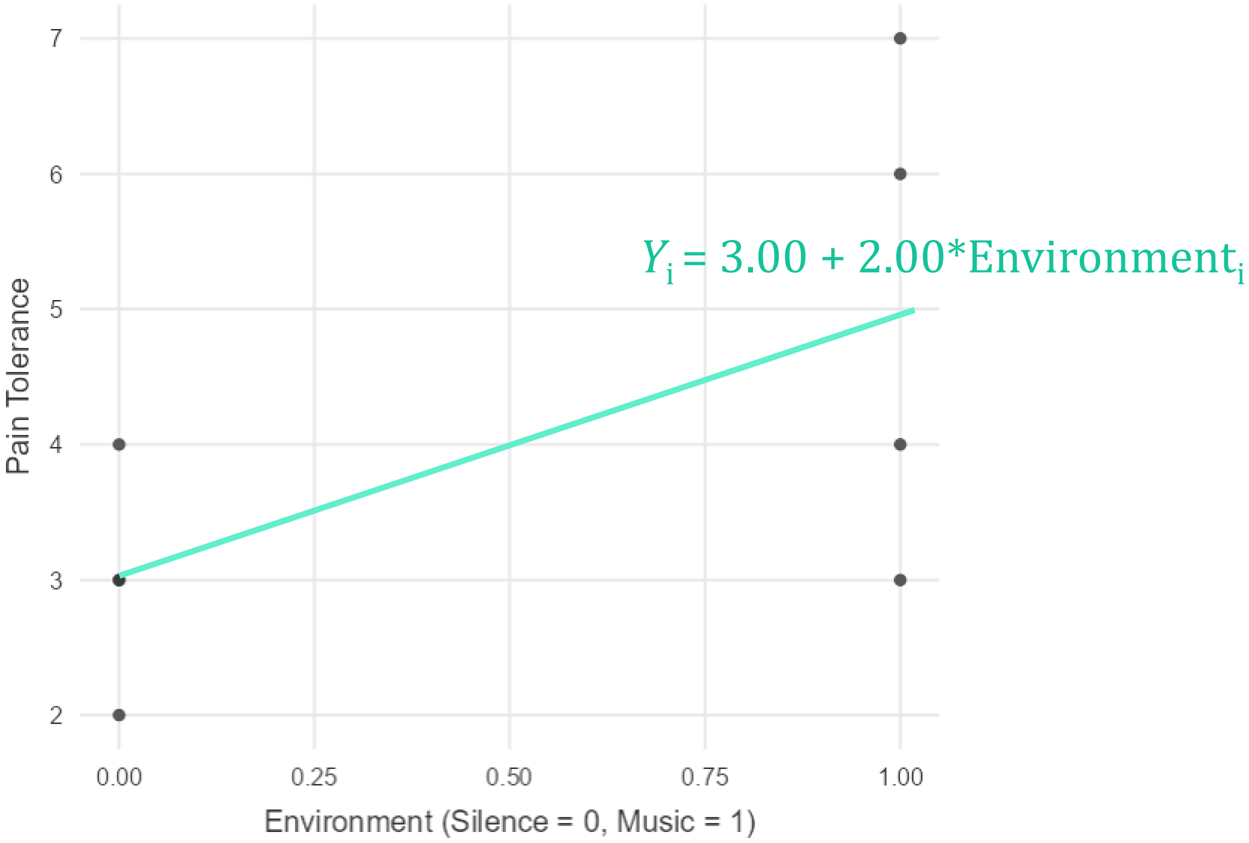
$$Y_i = 3.00 + 2.00{\star}\text{Environment}_i$$

- b_0 is the y-intercept
  - The value predicted by the model when the predictor(s) are equal to 0

Our example: $b_0 = 3$

- Since we used dummy coding, Environment = 0 means we are in the Silence condition.
- The y-intercept, in this case, is the mean for the Silence condition.

Interpreting Parameter Estimates



The estimate of the model is:

$$Y_i = 3.00 + 2.00\star\text{Environment}_i$$

- b_1 is the slope
  - The predicted change in Y per 1-unit change in our predictor(s)

Our example: $b_1 = 2$

- Since we used dummy coding, we coded the music condition to be 1 unit apart from the silence condition.
- Therefore, the slope is the difference between the mean of the music condition ($M = 5$) and the mean of the silence condition ($M = 3$)
- A test of the significance of the slope is a test of whether there is a significant difference between the means of each group

Evaluate Fit of Model A

Participant ID	Pain Tolerance (Y_i)	Environment (Silence = 0, Music = 1)	$Y' = 3 + 2\star\text{Environment}$ (Value Predicted by Model)	$e_i = Y_i - Y'_i$	$(e_i)^2 = (Y_i - Y'_i)^2$
1	3	0	$3 + 2\star 0 = 3$	3-3 = 0	0
2	3	0	3	0	0
3	4	0	3	1	1
4	2	0	3	-1	1
5	7	1	$3 + 2\star 1 = 5$	7-5 = 2	4
6	3	1	5	-2	4
7	6	1	5	1	1
8	4	1	5	-1	1

$SSE = \Sigma(Y_i - Y'_i)^2$

SSE for Model A:

- SSE(A) = 12

How much better did Model A do compared to Model C?

- $PRE = \frac{SSE(C) - SSE(A)}{SSE(C)} = \frac{SSR}{SSE(C)} = \frac{(20 - 12)}{20} = \frac{8}{20} = 0.40$

Model A accounts for 40% more error compared to Model C.

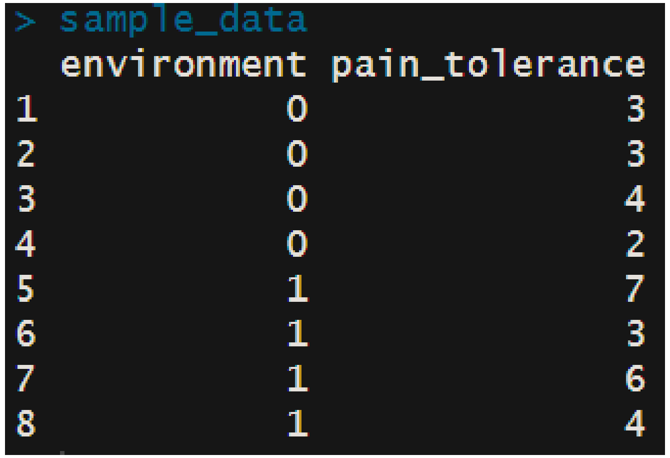
To obtain the *p*-value corresponding to our *F*-statistic, we will run the analysis in R.

ANOVA Summary Table

Source	SS	df	MS	*F*	PRE	*p*
Reduced	8	PA–PC = 1	MSR = SSR/$\text{df}_{\text{Reduce}}$ = 8.00	$F = \frac{MSR}{MSE}$ $F = 4.00$	0.40	?
Model A (Error)	12	n–PA = 6	MSE = SSE(A)/df_{Error} = 2.00			
Model C (Total)	20	n–PC = 7				

Performing the Analysis in R

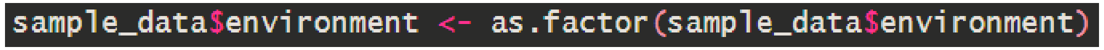
Set up the data so that Environment, the independent variable, is one column and Pain Tolerance, the dependent variable, is a second column.



```
> sample_data
  environment pain_tolerance
1           0              3
2           0              3
3           0              4
4           0              2
5           1              7
6           1              3
7           1              6
8           1              4
```

Environment should be a **factor**, and pain_tolerance should be **numeric**.

- Use as.numeric() to convert a variable to numeric
- Use as.factor() to convert a variable to a factor

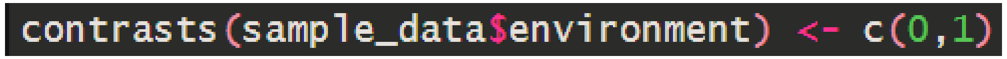


```
sample_data$environment <- as.factor(sample_data$environment)
```

Performing the Analysis in R

Now that we have converted our categorical predictor, Environment, into a factor, we can code it:

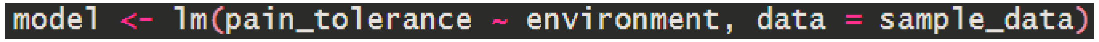
- You can pass a variable to the contrasts() function to see how it is currently coded
- Then, assign desired codes. We're using dummy codes for now.



```
contrasts(sample_data$environment) <- c(0,1)
```

Perform the model predicting pain tolerance scores from Environment as a predictor:

- The intercept is automatically estimated in addition to the slope when a predictor is included in the model using `lm`

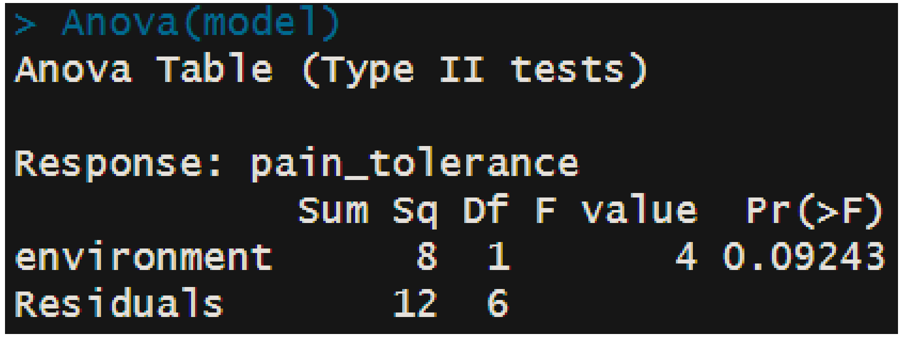


```
model <- lm(pain_tolerance ~ environment, data = sample_data)
```

Performing the Analysis in R

Look at the model output using:

- `Anova(model)` from the car package, and
- summary(model)



```
> Anova(model)
Anova Table (Type II tests)

Response: pain_tolerance
            Sum Sq Df F value  Pr(>F)
environment      8  1       4 0.09243
Residuals       12  6
```

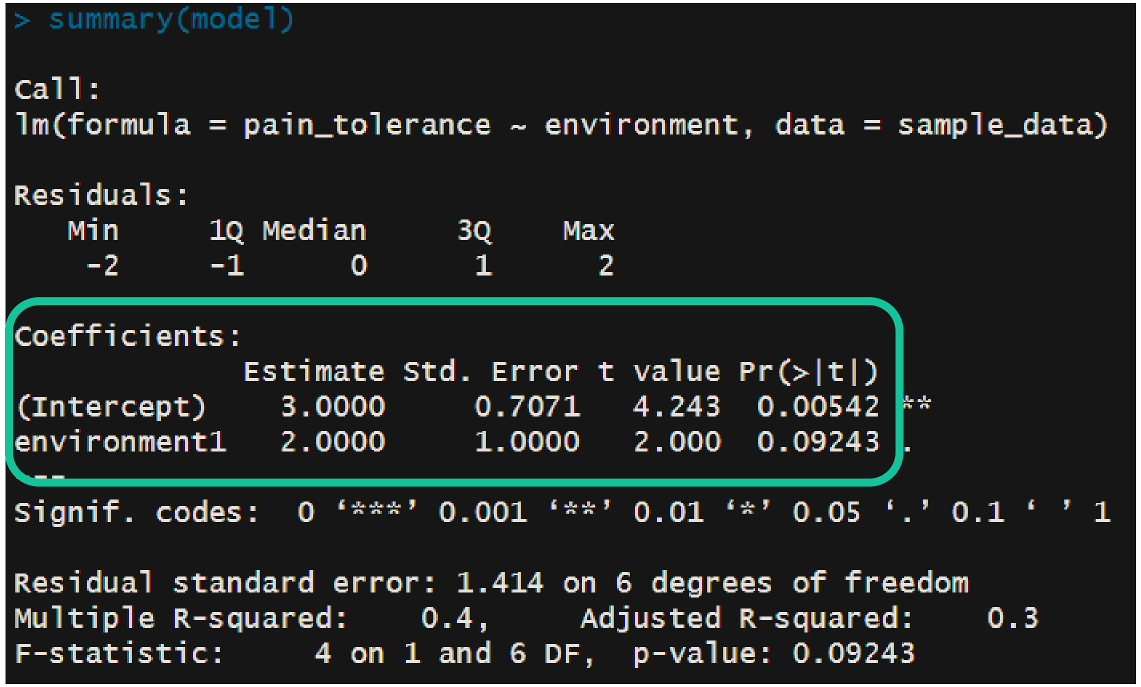
Anova() output:

Source	SS	df	MS	*F*	PRE	*p*
Reduced	8	1	8	4.00	0.40	.092
Model A (Error)	12	6	2			
Model C (Total)	20	7				

Performing the Analysis in R

Look at the model output using:

- `Anova(model)` from the car package, and
- summary(model)



```
> summary(model)

Call:
lm(formula = pain_tolerance ~ environment, data = sample_data)

Residuals:
   Min     1Q Median     3Q    Max 
    -2     -1      0      1      2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)    3.0000     0.7071   4.243  0.00542 **
environment1   2.0000     1.0000   2.000  0.09243 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 6 degrees of freedom
Multiple R-squared:    0.4,	Adjusted R-squared:    0.3 
F-statistic:     4 on 1 and 6 DF,  p-value: 0.09243
```

summary() output:

- Estimate
  - Intercept = b_0
  - Environment = b_1
- t value
  - $t = \frac{b_1}{SE_{b1}}$

- Std. error

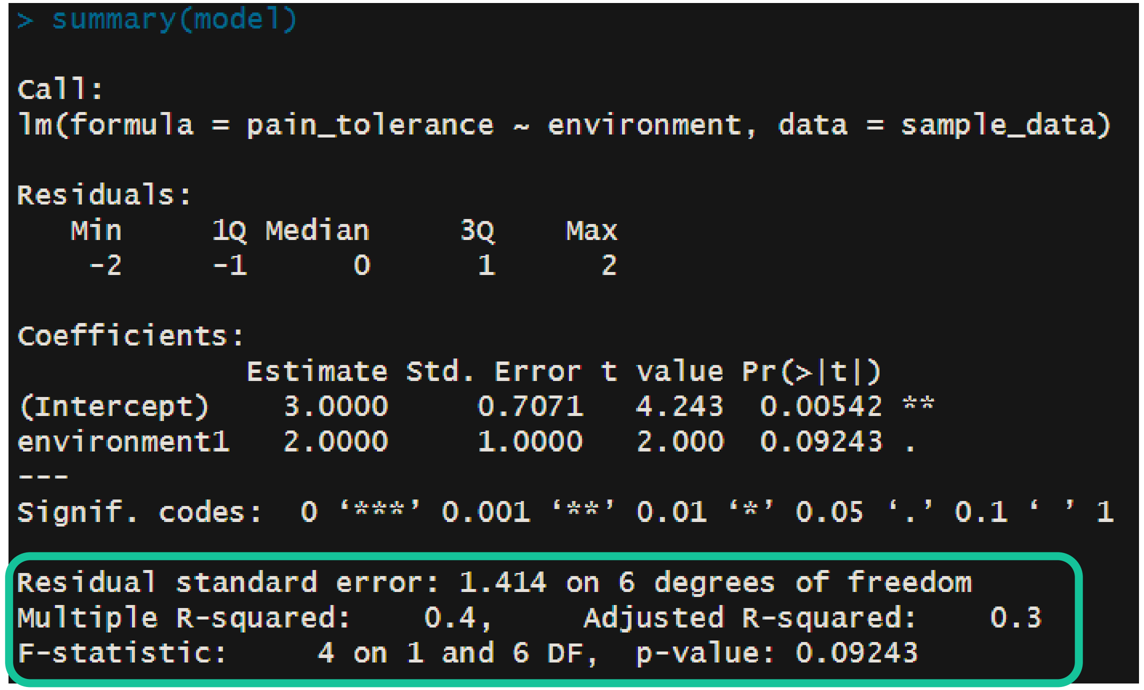
$$SE_{b1} = \sqrt{\frac{MSE}{SSx}}$$

- Pr (>|t|) is the p-value
  - If $p < .05$ → Significant
  - If $p > .05$ → Non-significant

Performing the Analysis in R

Look at the model output using:

- \`Anova(model)\` from the car package, and
- summary(model)



```
> summary(model)

Call:
lm(formula = pain_tolerance ~ environment, data = sample_data)

Residuals:
   Min     1Q Median     3Q    Max 
    -2     -1      0      1      2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)    3.0000     0.7071   4.243  0.00542 **
environment1   2.0000     1.0000   2.000  0.09243 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 6 degrees of freedom
Multiple R-squared:    0.4,	Adjusted R-squared:    0.3 
F-statistic:     4 on 1 and 6 DF,  p-value: 0.09243
```

summary() output:

- Multiple R-squared: When we are testing a 1-parameter difference between Modal A and Model C (PA-PC = 1), and Model C is the one parameter model ($Y_i = \beta_0 + \varepsilon_i$), then Multiple R-squared = PRE
  - Biased estimate of the true proportional reduction in error (tends to overestimate)

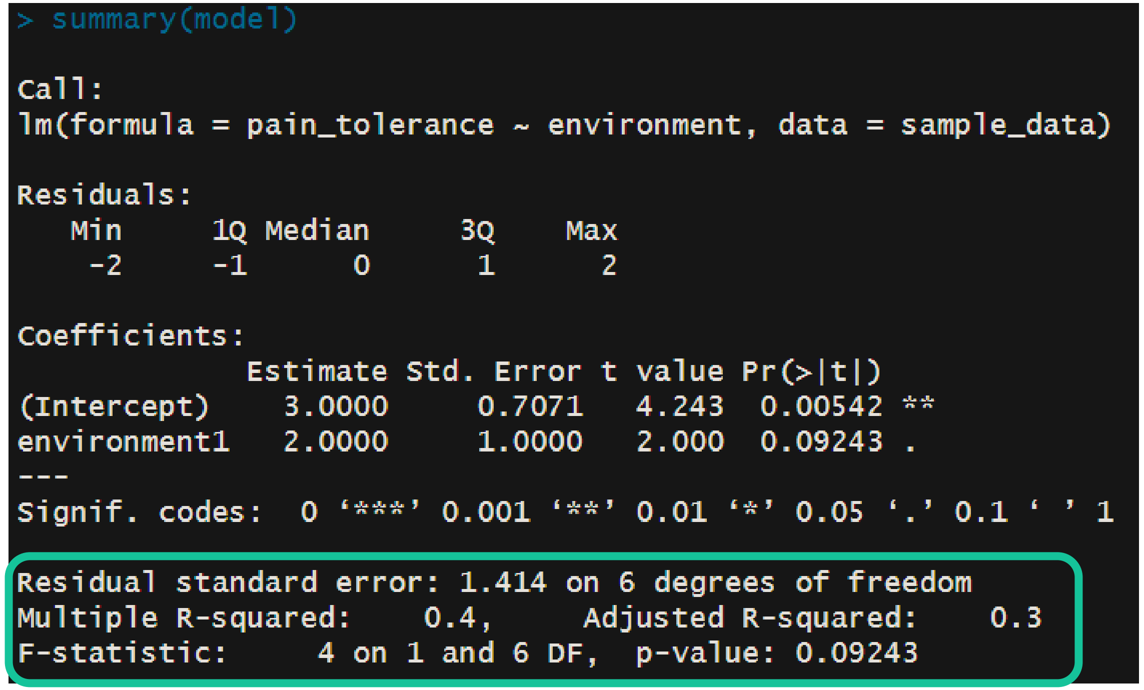
- Adjusted R-squared = $1 – (1 – R^2)[\frac{n-PC}{n-PA}]$
  - For our example: $1 – ((1\text{-}0.4)\star(\frac{(8-1)}{(8-2)})) = 0.3$
  - The penalty to R^2 increases as the number of parameters added to Model A increases

- $F(1, 6) = 4.00, p = .092$
  - The F-test for the overall model is the same as the F-test for our model comparison when Model C is the one parameter model

Performing the Analysis in R

Look at the model output using:

- `Anova(model)` from the car package, and
- summary(model)



```
> summary(model)

Call:
lm(formula = pain_tolerance ~ environment, data = sample_data)

Residuals:
   Min     1Q Median     3Q    Max 
    -2     -1      0      1      2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)    3.0000     0.7071   4.243  0.00542 **
environment1   2.0000     1.0000   2.000  0.09243 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 6 degrees of freedom
Multiple R-squared:    0.4,	Adjusted R-squared:    0.3 
F-statistic:     4 on 1 and 6 DF,  p-value: 0.09243
```

summary() output:

- Residual standard error
  - The standard deviation of the errors (aka, the residuals) = $\sqrt{\frac{SSE(A)}{df_{Error}}}$
    - For our example: $\sqrt{\frac{12}{6}} = 1.414$
    - A measure of how much the errors deviate from the model on average

Example of a Model with a Single Categorical Predictor

A researcher is interested in whether the presence of music, versus no music, in the background has an effect on people's pain tolerance. The researcher randomly assigns eight people to either sit in silence or sit in a room playing music while they complete a pain tolerance task (the number of minutes participants can leave their hands in an ice bath). The participants' scores are shown below:

Silence
3
3
4
2

$M = 3.00$, $SD = 0.82$

Music
7
3
6
4

$M = 5.00$, $SD = 1.83$

The researcher wants to test whether the environment (silence vs music) has a significant effect on pain tolerance.

The Null & Alternative Hypotheses

Model Comparison:

Model C: $Y_i = \beta_0 + \varepsilon_i$ PC = 1

Model A: $Y_i = \beta_0 + \beta_1 \text{Environment}_i + \varepsilon_i$ PA = 2

Null & Alternative Hypotheses:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Option 2: Contrast Coding **(Recommended Coding Method)**

Purpose: To test specific hypotheses about differences between means.

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. (This rule is relevant once you have **more than one** contrast code. More to come on this.).

- **Recommended for ease of interpretation, but not required**: 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]).

A major benefit of contrast codes is that they are **orthogonal**.

- This means the contrast codes are *independent of each other*, which means they do not pose issues related to redundancy between predictors (this will be relevant when there are 2+ predictors in the model).

Option 2: Contrast Coding

Contrast Coding for Our Example:

- Number of codes
  - $m - 1$ = 2-1 = 1 code

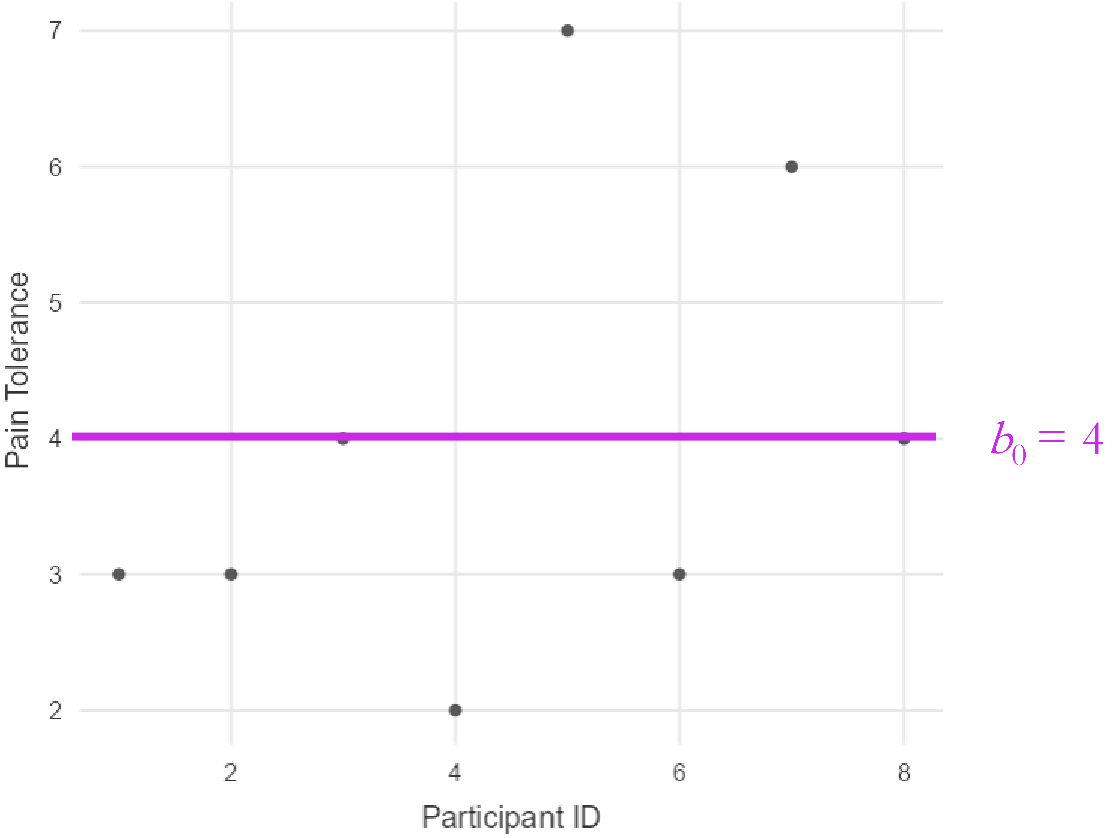
Rules of Contrast Coding:

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, but not required: 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]).

	Silence	**Music**
Contrast Code 1	-1/2	+1/2

Fit Model C to the Data

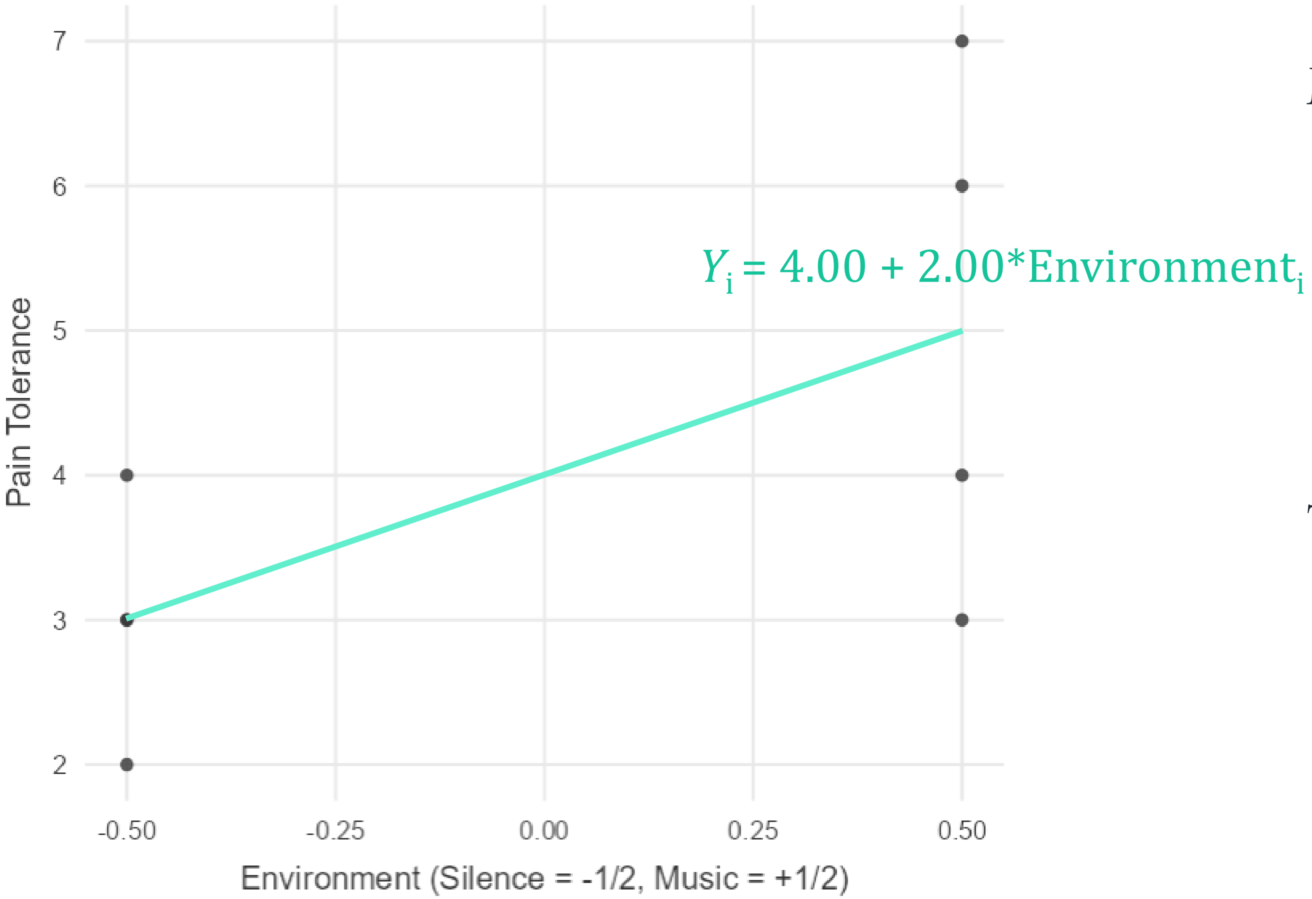


Model C: $Y_i = \beta_0 + \varepsilon_i$

- Model C has not changed. The best single value estimate of the data is still 4.00, the grand mean
  - b0 = 4.00

- SSE(C) = $\Sigma(Y_i - Y'_i)^2 = 20$
  - See table above for step-by-step calculation

Fit Model A to the Data



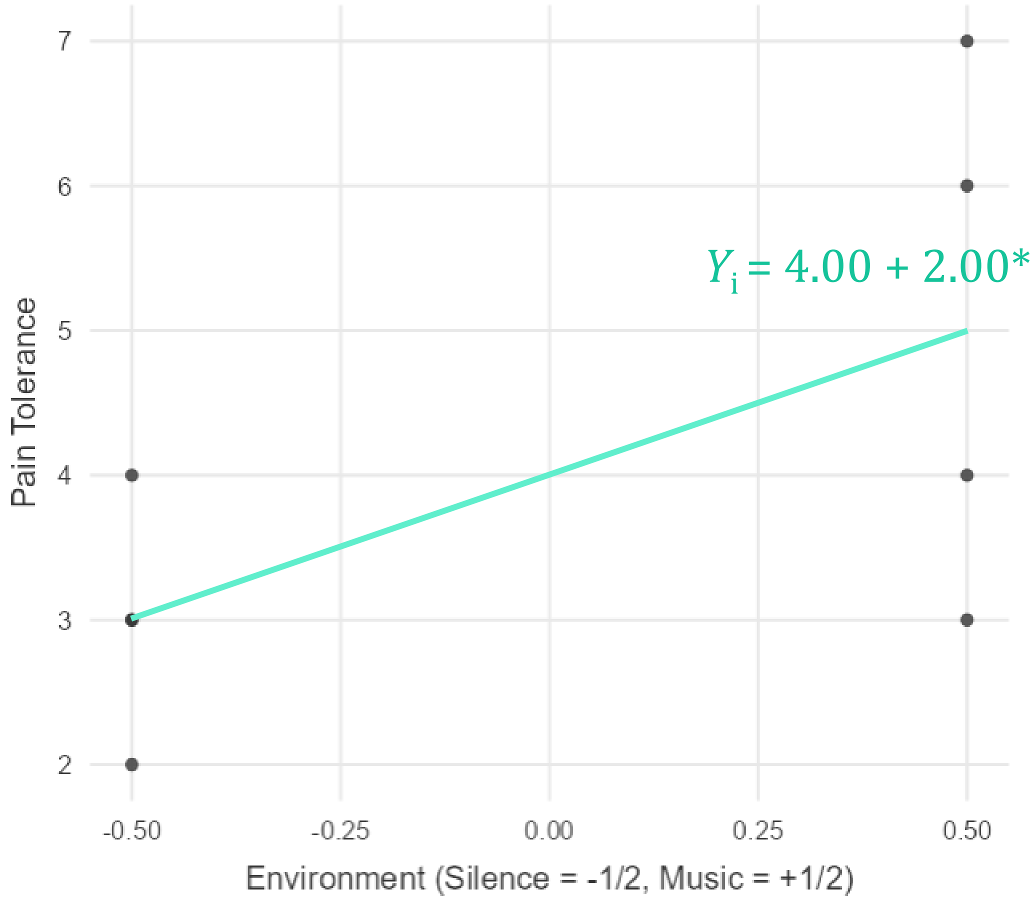
Model A: $Y_i = \beta_0 + \beta_1 \text{Environment}_i + \varepsilon_i$

- We have to estimate this model from the data: $Y_i = b_0 + b_1 \text{Environment}_i + e_i$
- The line that best minimizes SSE is the one that passes through the means of each group

The estimate of the model is:

$$Y_i = 4.00 + 2.00{\star}\text{Environment}_i$$

Interpreting Parameter Estimates



$Y_i = 4.00 + 2.00$*Environment$_i$

The estimate of the model is:

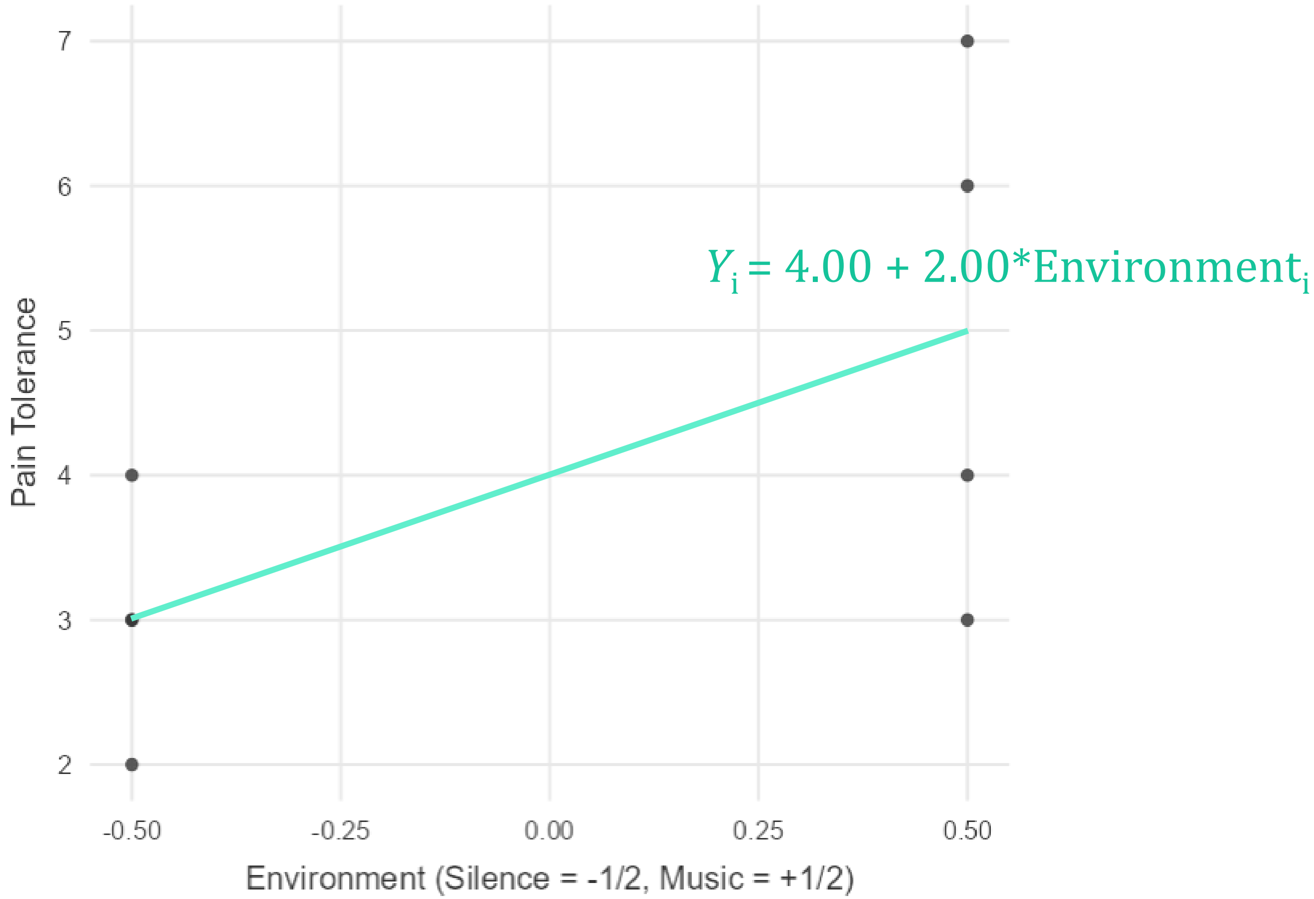
$$Y_i = 4.00 + 2.00{\star}\text{Environment}_i$$

- b_0 is the y-intercept
  - The value predicted by the model when the predictor(s) are equal to 0

Our example: $b_0 = 4$

- The y-intercept, in this case, is the mean of the group means
- AND the grand mean (because group sizes are equal)

Interpreting Parameter Estimates



The estimate of the model is:

$$Y_i = 4.00 + 2.00 \star \text{Environment}_i$$

- b_1 is the slope
  - The predicted change in Y per 1-unit change in our predictor(s)

Our example: $b_1 = 2$

- We coded the groups to be 1-unit apart (-1/2 = silence, +1/2 = music)
- Therefore, the slope is the difference between the mean of the music condition ($M = 5$) and the mean of the silence condition ($M = 3$)
- A test of the significance of the slope is a test of whether there is a significant difference between the means of each group

Evaluate Fit of Model A

Participant ID	Pain Tolerance (Y_i)	Environment (Silence = -1/2, Music = +1/2)	Y’ = 4 + 2★Environment (Value Predicted by Model)	$e_i = Y_i - Y'_i$	$(e_i)^2 = (Y_i - Y'_i)^2$
1	3	-1/2	4 + 2★(-1/2) = 3	3-3 = 0	0
2	3	-1/2	3	0	0
3	4	-1/2	3	1	1
4	2	-1/2	3	-1	1
5	7	1/2	4 + 2★(1/2) = 5	7-5 = 2	4
6	3	1/2	5	-2	4
7	6	1/2	5	1	1
8	4	1/2	5	-1	1

How do you think the SSE(A) will be affected by changing the coding scheme?

SSE for Model A:

- SSE(A) = 12

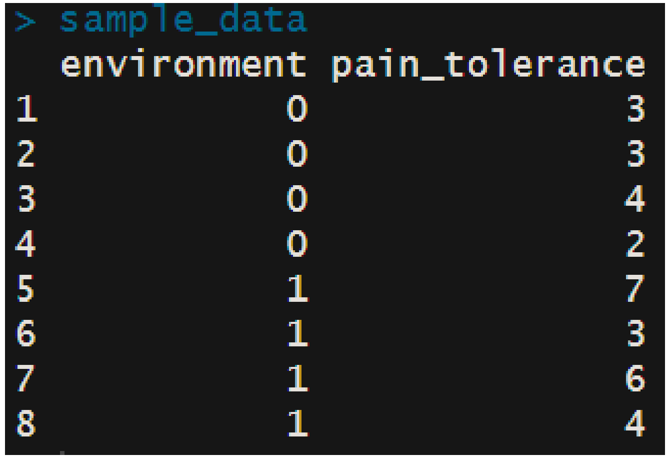
ANOVA Summary Table

To obtain the *p*-value corresponding to our *F*-statistic, we will run the analysis in R.

Source	SS	df	MS	*F*	PRE	*p*
Reduced	8	PA–PC = 1	MSR = SSR/df$_{\text{Reduce}}$ = 8.00	$F = \frac{MSR}{MSE}$ $F = 4.00$	0.40	?
Model A (Error)	12	n–PA = 6	MSE = SSE(A)/df$_{\text{Error}}$ = 2.00			
Model C (Total)	20	n–PC = 7				

Performing the Analysis in R

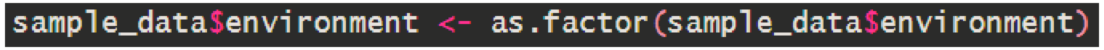
Set up the data so that Environment, the independent variable, is one column and Pain Tolerance, the dependent variable, is a second column.



```
> sample_data
  environment pain_tolerance
1           0              3
2           0              3
3           0              4
4           0              2
5           1              7
6           1              3
7           1              6
8           1              4
```

Environment should be a **factor**, and pain_tolerance should be **numeric**.

- Use as.numeric() to convert a variable to numeric
- Use as.factor() to convert a variable to a factor

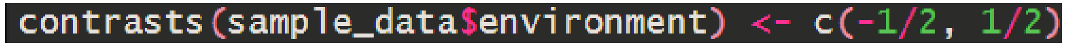


```
sample_data$environment <- as.factor(sample_data$environment)
```

Performing the Analysis in R

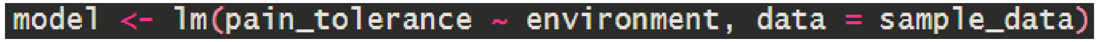
Contrast code the predictor by assigning the desired values to the contrasts() function in R:

- We're using contrast codes: -1/2 = silence, +1/2 = music.



```
contrasts(sample_data$environment) <- c(-1/2, 1/2)
```

Perform the model predicting pain tolerance scores from Environment as a predictor:



```
model <- lm(pain_tolerance ~ environment, data = sample_data)
```

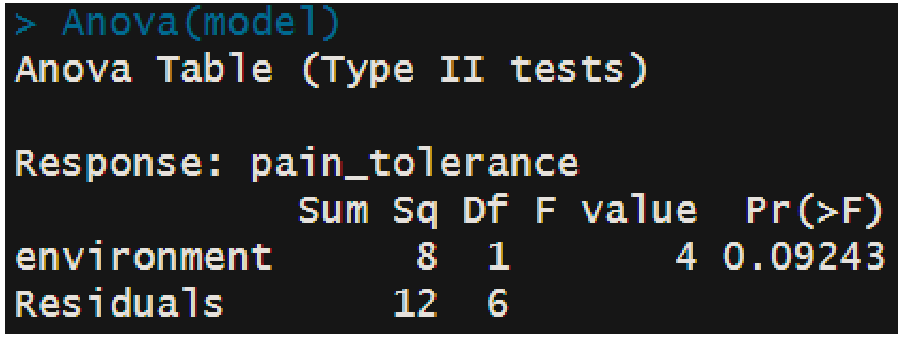
Performing the Analysis in R

Look at the model output using:

- \`Anova(model)\` from the car package, and
- summary(model)

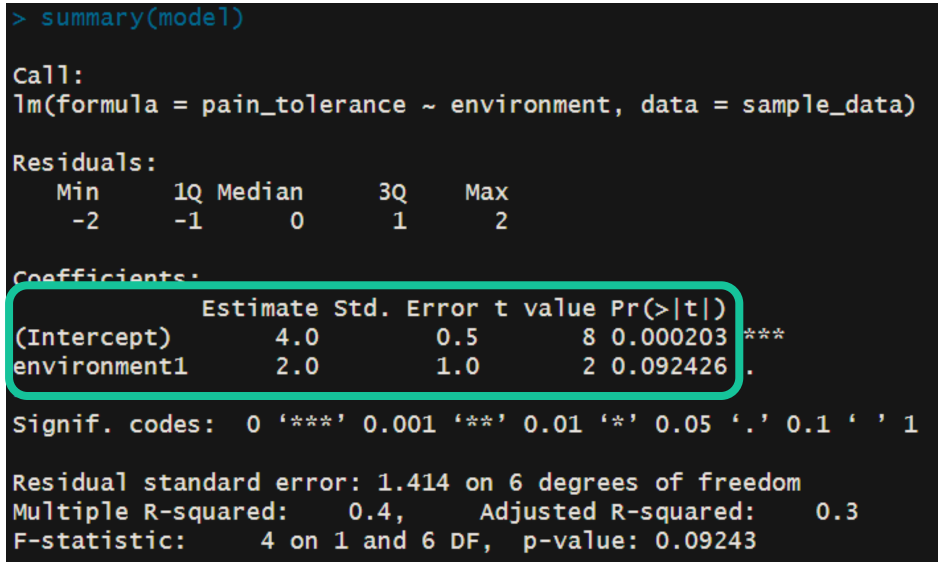
Notice there has been no change in the Anova() output.

- And *almost* no changes in the summary() output.
- What is the only difference in the output resulting from the new coding values?
  - The parameter estimates



```
> Anova(model)
Anova Table (Type II tests)

Response: pain_tolerance
            Sum Sq Df F value  Pr(>F)
environment      8  1       4 0.09243
Residuals       12  6
```



```
> summary(model)

Call:
lm(formula = pain_tolerance ~ environment, data = sample_data)

Residuals:
   Min     1Q Median     3Q    Max 
    -2     -1      0      1      2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.0        0.5       8 0.000203 ***
environment1      2.0        1.0       2 0.092426 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 6 degrees of freedom
Multiple R-squared:    0.4,	Adjusted R-squared:    0.3 
F-statistic:     4 on 1 and 6 DF,  p-value: 0.09243
```

Power Analysis

Power is the probability that a particular analysis will detect a significant result if there truly is one.

- i.e., If the null hypothesis is false, there is a ___% chance of correctly rejecting it.

The desired power level is set by the researcher.

- 80% is a commonly used minimum desired power level

It is recommended to perform an **a priori power analysis** to determine the sample size one needs to have a well-powered study before collecting one's data.

A Priori Power Analysis

Performing an a priori power analysis requires knowing:

- The model comparison being tested
- Alpha level
- An estimate of the true effect size
- Desired power level

Ideally, the researcher's best guess of the true effect size would be informed by previous literature that has studied similar effects or relationships.

- If an effect size is not readily available from the literature, one can use conventions for "small," "medium," and "large" effects
- There are several different types of effect size measures with their own conventions
  - Conventions are not necessarily equivalent across effect size measures

Cohen's d
Small = 0.2
Medium = 0.5
Large = 0.8

R-Squared
Small = 0.02
Medium = 0.13
Large = 0.26

Proposing an Effect Size

- For our power analyses, we'll use **R-squared** (also called eta-squared in many statistical softwares) as the measure of effect size

$$R^2 = \frac{SSR}{SS_{Total}}$$

- SSR is the sum of squares reduced corresponding to the model comparison that was specified
- $\mathrm{SS}_{\mathrm{Total}}$ is the total sum of squares on the outcome variable

- *Note*: This value is equal to PRE when Model C is the one parameter model.

Performing an A Priori Power Analysis

- For example, let's propose for our model comparison that we estimate that Model A will explain 13% more variance in Cooperation scores than Model C (a medium effect size).

R-Squared
Small = 0.02
Medium = 0.13
Large = 0.26

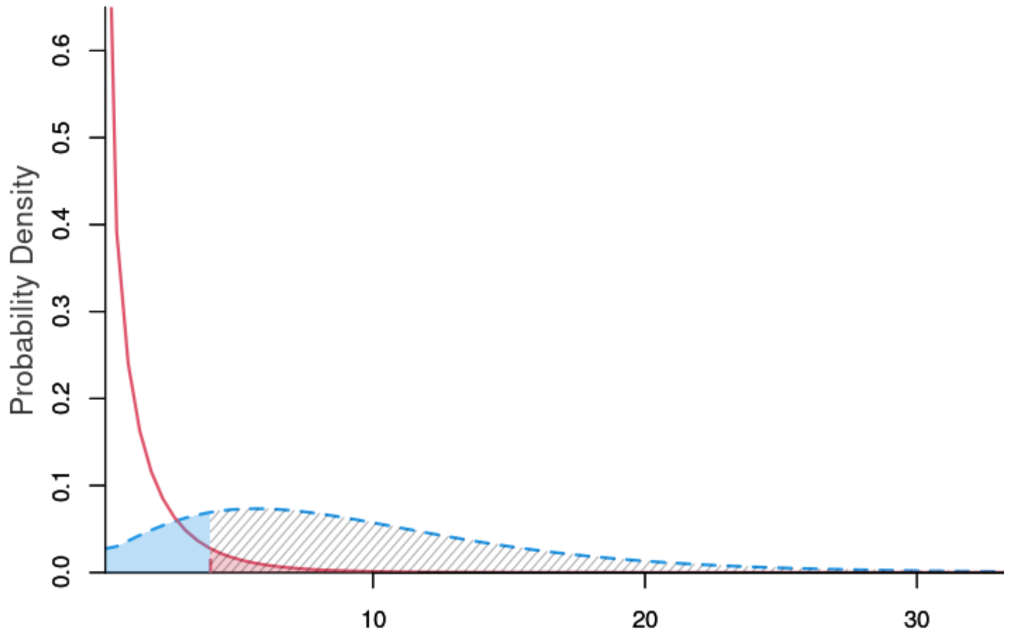
Model C: $Y_i = \beta_0 + \varepsilon_i$

Model A: $Y_i = \beta_0 + \beta_1 \text{Environment}_i + \varepsilon_i$

- What sample size would we need to have an 80% chance of detecting this medium sized effect?

Performing an A Priori Power Analysis

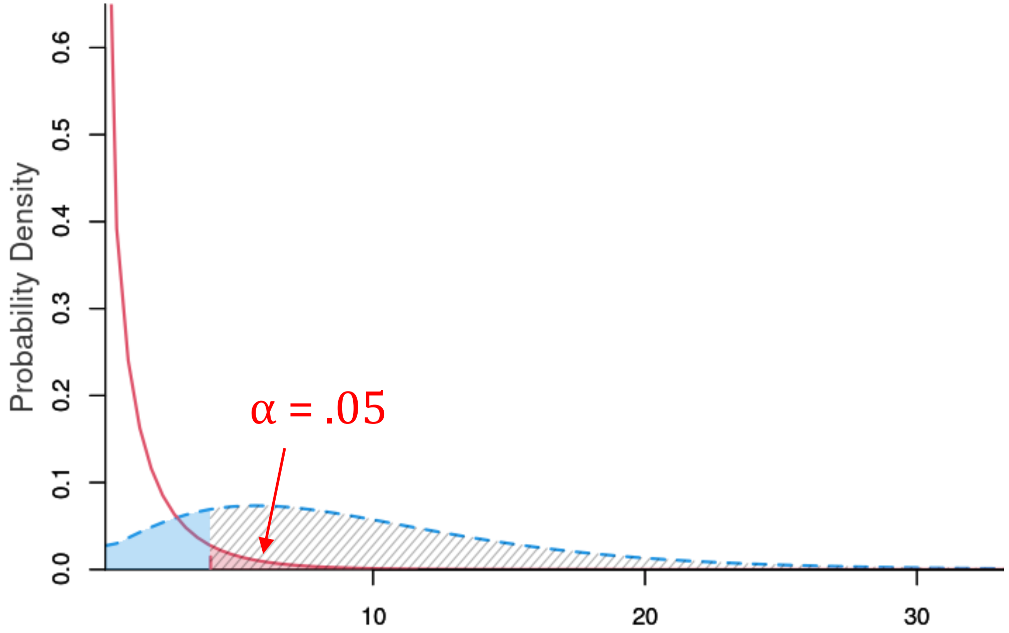
To perform an a priori power analysis, one must specify **the sampling distribution under the conditions of the null hypothesis** and **the sampling distribution that would result from the proposed effect size**.



Performing an A Priori Power Analysis

The red curve is the *F*-sampling distribution representing the results we would expect to get *if the null hypothesis is true*.

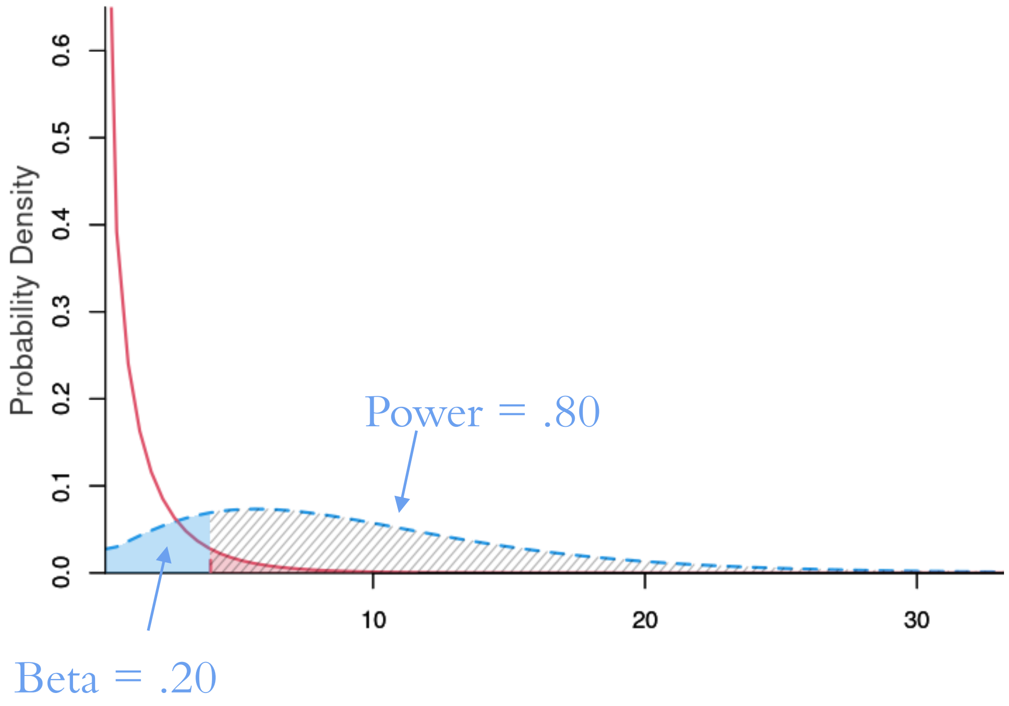
- Alpha determines the size of the **rejection region** where
- Landing in the rejection region means the results are **significant**



Performing an A Priori Power Analysis

The blue curve is the *F*-sampling distribution representing the results we would expect to get *if the true effect size is* $R^2 = 0.13$

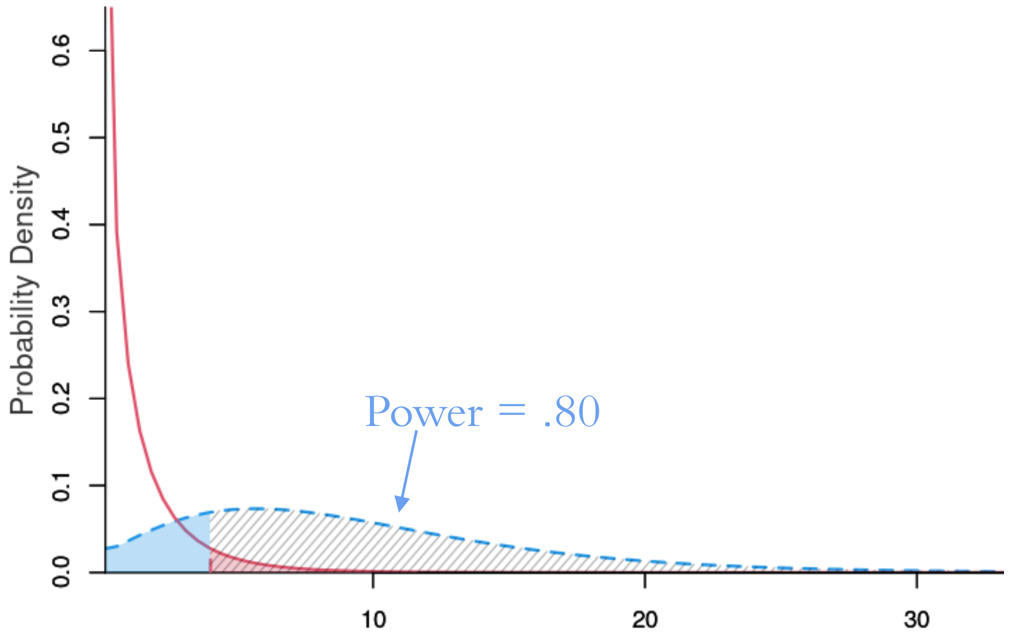
- The region of the blue distribution that overlaps with the rejection region of the red distribution corresponds to the **power of the analysis**
- The part of the blue distribution that *falls short of* the rejection region is called **beta** and corresponds to the chance of making a Type II error (1 – power)



Performing an A Priori Power Analysis

We can solve for the sample size needed to achieve 80% power in the proposed scenario using R or an online calculator:

- In lab, you'll discuss using the `pwrss.f.reg` function in R to perform a power analysis
- The creators of this function also created an online shinyapp: https://pwrss.shinyapps.io/index/


Probability Density
Power = .80