Categorical Predictors in Linear Regression

Overview

- Review of categorical predictor examples
- Setup of our working example
- Introduction to using categorical variables in linear regression

Goals

- Understand how to use categorical variables as predictors
- Understand how to makes inferences about categorical predictors

What is a Categorical Predictor?

A categorical predictor is a variable where the values denote membership to a specific group.

Examples of Categorical Predictors?

Some common categorical predictors are:

- Biological sex assigned at birth (Male or Female)
- Educational Level (Some HS, HS, College, Graduate)
- Employment Status (Full Time or Part Time)

Can anyone think of any others?

Dichotomous vs Multicategorical Predictors

We will distinguish between categorical predictors that have two groups and those that have three or more groups:

- **Dichotomous or Binary Predictors**: Two groups
- **Multicategorical Predictors**: Three or more groups

Dichotomous (Binary) Predictors

Our Working Example

We want to understand if an employee's previous experience with AI-based technology affects the frequency with which they adopt to the organization's new gen AI chatbot.

Our data generating process (theoretical model):

$$Y_{\text{Freq Use.}} = \beta_0 + \beta_1 X_{\text{Prev. Exp. Tech.}}$$

Where the variable **previous experience with technology** is a dichotomous variable with the values: Yes or No.

Representing a Dichotomous Predictor with Numbers

To be able to use a dichotomous predictor in a statistical model, we have to represent the categories as numbers:

Prev. Exp.	Code 1	Code 2	Code 3	Code 4
No	-0.5	-1	1	0
Yes	0.5	1	2	1

Indicator Coding

We can use any coding scheme as long as observations in the same group are assigned the same number and these numbers are different across groups. There is, however, one coding scheme that makes the most sense for regression: **Indicator Coding**.

Prev. Exp.	Indicator Code
No	0
Yes	1

Storing a Dichtomous Predictor in Your Data

It is good practice to have the numerical and categorical representation of your categorical variable in your dataset:

```
# A tibble: 10 × 3
   prev_exp_cat prev_exp freq_use
   <chr>           <dbl>    <dbl>
 1 No                  0        3
 2 No                  0        1
 3 No                  0        3
 4 No                  0        2
 5 No                  0        4
 6 No                  0        4
 7 No                  0        1
 8 Yes                 1        6
 9 Yes                 1        3
10 Yes                 1        2
```

Interpreting the Regression Coefficients

The reason we use indicator coding is because it makes the regression coefficients very easy to interpret:

$$\beta_0 = \overline{Y}_{\text{Group: 0}}$$

$$\beta_1 = \overline{Y}_{\text{Group: 1}} - \overline{Y}_{\text{Group: 0}}$$

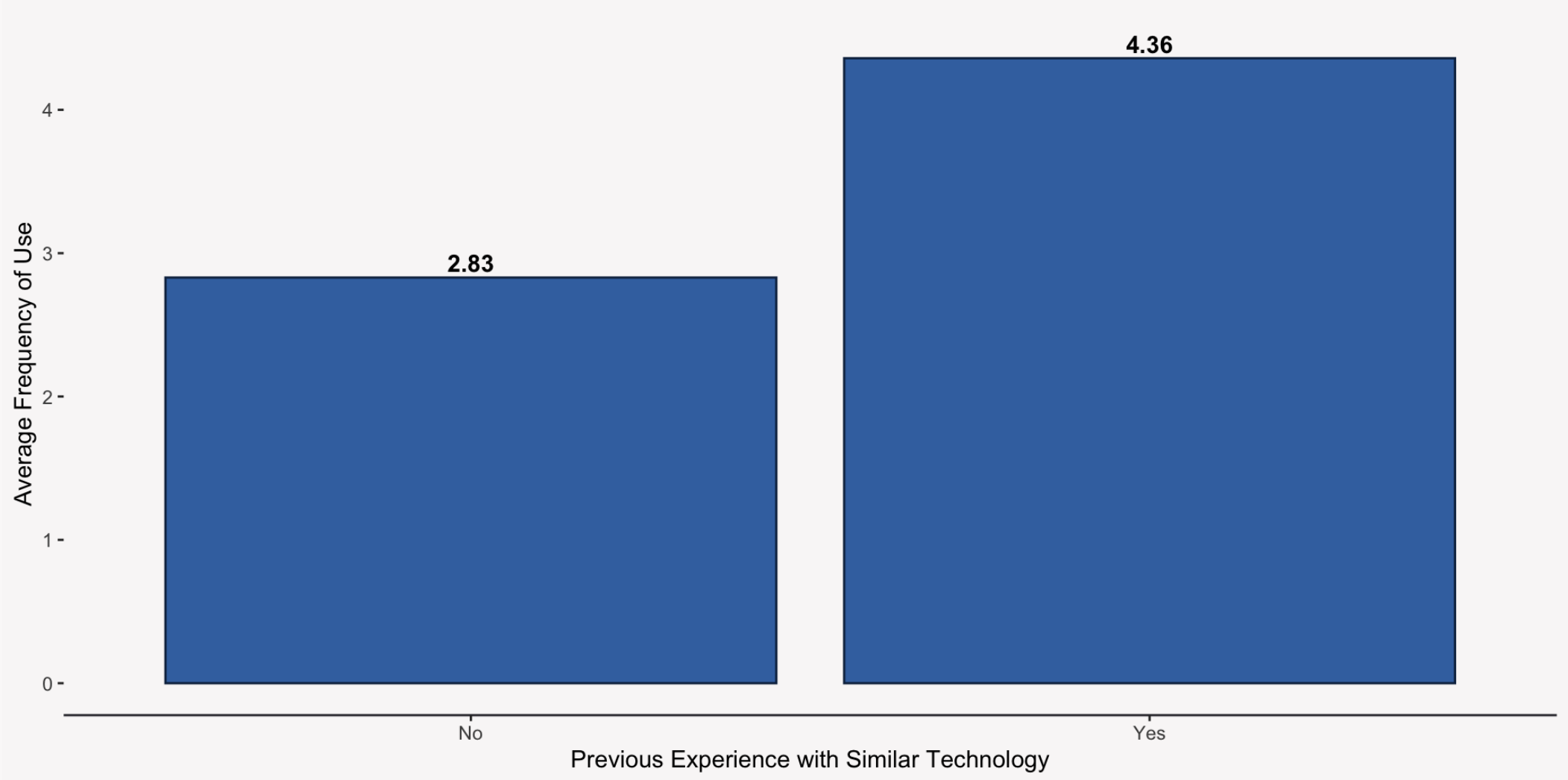
Estimating Mean Differences with Regression

```r
model <- lm(freq_use ~ prev_exp_cat, data = data_lecture)
summary(model)
```

```
Call:
lm(formula = freq_use ~ prev_exp_cat, data = data_lecture)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3600 -0.8343  0.1657  1.1657  3.1657

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.83429    0.06485   43.71   <2e-16 ***
prev_exp_catYes  1.52571    0.11839   12.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.213 on 498 degrees of freedom
Multiple R-squared:  0.2501,    Adjusted R-squared:  0.2486
F-statistic: 166.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

Mean Differences in Our Data

The table below contains the averages for `freq_use` by `prev_exp` group:



Changing the Reference Group

The reference group is the group that is assigned the value of 0. Nothing happens to the overall model fit when you switch the reference groups, the regression coefficients will change, however:

```
Call:
lm(formula = freq_use ~ prev_exp_cat, data = data_lecture)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3600 -0.8343  0.1657  1.1657  3.1657

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.36000    0.09906   44.02   <2e-16 ***
prev_exp_catNo -1.52571    0.11839  -12.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.213 on 498 degrees of freedom
Multiple R-squared:  0.2501,    Adjusted R-squared:  0.2486
F-statistic: 166.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

Comparison to a T-Test

When a regression model only includes a single dichotomous predictor, it is equivalent to a Student's T-Test—a common statistical method used to test the means of two groups are significantly different:

Method	95% Conf. Int. Lower	95% Conf. Int. Upper	T Value
T-Test	-1.76	-1.29	-12.89
Regression	-1.76	-1.29	-12.89

Including a Quantitative Predictor

You can, and often will, include quantitative (continuous) predictor in a model that contains a categorical predictor. The interpretations of the regression intercept and slope for the categorical variable change, however.

- **Intercept**: The estimated mean of the reference group when all other predictor variables equal 0.
- **Slope for Categorical Variable**: The mean difference between the two groups at different levels of the predictor.

Multiple Groups

Our Revised Example

We are going to assume the same data generating process as earlier, but now the categorical variable **previous experience with technology** has three different categories: low, moderate, or high.

$$Y_{\text{Freq Use.}} = \beta_0 + \beta_1 X_{\text{Mod. Exp.}} + \beta_2 X_{\text{High Exp.}}$$

Indicator Coding for Multicategorical Predictors

Indicator coding for multicategorical predictors works just like indicator coding for dichotomous variables with one exception: **we need multiple indicator variables**.

Prev. Exp.	Indicator Code: Moderate	Indicator Code: High
low	0	0
moderate	1	0
high	0	1

Storing a Mulitcategorical Predictor in Your Data

To numerically represent a multicategorical variable, we need to create **g - 1 indicator variables**, where **g equals the number of groups**.

```
# A tibble: 10 × 4
   prev_exp_cat prev_exp_mod prev_exp_high freq_use
   <fct>               <dbl>         <dbl>    <dbl>
 1 low                     0             0        3
 2 low                     0             0        4
 3 low                     0             0        3
 4 low                     0             0        3
 5 moderate                1             0        3
 6 moderate                1             0        3
 7 moderate                1             0        3
 8 high                    0             1        6
 9 high                    0             1        5
10 high                    0             1        3
```

Setting a Reference Group

The reference group is the group that receives a 0 across all of the indicator variables for the categorical variable. In our example, the reference group is `low` as observations in the `low` group receive values of 0 across both the `prev_exp_mod` and `prev_exp_high` indicator variables.

There are no real statistical consequences for your choice of reference group, I generally recommend picking the group with the largest group size or the group you want to compare to all of the remaining groups.

Interpreting Regression Coefficients for Indicator Variables

The reason we use indicator coding is because it makes the regression coefficients very easy to interpret:

$$\beta_0 = \overline{Y}_{\text{Group: Low (Reference)}}$$

$$\beta_{\text{Mod.}} = \overline{Y}_{\text{Group: Moderate}} - \overline{Y}_{\text{Group: Low}}$$

$$\beta_{\text{High}} = \overline{Y}_{\text{Group: High}} - \overline{Y}_{\text{Group: Low}}$$

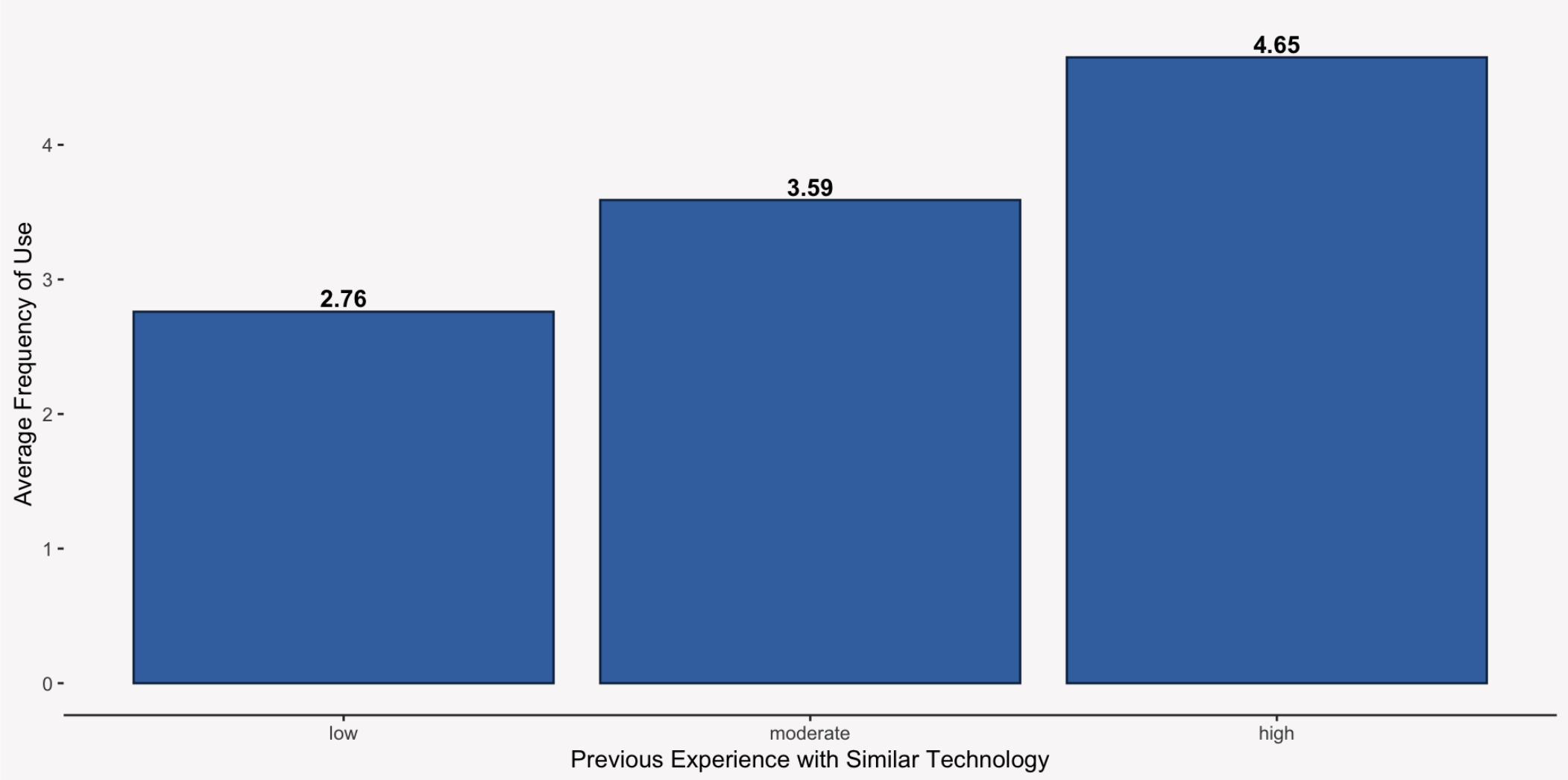
Model Summary: freq_use ~ prev_exp_cat

```r
model <- lm(freq_use ~ prev_exp_cat, data = data_lecture_multicat)
summary(model)
```

```
Call:
lm(formula = freq_use ~ prev_exp_cat, data = data_lecture_multicat)

Residuals:
   Min     1Q Median     3Q    Max
 -2.65  -0.76   0.24   1.24   3.24

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.76000    0.06098  45.262  < 2e-16 ***
prev_exp_catmoderate 0.82667    0.11133   7.425  3.9e-13 ***
prev_exp_cathigh     1.89000    0.12936  14.611  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.141 on 597 degrees of freedom
Multiple R-squared:  0.2762,    Adjusted R-squared:  0.2737
```

Mean Differences in Our Data

The table below contains the averages for `freq_use` by `prev_exp` group:



Using the F Test

Remember the F test tests the hypothesis that **all regression slope coefficients are equal to 0**. When your model only includes a single multicategorical predictor, this is equivalent to testing if **the means across the different groups are all equal** as the regression slopes are tests of group mean difference.

A significant F means that **at least one group mean** is different from the other group means.

Comparison to a One-Way ANOVA

When the regression model includes only a single multicategorical predictor, it is equivalent to an ANOVA—a common statistical method used to test for mean differences across three or more groups:

Method	F value
ANOVA	113.89
Regression	113.89

How Do We Make Multiple Comparisons?

Indicator coding will **only** give you a subset of the possible mean comparisons you can make:

- Moderate - Low: 0.83
- High - Low: 1.89
- Moderate - High: ??

To get all the possible comparisons, you will need to change your reference group accordingly and refit your regression model.

Changing `prev_exp` Reference Group to `high`

```r
data_lecture_multicat <-
  data_lecture_multicat |>
  dplyr::mutate(
    prev_exp_cat = relevel(prev_exp_cat, ref = "high")
  )

lm(freq_use ~ prev_exp_cat, data = data_lecture_multicat) |> summary()
```

```
Call:
lm(formula = freq_use ~ prev_exp_cat, data = data_lecture_multicat)

Residuals:
   Min     1Q Median     3Q    Max
 -2.65  -0.76   0.24   1.24   3.24

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.6500     0.1141   40.76  < 2e-16 ***
prev_exp_catlow       -1.8900     0.1294  -14.61  < 2e-16 ***
prev_exp_catmoderate  -1.0633     0.1473   -7.22 1.59e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.141 on 597 degrees of freedom
Multiple R-squared:  0.2762,    Adjusted R-squared:  0.2737
```

The Bonferonni Correction

When you start making multiple comparisons, you are essentially reusing your data over and over again, which **increases the chance of finding a significant finding by chance**.

To protect against this, you need to make it harder to find a significant result by changing the threshold at which you would have declared a finding significant to:

$$\text{New Threshold} = \frac{\text{Old Threshold } (.05)}{\#\text{ of Comparisons}}$$

Including a Quantitative Predictor Variable

When you include a quantitative predictor variable alongside your categorical predictor, you have to interpret the regression coefficients for the categorical variable as group mean differences at a specific level of the quantitative variable or group mean differences while controlling for the effects of the other variables.

General Considerations with Categorical Predictors

Do Not Artificially Group Your Data

If you have a continuous (quantitative) predictor variable, **you should not** try to transform it into a categorical predictor by creating artificial groups based on the responses to the continuous predictors unless:

- You are truly interested in group differences
- Responses to the continuous variable are clustering in groups

Do Not Standardize Categorical Predictors

Standardized categorical predictors do not make any sense as standardizing changes the scale of the variable and the scale matters for categorical predictors!