CATEGORICAL X

Research Methods in Psychology I & II ▪ Department of Psychology ▪ Colorado State University

BY THE END OF THIS UNIT YOU WILL:

1. Know how to create dummy codes to represent a categorical predictor.
2. Know how to include dummy coded indicators as predictors in a MLR.
3. Know how to interpret the regression coefficients associated with dummy coded indicators.
4. Know how to plot the results of a MLR model with categorical predictors.

Why do we need special techniques for categorical predictors?

Often times we will find a need to include qualitative predictors in our models—variables such as gender, race/ethnicity, political party affiliation, etc. We can easily incorporate these sorts of variables by using special coding techniques. We will learn about the most common coding method — dummy coding.

"Well, I guess we're the control group."

What are Categorical Variables?

- Categorical variables are qualitative in nature (rather than quantitative).
- The values denote ordered or unordered categories. Unordered (i.e., nominal) variables have no reasonable order that can be applied (e.g., gender, marital status, political party affiliation). Ordered (i.e., ordinal) variables have a reasonable order (e.g., level of education [HS dropout, HS grad, college grad], developmental period [child, teen, adult], level of liberalism (conservative, moderate, liberal), but they are not on a ratio or interval scale.
- When a variable has two categories, it is dichotomous (or binary). When a variable has more than two categories, it is polychotomous.

Inquiry Activity

Imagine a variable that represents political party that is coded as follows: 1=Republican, 2=Democrat, 3=Reform Party, 4=Green Party. Would it make sense to include this variable "as is" in a regression model to predict views on a certain political issue? In this model, what would the slope for political party capture?

Methods for Handling Categorical Predictors in MLR

By simple recoding of the categorical predictor(s), we can easily incorporate these types of variables as independent variables in our MLR models. In this unit, we will consider the most common method — dummy coding. There are other types that you may encounter as well (e.g., effect coding), but once you learn one method you can easily understand and apply others.

Dummy Coded Categorical Variables

Dummy coded variables summarize all the information in a categorical variable with *k* categories, via *k*-1 indicators. For example, a categorical variable with two categories (e.g., sex — male or female) can be represented by 1 dummy coded indicator. A categorical variable with three categories (e.g., race/ethnicity — Black, Hispanic, White) can be represented by 2 dummy coded indicators.

One category serves as the reference group. This category is assigned a 0 for all dummy coded indicators. The remaining categories have a 1 for the respective category and a 0 for all others. Let's form dummy coded indicators for sex and for race/ethnicity — we'll use female as the reference group for sex, and non-Hispanic White as the reference group for race/ethnicity.

SEX EXAMPLE

Original Variable — SEX: Male, Female

New Indicator — MALE: 1=Male, 0=Female

ID	SEX	MALE
1	MALE	1
2	FEMALE	0
3	FEMALE	0
4	MALE	1
5	FEMALE	0
6	MALE	1

RACE/ETHNICITY EXAMPLE

Original Variable — RACE: Black, Hispanic, Non-Hispanic White

New Indicators — BLACK: 1=Black, 0=not Black; HISPANIC: 1=Hispanic, 0=not Hispanic

ID	RACE	BLACK	HISPANIC
1	BLACK	1	0
2	HISPANIC	0	1
3	BLACK	1	0
4	NHWHITE	0	0
5	HISPANIC	0	1
6	NHWHITE	0	0

TIP: Always give your dummy coded indicator a name that corresponds to the group coded 1.

A New Data Example: Effectiveness of a Sleep Intervention

A team of sleep researchers sought to study the effects of a 6-week sleep intervention aimed to improve participant's sleep hygiene. Sleep hygiene encompasses a variety of practices and habits that are necessary to have good nighttime sleep quality and full daytime alertness. The team formulated three different versions of the intervention. The first version (condition 1) provided participants with a self-help book on the topic of sleep hygiene. The second version (condition 2) brought participants together once per week in groups of 10-12 to teach the principles of sleep hygiene in a classroom setting. The final version (condition 3) also used the group-based classroom setting of condition 2, but in addition, each participant's partner was invited to also take part in the group sessions. Six-hundred male and female adults living with an intimate partner and suffering from a sleep disorder were recruited to take part in the study, the participants were randomly assigned to one of the three conditions.

The data set includes the following variables:

- **sex:** 1=male, 2=female
- **age:** Participant's age in years
- **anxiety:** Participant's level of general anxiety measured at the start of the study via a multi-item scale. The scale (average of all items) ranges from 1 to 7, where a higher score indicates a higher level of anxiety.
- **prior:** An indicator of whether or not the participant had previously participated in some type of sleep intervention, 1 = yes, 0 = no.
- **hygiene:** Participant's sleep hygiene at week 6. It ranges from 0 to 10, and higher means better sleep practices.
- **support:** Participant's perception that their partner is supportive of their struggles with sleep and their efforts to improve sleep. It is a multi-item scale that ranges from 1 to 5, where higher indicates more support.
- **sleep:** Participant's average sleep efficiency during the month following the intervention, calculated as time spent in bed *asleep* (minus all the awakenings), divided by the total time spent in bed. It is expressed as a percentage.
- **lifesat:** Participant's sense of life satisfaction measured 30 days after the completion of the intervention. It is a multi -item scale that ranges from 1 to 7, where a higher score indicates more satisfaction.
- **cond:** Treatment condition, 1 = self-help, 2 = group-based intervention, 3 = group-based plus partner participation.

Sleep_Notebook.Rmd*

```
---
title: "R Notebook for Sleep Study Data"
output: html_notebook
---


# Load libraries
```{r}

library(tidyverse)
library(psych)
library(olsrr)
library(descriptr)
library(car)
library(modelr)

```


Import data
```{r}

slp <- read_csv("slpdata.csv")

```
```

Describe the Data

```
describe(slp)
```

```
        vars   n   mean     sd median trimmed    mad   min    max  range  skew kurtosis   se
cond       1 600   2.00   0.82   2.00    2.00   1.48  1.00   3.00   2.00  0.00    -1.50 0.03
prior      2 600   0.72   0.45   1.00    0.78   0.00  0.00   1.00   1.00 -1.00    -1.01 0.02
age        3 600  44.94  12.87  45.20   45.12  16.46 20.00  67.80  47.80 -0.10    -1.14 0.53
anxiety    4 600   3.88   0.90   3.86    3.89   0.93  1.05   6.84   5.79 -0.07    -0.06 0.04
hygiene    5 600   5.99   1.57   6.05    6.04   1.57  1.68   9.74   8.06 -0.23    -0.29 0.06
support    6 600   3.04   0.68   2.96    3.02   0.73  1.09   4.91   3.82  0.21    -0.51 0.03
sleep      7 600  68.88  12.14  69.00   69.09  11.86 34.00  99.00  65.00 -0.16    -0.17 0.50
lifesat    8 600   4.06   0.92   4.05    4.04   0.96  1.68   6.61   4.93  0.13    -0.23 0.04
sex        9 600   1.41   0.49   1.00    1.39   0.00  1.00   2.00   1.00  0.36    -1.87 0.02
id        10 600 300.50 173.35 300.50  300.50 222.39  1.00 600.00 599.00  0.00    -1.21 7.08
```

Exploration #1: Sex differences in Sleep Hygiene for Condition 1

Let's begin by examining the difference in sleep hygiene following the program between males and females participating in condition 1. First, we need to create a dummy indicator for sex. We will also create a factor version of this variable for later plotting. Second, we need to filter the data to keep only people in condition 1.

```
slp <- mutate(slp,
        female = ifelse(sex == 1, 0, 1),
        female.f = factor(female, levels = c(0,1), labels = c("male", "female")))

# check to make sure new variable was correctly created
cross_table(slp$sex, slp$female)
```

```
   Cell Contents
|---------------|
|     Frequency |
|       Percent |
|       Row Pct |
|       Col Pct |
|---------------|

Total Observations:  600

-------------------------------------------------------------
|              |                  female                    |
-------------------------------------------------------------
|          sex |            0 |            1 |    Row Total |
-------------------------------------------------------------
|            1 |          353 |            0 |          353 |
|              |        0.588 |            0 |              |
|              |            1 |            0 |         0.59 |
|              |            1 |            0 |              |
-------------------------------------------------------------
|            2 |            0 |          247 |          247 |
|              |            0 |        0.412 |              |
|              |            0 |            1 |         0.41 |
|              |            0 |            1 |              |
-------------------------------------------------------------
| Column Total |          353 |          247 |          600 |
|              |        0.588 |        0.412 |              |
-------------------------------------------------------------
```

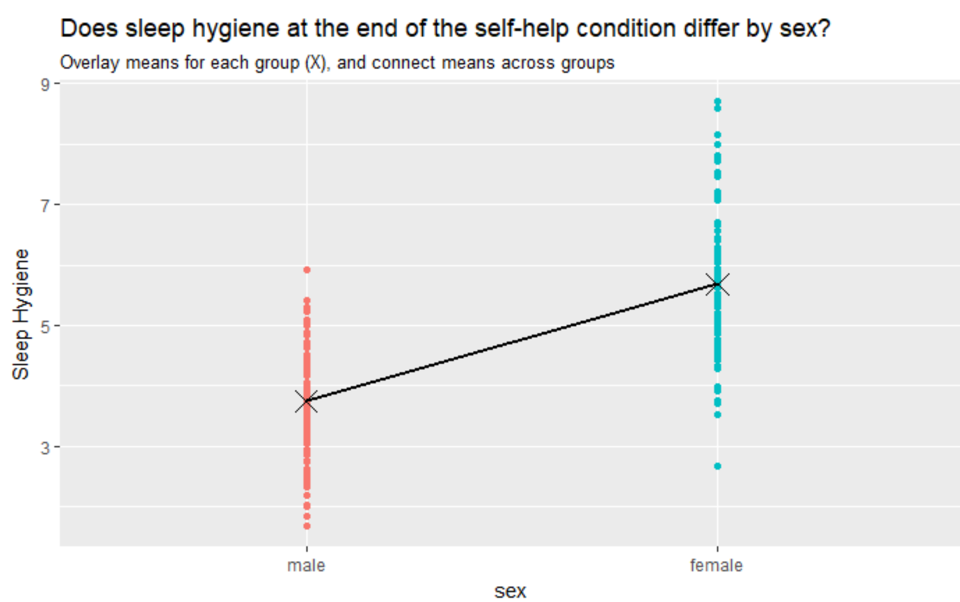
```
slp_cond1 <- filter(slp, cond == 1)
```

Summarize Sleep Hygiene in Condition 1 by Sex

```
slp_cond1 %>%
  group_by(female.f) %>%
  summarize(hyg_mean = mean(hygiene)) %>%
  ungroup()
```

female.f <fctr>	hyg_mean <dbl>
male	3.750943
female	5.686064

```
ggplot(slp_cond1, aes(x = female.f, y = hygiene, group = female.f, color = female.f)) +
  geom_point(show.legend = FALSE) +
  stat_summary(fun.y=mean, geom="line", lwd = 1, colour = "black", aes(group=1))  +
  stat_summary(fun.y=mean, size = 5, shape = 4, geom="point", colour = "black", show.legend = FALSE) +
  labs(title = "Does sleep hygiene at the end of the self-help condition differ by sex?",
       subtitle = "Overlay means for each group (X), and connect means across groups",
       x = "sex", y = "Sleep Hygiene")
```



Fit a Simple Linear Regression to Predict Sleep Hygiene with Sex

```
m1 <- lm(data =  slp_cond1, hygiene ~ female)
ols_regress(m1)
```

```
qt(c(.025, .975), 198)     [1] -1.972017  1.972017
```

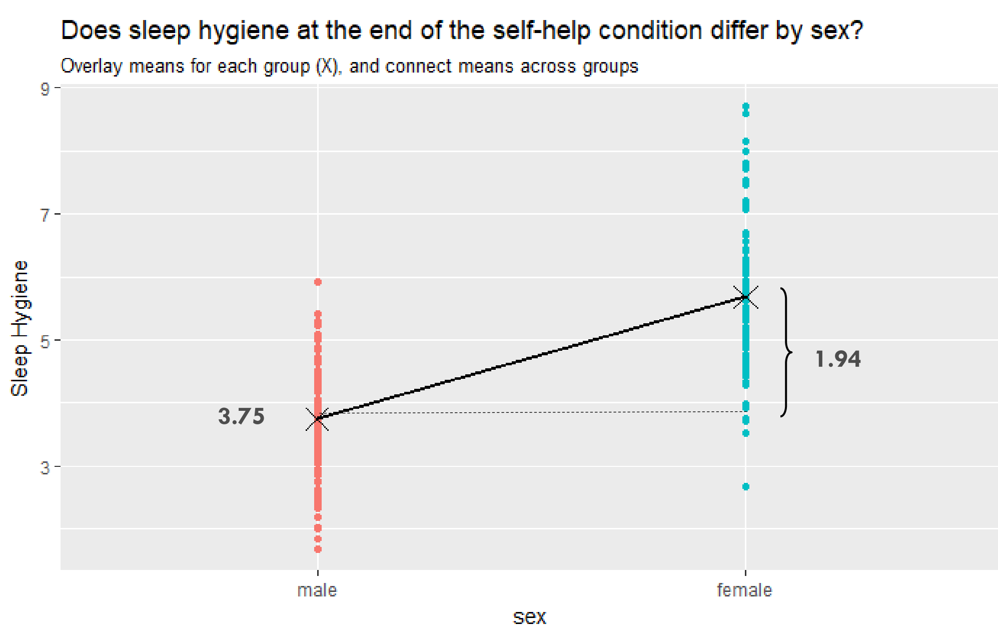
```
                        Model Summary
---------------------------------------------------------------
R                       0.682       RMSE              1.042
R-Squared               0.465       Coef. Var        22.349
Adj. R-Squared          0.462       MSE               1.085
Pred R-Squared          0.454       MAE               0.823
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
----------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square       F         Sig.
----------------------------------------------------------------------
Regression   186.561          1        186.561    171.964    0.0000
Residual     214.807        198          1.085
Total        401.367        199
----------------------------------------------------------------------

                                  Parameter Estimates
--------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig     lower    upper
--------------------------------------------------------------------------------------
(Intercept)     3.751         0.101                 37.077    0.000    3.551    3.950
     female     1.935         0.148        0.682    13.113    0.000    1.644    2.226
--------------------------------------------------------------------------------------
```

The intercept is, as usual, the predicted score when x = 0, so predicted sleep hygiene for males.

The slope is, as usual, the predicted change in y for a one-unit increase in x, so the expected difference in sleep hygiene for a female compared to a male. Females, on average, have a sleep hygiene score that is 1.9 units higher than males. The p-value is less than alpha, indicating that females in condition 1 had significantly better sleep hygiene than males in condition 1.



female.f <fctr>	hyg_mean <dbl>
male	3.750943
female	5.686064

$$\hat{hygiene} = 3.75 + (1.94 \cdot female_i)$$

For females: 3.75 + (1.94 ▪ 1) = 5.69

For males: 3.75 + (1.94 ▪ 0) = 3.75

Exploration #2: Condition Differences in Sleep Efficiency among Females

Now, consider the differences in sleep efficiency across the three conditions (self-help, group, group + partner inclusion) for females in the study. First, we need to create a set of dummy coded indicators to represent condition. We will select condition 1 (self-help) as the reference group. Second, we will filter the dataset to include only females.

```
slp <- mutate(slp,
        cond2 = ifelse(cond == 2, 1, 0),
        cond3 = ifelse(cond == 3, 1, 0),
        cond.f = factor(cond, levels = c(1,2,3), labels = c("self help", "group-based", "group + partner")))

cross_table(slp$cond, slp$cond2)
cross_table(slp$cond, slp$cond3)
```

```
   Cell Contents
|---------------|
|     Frequency |
|       Percent |
|       Row Pct |
|       Col Pct |
|---------------|

Total Observations:  600

-------------------------------------------------------------
|              |                   cond2                    |
-------------------------------------------------------------
|         cond |            0 |            1 |    Row Total |
-------------------------------------------------------------
|            1 |          200 |            0 |          200 |
|              |        0.333 |            0 |              |
|              |            1 |            0 |         0.33 |
|              |          0.5 |            0 |              |
-------------------------------------------------------------
|            2 |            0 |          200 |          200 |
|              |            0 |        0.333 |              |
|              |            0 |            1 |         0.33 |
|              |            0 |            1 |              |
-------------------------------------------------------------
|            3 |          200 |            0 |          200 |
|              |        0.333 |            0 |              |
|              |            1 |            0 |         0.33 |
|              |          0.5 |            0 |              |
-------------------------------------------------------------
| Column Total |          400 |          200 |          600 |
|              |        0.666 |        0.333 |              |
-------------------------------------------------------------
```

```
   Cell Contents
|---------------|
|     Frequency |
|       Percent |
|       Row Pct |
|       Col Pct |
|---------------|

Total Observations:  600

-------------------------------------------------------------
|              |                   cond3                    |
-------------------------------------------------------------
|         cond |            0 |            1 |    Row Total |
-------------------------------------------------------------
|            1 |          200 |            0 |          200 |
|              |        0.333 |            0 |              |
|              |            1 |            0 |         0.33 |
|              |          0.5 |            0 |              |
-------------------------------------------------------------
|            2 |          200 |            0 |          200 |
|              |        0.333 |            0 |              |
|              |            1 |            0 |         0.33 |
|              |          0.5 |            0 |              |
-------------------------------------------------------------
|            3 |            0 |          200 |          200 |
|              |            0 |        0.333 |              |
|              |            0 |            1 |         0.33 |
|              |            0 |            1 |              |
-------------------------------------------------------------
| Column Total |          400 |          200 |          600 |
|              |        0.666 |        0.333 |              |
-------------------------------------------------------------
```

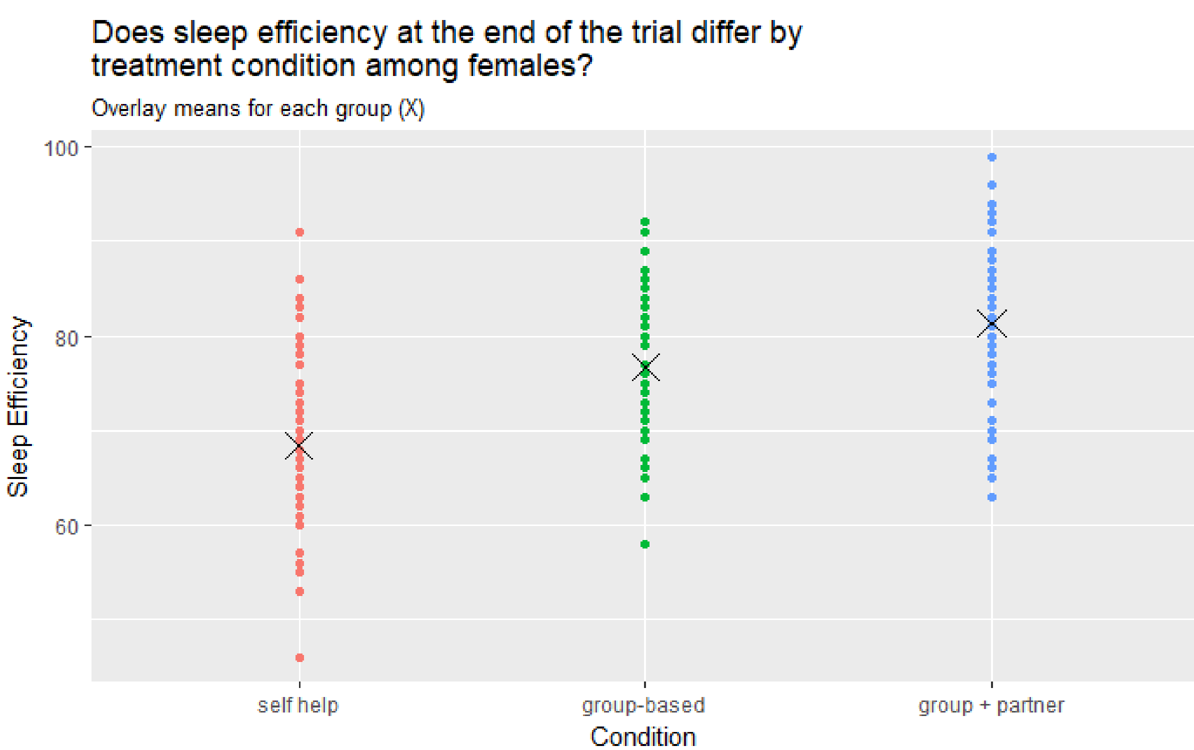
```
slp_females <- filter(slp, female == 1)
```

Summarize Sleep Efficiency across Conditions for Females

```
slp_females %>%
  group_by(cond.f) %>%
  summarize(sleep_mean = mean(sleep)) %>%
  ungroup()

ggplot(slp_females, aes(x = cond.f, y = sleep, group = cond.f, color = cond.f)) +
  geom_point(show.legend = FALSE) +
  stat_summary(fun.y=mean, size = 5, shape = 4, geom="point", colour = "black", show.legend = FALSE) +
  labs(title = "Does sleep efficiency at the end of the self-help condition differ by \ntreatment condition for females?",
     subtitle = "Overlay means for each group (X)",
     x = "Condition", y = "Sleep Efficiency")
```

cond.f <fctr>	sleep_mean <dbl>
self help	68.34043
group-based	76.70000
group + partner	81.36145



Fit a Multiple Linear Regression for Females

We will regress sleep efficiency on the two dummy coded indicators for condition among female participants.

```
m2 <- lm(data =  slp_females, sleep ~ cond2 + cond3)
ols_regress(m2)
```

```
qt(c(.025, .975), 244)          [1] -1.969734  1.969734
qf(.95, df1 = 2, df2 = 244)     [1] 3.032816
```

The significant F test for the overall model indicates that condition explains a significant amount of variability in sleep efficiency. The three means significantly differ.

```
                        Model Summary
--------------------------------------------------------------
R                        0.595       RMSE              7.605
R-Squared                0.354       Coef. Var        10.128
Adj. R-Squared           0.349       MSE              57.832
Pred R-Squared           0.338       MAE               6.001
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
----------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F          Sig.
----------------------------------------------------------------------
Regression     7728.252        2       3864.126    66.817     0.0000
Residual      14110.963      244         57.832
Total         21839.215      246
----------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta        t        Sig      lower    upper
-----------------------------------------------------------------------------------------
(Intercept)      68.340         0.784                  87.128    0.000     66.795   69.885
      cond2       8.360         1.201        0.401      6.963    0.000      5.995   10.724
      cond3      13.021         1.145        0.654     11.368    0.000     10.765   15.277
-----------------------------------------------------------------------------------------
```

The intercept is, as usual, the predicted score when all x's = 0, so predicted sleep efficiency for females in condition 1 (self-help).

The slope for cond2 is the predicted change in y for a one-unit increase in cond2, so the expected difference in sleep efficiency for females in condition 2 compared to condition 1. The p-value is less than alpha, indicating that females in condition 2 had significantly better sleep efficiency than females in condition 1. Likewise, the slope for cond3 is the expected difference in sleep efficiency for females in condition 3 compared to condition 1. The p-value is less than alpha, indicating that females in condition 3 had significantly better sleep efficiency than females in condition 1.

$$\hat{sleep}_i = 68.34 + (8.36 \cdot cond2_i) + (13.02 \cdot cond3_i)$$

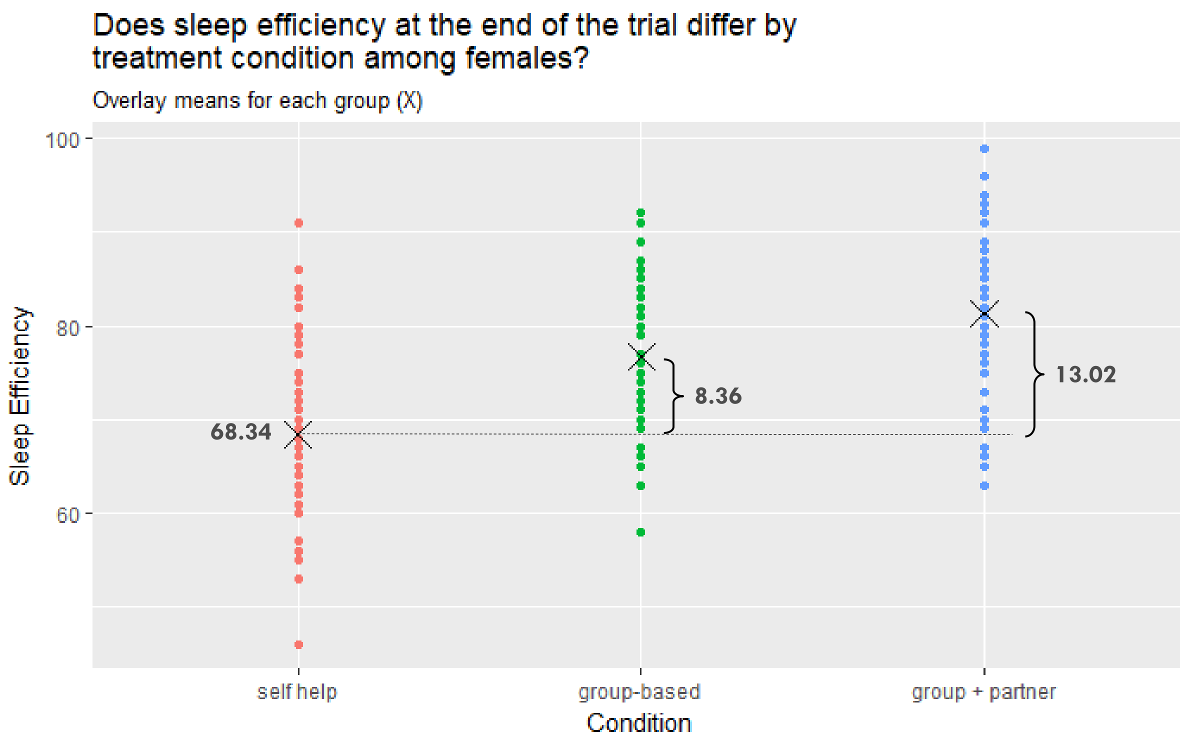
For females in condition 1: 68.34 + (8.36 ▪ 0) + (13.02 ▪ 0) = 68.34

For females in condition 2: 68.34 + (8.36 ▪ 1) + (13.02 ▪ 0) = 76.70

For females in condition 3: 68.34 + (8.36 ▪ 0) + (13.02 ▪ 1) = 81.36

cond.f <fctr>	sleep_mean <dbl>
self help	68.34043
group-based	76.70000
group + partner	81.36145

Visualization of Effects



$$\hat{sleep}_i = 68.34 + (8.36 \cdot cond2_i) + (13.02 \cdot cond3_i)$$

How Can we Compare Conditions 2 and 3?

Our first model considered condition 1 as the reference group. Therefore, we obtained a comparison of condition 2 to 1 and condition 3 to 1. But what if we want to compare condition 2 to 3? One way to accomplish this is to change the reference group and refit the model. This requires the creation of a new dummy coded indicator. Here, we create a dummy indicator called cond1 (condition 1 = 1, all others equal 0). Then fit the model with cond1 and cond2, making condition 3 the reference group.

```
slp_females <- mutate(slp_females, cond1 = ifelse(cond == 1, 1, 0))

m3 <- lm(data =  slp_females, sleep ~ cond1 + cond2)
ols_regress(m3)
```

```
                        Model Summary
--------------------------------------------------------------
R                       0.595       RMSE                7.605
R-Squared               0.354       Coef. Var          10.128
Adj. R-Squared          0.349       MSE                57.832
Pred R-Squared          0.338       MAE                 6.001
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
----------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression    7728.252         2       3864.126    66.817     0.0000
Residual     14110.963       244         57.832
Total        21839.215       246
----------------------------------------------------------------------

                                  Parameter Estimates
---------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta       t        Sig       lower      upper
---------------------------------------------------------------------------------------------
(Intercept)      81.361         0.835                 97.471    0.000     79.717     83.006
      cond1     -13.021         1.145       -0.672   -11.368    0.000    -15.277    -10.765
      cond2      -4.661         1.234       -0.223    -3.777    0.000     -7.092     -2.231
---------------------------------------------------------------------------------------------
```

Notice that the overall model fit is exactly the same as the previous model.

The intercept is now the predicted sleep for females in condition 3 (the group where all predictors equal 0).

The slope for cond1 is the predicted change in y for a one-unit increase in cond1, so the expected difference in sleep efficiency for females in condition 1 compared to condition 3 (females in condition 1 had worse sleep efficiency than females in condition 3). Likewise, the slope for cond2 is the expected difference in sleep efficiency for females in condition 2 compared to condition 3 (females in condition 2 had worse sleep efficiency than females in condition 3).

$$sl\hat{e}ep_i = 81.36 + (-13.02 \cdot cond1_i) + (-4.66 \cdot cond2_i)$$

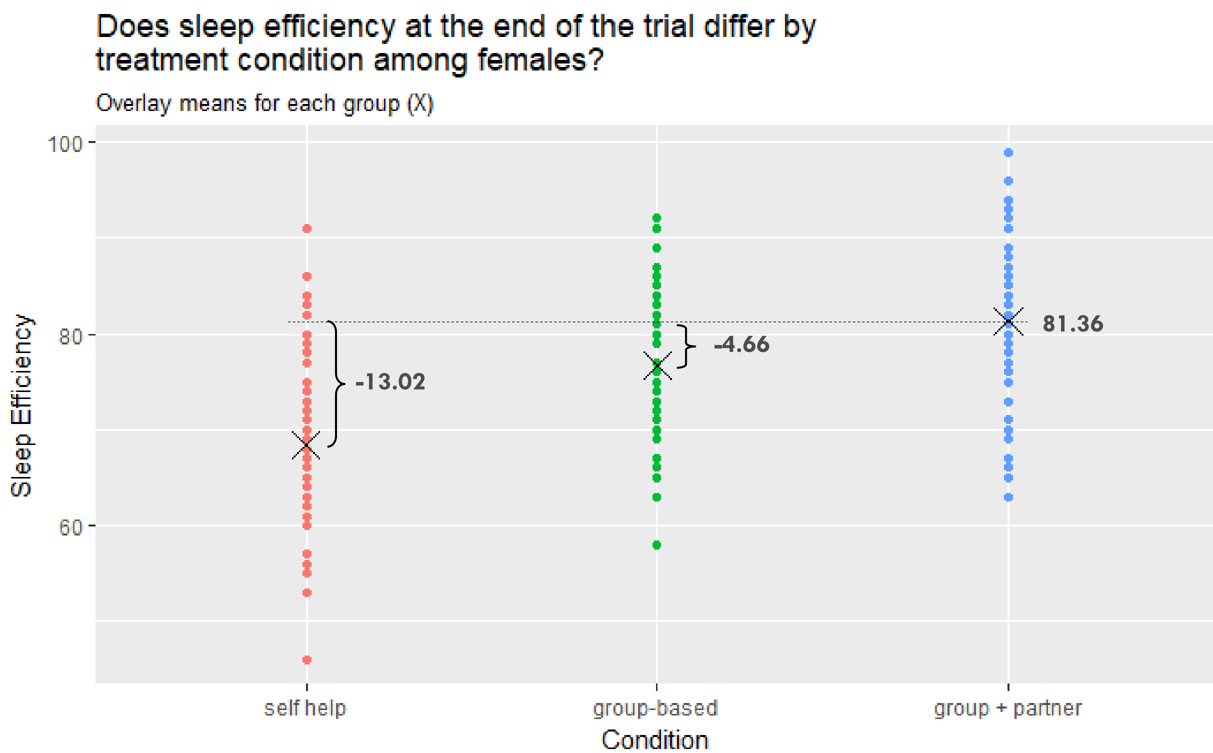
For females in condition 1: 81.36 + (-13.02 ▪ 1) + (-4.66 ▪ 0) = 68.34

For females in condition 2: 81.36 + (-13.02 ▪ 0) + (-4.66 ▪ 1) = 76.70

For females in condition 3: 81.36 + (-13.02 ▪ 0) + (-4.66 ▪ 0) = 81.36

cond.f <fctr>	sleep_mean <dbl>
self help	68.34043
group-based	76.70000
group + partner	81.36145

Visualization of Effects with Switched Reference Group



$$s\hat{leep}_i = 81.36 + (-13.02 \cdot cond1_i) + (-4.66 \cdot cond2_i)$$

Alternative Method to Compare Condition 2 and 3

Instead of refitting the model with the alternative set of dummy coded indicators, we can fit a linear constraint to the original model. The function called linearHypothesis from the car (Companion to Applied Regression) package assists us with this task.

```
linearHypothesis(m2, "cond2 = cond3")
```

```
Linear hypothesis test

Hypothesis:
cond2 - cond3 = 0

Model 1: restricted model
Model 2: sleep ~ cond2 + cond3

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    245 14936
2    244 14111  1    825.14 14.268 0.0001993 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F^* = 14.268—this exceeds the critical value of F, and accordingly, the p-value is less than alpha. Therefore, we reject the null hypothesis. The sleep efficiency for females in condition 3 is not equal to females in condition 2. The difference between the two slopes is 4.66—females in condition 3 have an average sleep efficiency score that is 4.66 units higher than females in condition 2. Notice that the t^* for the cond2 slope in the refit MLR (-3.777) squared equals this F^* (14.268) — these are equivalent methods for accomplishing the same task.

The F-test has degrees of freedom equal to the number of constraints for the numerator, and the SSE of the initial model for the denominator. Therefore the critical value of F is 3.88. The null hypothesis is that the slope for condition 2 is equal to the slope for condition 3.

```
qf(.95, df1 = 1, df2 = 244)
```

```
[1] 3.879852
```

The Danger of Multiple Comparisons

There are two types of errors we can make when we conduct a hypothesis test:

- Type I error: Rejecting H0 when it's true—concluding there's a difference when there isn't.
- Type II error: Failing to reject H0 when it's false—concluding there is no difference when there really is one.

We minimize Type I error when we set alpha=0.05 for our tests, but as we conduct multiple tests, the Type I error for the "family of tests" grows. We need a way of minimizing the Type I error when we make multiple comparisons for a set of groups.

Given *k* groups, we can make (*k**(*k*-1))/2 comparisons. For our current example, that's (3*2)/2=3: Condition 1 to Condition 2, Condition 1 to Condition 3, and Condition 2 to Condition 3. With all of these comparison, we're using up more df for the set than we have available. Instead of using alpha=0.05 for each individual test, one reasonable solution is to apply a Bonferroni correction. Here, we use our desired alpha (e.g., .05) for the entire family of tests when comparing groups in a set. This is called minimizing the family-wise error rate.

To calculate the new alpha, divide the desired alpha by the number of comparisons that you would like to make. For example, if we set alpha to .05, and we want to make all three comparisons, then take .05/3 = .017. We can look up the critical t for alpha=.017 (divide by 2 to put half in each tail for a two tailed test). We can also obtain the equivalent critical F if we are using the linear constraint method. The df remains the same as for the usual MLR model. In assessing the significance for each of the group comparisons, we will use this critical t/F rather than the usual critical t/F.

Effect of making multiple comparisons on the Type I error for the entire family of tests (α=0.05)	
# tests	# wrong
1	0.05
2	0.10
5	0.25
10	0.50
20	1.00
50	2.50
100	5.00

The Danger of Multiple Comparisons Continued

```
# take alpha and divide by 3
new_alpha <-  .05/3

# divide new alpha in half for two-tails
new_alpha2 <- new_alpha/2

# lower & upper quantile
new_alpha2
1-new_alpha2

qt(c(.0083, .9917), df = 244)

qf(.9833, df1 = 1, df2 = 244)
```

```
[1] 0.008333333
[1] 0.9916667
[1] -2.412097  2.412097
[1] 5.807389
```

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	81.361	0.835		97.471	0.000	79.717	83.006
cond1	-13.021	1.145	-0.672	-11.368	0.000	-15.277	-10.765
cond2	-4.661	1.234	-0.223	-3.777	0.000	-7.092	-2.231

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	68.340	0.784		87.128	0.000	66.795	69.885
cond2	8.360	1.201	0.401	6.963	0.000	5.995	10.724
cond3	13.021	1.145	0.654	11.368	0.000	10.765	15.277

```
Linear hypothesis test

Hypothesis:
cond2 - cond3 = 0

Model 1: restricted model
Model 2: sleep ~ cond2 + cond3

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    245 14936
2    244 14111  1    825.14 14.268 0.0001993 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this particular case, all absolute values of t*, and the F* for the constraint, exceeds the more conservative critical value, so the interpretations would remain the same. That is, all comparisons for differences in sleep efficiency are significant.

Including Additional Predictors

We can easily include additional predictors in a model with a categorical predictor. We will add age (centered at 30), anxiety at the start of the study (centered at the mean), and a binary (already dummy coded) indicator of involvement with prior sleep interventions (recall that 0 = no prior involvement, 1= prior involvement).

```
# center predictors
slp_females <- slp_females %>%
  mutate(age30 = age-30,
         anxiety_m = anxiety - mean(anxiety))

# fit model
m4 <- lm(data = slp_females, sleep ~ cond2 + cond3 + prior + age30 + anxiety_m)
ols_regress(m4)
```

```
qt(c(.025, .975), 241)          [1] -1.969856   1.969856
qf(.95, df1 = 5, df2 = 241)     [1] 2.251492
```

```
                        Model Summary
------------------------------------------------------------
R                        0.662       RMSE              7.132
R-Squared                0.439       Coef. Var         9.498
Adj. R-Squared           0.427       MSE              50.861
Pred R-Squared           0.412       MAE               5.513
------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
                 Sum of
                Squares         DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression     9581.749          5       1916.350    37.678    0.0000
Residual      12257.466        241         50.861
Total         21839.215        246
----------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta       t        Sig      lower    upper
----------------------------------------------------------------------------------------
(Intercept)     66.605         1.120                 59.450    0.000     64.398   68.812
      cond2      9.219         1.140        0.442     8.086    0.000      6.973   11.465
      cond3     13.221         1.083        0.664    12.210    0.000     11.088   15.354
      prior      0.114         1.039        0.005     0.110    0.913     -1.932    2.161
      age30      0.096         0.032        0.142     2.947    0.004      0.032    0.160
  anxiety_m     -2.712         0.505       -0.262    -5.372    0.000     -3.707   -1.718
----------------------------------------------------------------------------------------
```

The F* is 37.678 — this exceeds our critical value of F (the p-value is less than alpha). Our model predicts a significant portion of the variability in sleep efficiency (about 44%).

Now, the coefficients for cond2 and cond3 are not capturing simple mean differences between each condition and the reference group, but rather adjusted mean differences — that is, adjusted for the other control variables. Here's how each slope is interpreted.

intercept: the predicted sleep efficiency score for a 30 year old female in condition 1 with no prior intervention history and an average level of anxiety.

cond2: holding constant prior involvement in a sleep intervention, age, and anxiety, we expect females in condition 2 to have a sleep efficiency score 9.22 units higher than females in condition 1.

cond3: holding constant prior involvement in a sleep intervention, age, and anxiety, we expect females in condition 3 to have a sleep efficiency score 13.22 units higher than females in condition 1.

prior: holding constant condition, age, and anxiety, we expect females with prior intervention involvement to have a sleep efficiency score .11 units higher than females without prior intervention involvement— however, this slope is not significantly different from 0.

age30: holding constant condition, prior intervention involvement, and anxiety, each one unit increase in age is associated with a .10 unit increase in sleep efficiency.

anxiety_m: holding constant condition, prior intervention involvement, and age, each one unit increase in anxiety prior to the intervention is associated with a 2.71 unit decrease in sleep efficiency.

Note that all predictors except prior are statistically significant — that is the |t*| exceeds our calculated |critical t|, and you can see that the p-values are small (less than our selected alpha of .05).

Plot the results

Let's create a plot to present the adjusted means of sleep efficiency across the three conditions. We will hold prior condition at 1, age at 30, and anxiety at the mean. Notice that in the data_grid function, we use a group_by statement first, so that when the combos of predictor values are constructed it is done by group (i.e., condition). Without this, the combo of a score of 1 on cond2 and 1 on cond3 will be created, but that isn't a possible score — nobody can be in both condition 2 and condition 3.

```
predgrid <- slp_females %>%
  group_by(cond.f) %>%
  data_grid(cond2, cond3,
                    prior = 1,
                    age30 = 0,
                    anxiety_m = 0) %>%
  ungroup()

predictions <- predict(m4, predgrid, interval = "confidence") %>%
  as_data_frame()

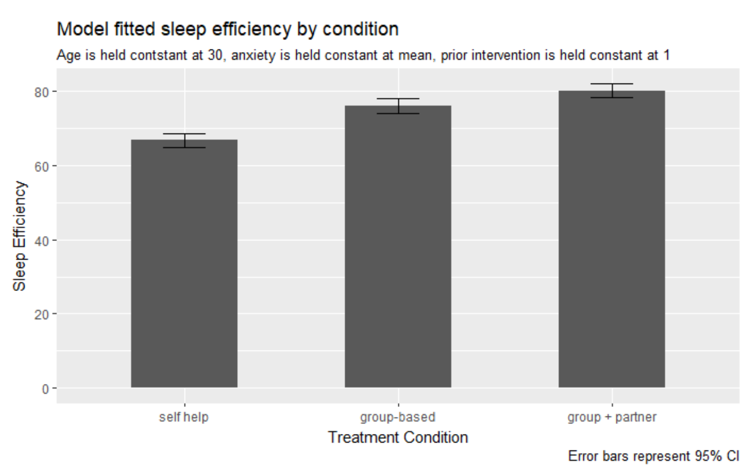
adjmeans <- cbind(predgrid, predictions)

ggplot(adjmeans, aes(x=cond.f, y=fit, ymin = lwr, ymax = upr)) +
  geom_bar(stat = "identity", width = .5) +
  geom_errorbar(color="black", width = .2) +
  labs(title = "Model fitted sleep efficiency by condition",
       subtitle = "Age is held contstant at 30, anxiety is held constant at mean, prior intervention is held constant at 1",
       x = "Treatment Condition", y = "Sleep Efficiency",
       caption = "Error bars represent 95% CI")
```

Sleep_Notebook.Rmd* predgrid adjmeans

Filter

	cond.f	cond2	cond3	prior	age30	anxiety_m	fit	lwr	upr
1	self help	0	0	1	0	0	66.71960	64.88963	68.54957
2	group-based	1	0	1	0	0	75.93839	73.97799	77.89878
3	group + partner	0	1	1	0	0	79.94026	78.10508	81.77544



One Last Plot that Incorporates Two Categorical Variables

We can modify the last plot to show differences by prior involvement status as well. This isn't of huge interest to us because prior involvement doesn't seem to be important, but it might be useful to you for other applications, so let's see how it works.

```
predgrid <- slp_females %>%
  group_by(cond.f) %>%
  data_grid(cond2, cond3, prior,
            age30 = 0,
            anxiety_m = 0) %>%
  ungroup()

predictions <- predict(m4, predgrid, interval = "confidence") %>%
  as_data_frame()

adjmeans <- cbind(predgrid, predictions)

ggplot(adjmeans, aes(x=cond.f, y=fit, ymin = lwr, ymax = upr, group = factor(prior), fill = factor(prior))) +
  geom_bar(position="dodge", stat = "identity", width = .5) +
  geom_errorbar( position = position_dodge(.5), colour="black", width = .2) +
  guides(fill = guide_legend("Prior Involvement")) +
  labs(title = "Model fitted sleep efficiency by condition and prior involvement",
     subtitle = "Age is held contstant at 30, anxiety is held constant at mean",
     x = "Treatment Condition", y = "Sleep Efficiency",
     caption = "Error bars represent 95% CI")
```

	cond.f	cond2	cond3	prior	age30	anxiety_m	fit	lwr	upr
1	self help	0	0	0	0	0	66.60532	64.39836	68.81228
2	self help	0	0	1	0	0	66.71960	64.88963	68.54957
3	group-based	1	0	0	0	0	75.82411	73.36162	78.28659
4	group-based	1	0	1	0	0	75.93839	73.97799	77.89878
5	group + partner	0	1	0	0	0	79.82598	77.39077	82.26119
6	group + partner	0	1	1	0	0	79.94026	78.10508	81.77544