WELCOME TO PSY 653 LAB!

MODULE 04:

CATEGORICAL PREDICTORS IN REGRESSION & NONLINEAR REGRESSION

* Thanks to Gemma Wallace for her help with these slides

OBJECTIVES

- Part 1: Categorical predictors in regression models
- Part 2: Nonlinear regression with continuous variables

PART 1: CATEGORICAL VARIABLES IN REGRESSION MODELS

DEMO OVERVIEW

We are practicing three different ways of examining categorical predictors in a regression framework:

1) Dummy coding
2) Effect coding
3) Contrast coding

Note: our outcome variables are continuous in each of these examples

CREATE A NEW R-PROJECT AND R-NOTEBOOK!

Download the “slpdata.csv” file from Canvas and save it into your R-project file

LOAD LIBRARIES

````
# Load Libraries
```{r}
library(tidyverse)
library(olsrr)
library(psych)
```
````

READ IN DATA

````
# Read in dataset
```{r}
slp <- read_csv("slpdata.csv")
```

Parsed with column specification:
cols(
  cond = col_double(),
  prior = col_double(),
  age = col_double(),
  anxiety = col_double(),
  hygiene = col_double(),
  support = col_double(),
  sleep = col_double(),
  lifesat = col_double(),
  sex = col_double(),
  id = col_double()
)
````

This is the same slpdata we've used in previous lab activities.

REDUCE DATASET TO JUST VARIABLES OF INTEREST

```
# Reduce dataset to just variables of interest
```{r}
slp <- select(slp, cond, hygiene)
```
```

DESCRIBE THE VARIABLES

```
# Describe variables
```{r}
describe(slp)
```
```

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
cond	1	600	2.00	0.82	2.00	2.00	1.48	1.00	3.00
hygiene	2	600	5.99	1.57	6.05	6.04	1.57	1.68	9.74

2 rows | 1-10 of 13 columns

OUR RESEARCH QUESTION FOR PARTS 1-3:

To what extent do treatment condition predict sleep hygiene?

We will show three different ways to approach this question using different coding methods for the categorical predictor variables: dummy coding, effect coding, and contrats coding

These methods are similar, but the coding and interpretations are slightly different.

PART 1: DUMMY CODING

WHAT IS **DUMMY CODING** AND WHY USE IT?

It is one of the most common and simplest approaches to evaluating categorical predictors in psychology

Dummy coding allows you to compare the mean difference between two levels of a categorical variable: the level that is coded as a 1 versus the level that is coded as 0

- You can specify any level of the variable to be the reference group (i.e., the level coded as 0)
- Create a new “dummy coded” binary variable for every comparison you want to make between two groups

DUMMY CODING

- For dummy coding, we will be converting categorical variables into a series of binary variables.
- For all but one of the levels of the categorical variable, a new variable will be created that has a value of 1 for each observation at that level and 0 for all others.

Level of race	New variable 1 (x1)	New variable 2 (x2)	New variable 3 (x3)
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (white)	0	0	0

Reference category

https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/

SPECIFY DUMMY CODES

Syntax of ifelse() function

```
ifelse(test_expression, x, y)
```

TRUE | FALSE

```
## Specify dummy codes for categorical predictors
```{r}

slp <- mutate(slp,
              cond2 = ifelse(cond == 2, 1, 0),
              cond3 = ifelse(cond == 3, 1, 0))
```
```

We created two new variables:
cond2 is a dummy coded binary variable in which condition 2 is coded as 1 and condition 1 is coded as 0. This variable allows us to compare the mean difference in Y between conditions 2 and 1.

cond3 is a is a dummy coded binary variable in which condition 3 is coded as 1 and condition 1 is coded as 0. This variable allows us to compare the mean difference in Y between conditions 3 and 1.

RUN MODEL WITH DUMMY CODED CONDITION VARIABLE

```
## Run model
```{r}
m1 <- lm(hygiene ~ cond2 + cond3, data = slp)
ols_regress(m1)
```
```

```r
## Run model
```{r}
m1 <- lm(hygiene ~ cond2 + cond3, data = slp)
ols_regress(m1)
```
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.630       RMSE                1.224
R-Squared               0.397       Coef. Var          20.418
Adj. R-Squared          0.395       MSE                 1.498
Pred R-Squared          0.391       MAE                 0.973
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                 ANOVA
-----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
-----------------------------------------------------------------------
Regression     589.467          2        294.733    196.774     0.0000
Residual       894.205        597          1.498
Total         1483.671        599
-----------------------------------------------------------------------

                            Parameter Estimates
--------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta      t        Sig     lower    upper
--------------------------------------------------------------------------------------
(Intercept)      4.660         0.087               53.853    0.000    4.490    4.830
      cond2      1.627         0.122        0.488  13.293    0.000    1.386    1.867
      cond3      2.374         0.122        0.712  19.399    0.000    2.134    2.615
--------------------------------------------------------------------------------------
```

Interpretations

Intercept: The predicted sleep hygiene score when all x variables are zero, so participants in Condition 1.

cond2: the predicted difference in sleep hygiene score between participants in condition 2 compared to condition 1.

cond3: the predicted difference in sleep hygiene score between participants in condition 3 compared to condition 1.

PART 2: EFFECT CODING

WHAT IS EFFECT CODING AND WHY USE IT?

Similar to dummy coding, except here, you are comparing one level of a categorical predictor to the *mean* of all of the levels.

Instead of asking “are two conditions different from each other?” using dummy coding, effect coding asks “is this condition different from average?”

While the "rule" in dummy coding is that only values of 0 and 1 are valid, the "rule" in effect coding is that all of the values in any new variable must sum to zero.

SPECIFY EFFECT CODING VARIABLES

Syntax of ifelse() function

```
ifelse(test_expression, x, y)
```

TRUE FALSE

````
```{r}
slp <- mutate(slp,

              cond2.ec = ifelse(cond == 2, 1, 0),
              cond3.ec = ifelse(cond == 3, 1, 0),

              cond2.ec = ifelse(cond == 1, (-1), cond2.ec),
              cond3.ec = ifelse(cond == 1, (-1), cond3.ec))
```
````

We created two new variables:

<u>cond2.ec</u> is a effect coded variable in which condition 2 is coded as 1, condition 1 is coded as -1, and condition 3 is coded as 0. This variable allows us to compare the mean difference in Y between condition 2 and the average score across all conditions.

<u>cond3.ec</u> is a is a dummy coded binary variable in which condition 3 is coded as 1, condition 1 is coded as -1, and condition 2 is coded as 0. This variable allows us to compare the mean difference in Y between condition 3 and the average score across all conditions.

```
## Run model without interaction first
```{r}
m2_no_interaction <- lm(hygiene ~ cond2.ec + cond3.ec, data = slp)
ols_regress(m2_no_interaction)
```
```

Model Summary

R	0.630	RMSE	1.224
R-Squared	0.397	Coef. Var	20.418
Adj. R-Squared	0.395	MSE	1.498
Pred R-Squared	0.391	MAE	0.973

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	589.467	2	294.733	196.774	0.0000
Residual	894.205	597	1.498		
Total	1483.671	599			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	5.994	0.050		119.969	0.000	5.896	6.092
cond2.ec	0.293	0.071	0.152	4.149	0.000	0.154	0.432
cond3.ec	1.041	0.071	0.540	14.726	0.000	0.902	1.179

Interpretations

Intercept: In effect coding, the intercept is the grand mean of sleep hygiene across all the three treatment groups

cond2: the predicted difference in sleep hygiene score between participants in condition 2 compared to the mean of all three treatment conditions.

cond3: the predicted difference in sleep hygiene score between participants in condition 3 compared to the mean of all three treatment conditions.

PART 3: CONTRAST CODING

WHAT IS CONTRAST CODING?

- Contrast coding is used to compare specific groups within your variable
- It is required that the contrasts are orthogonal. Remember, contrasts are orthogonal if:
  - The number of contrasts is equal to the df (# of groups – 1)
  - There are at least three groups
  - The pairwise products of the corresponding coefficients for each term sum to zero
- **Contrast 1**: Compares Condition 1 to Condition 2 (Ignoring condition 3)
- **Contrast 2**: Compares Condition 3 to Condition 1 & Condition 2

	Cond1	Cond2	Cond3
Contrast 1	.5	-.5	0
Contrast 2	-.5	-.5	1

These are orthogonal because: **(.5 * -.5) + (-.5 * -.5) + (0 * 1) = 0**

SPECIFY CONTRAST CODES

```
# Contrast Coding
```{r}

slp <- mutate(slp,
              contrast_1v2 = ifelse( cond == 1,  .5, ifelse(cond == 2, -.5,   0)),   # Compare condition 1 and condition 2
              contrast_3v12 = ifelse(cond == 1, -.5, ifelse(cond == 2, -.5,   1)))   # Compare condition 3 to condition 1 & 2
```
```

We created two new variables:

<u>Contrast_1v2</u> is a contrast coded variable in which **condition 1** is coded as .5, **condition 2** is coded as -.5 and **condition 3** is coded as 0. This contrast compares condition 1 to condition 2.

<u>Contrast_3v12</u> is a contrast coded variable in which **condition 1** is coded as -.5, **condition 2** is coded as -.5, and **condition 3** is coded as 1. This contrast compares condition 3 to the average of conditions 1 & 2

Contrast Coded Model

```{r}
m3 <- lm(hygiene ~ contrast_1v2 + contrast_3v12, data = slp)
ols_regress(m3)
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.630       RMSE                 1.224
R-Squared               0.397       Coef. Var           20.418
Adj. R-Squared          0.395       MSE                  1.498
Pred R-Squared          0.391       MAE                  0.973
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
-----------------------------------------------------------------------
Regression     589.467          2        294.733    196.774     0.0000
Residual       894.205        597          1.498
Total         1483.671        599
-----------------------------------------------------------------------

                                  Parameter Estimates
---------------------------------------------------------------------------------------------
        model      Beta    Std. Error    Std. Beta          t        Sig       lower     upper
---------------------------------------------------------------------------------------------
  (Intercept)     5.994         0.050                 119.969      0.000       5.896     6.092
 contrast_1v2    -1.627         0.122       -0.422    -13.293      0.000      -1.867    -1.386
contrast_3v12     1.041         0.071        0.468     14.726      0.000       0.902     1.179
---------------------------------------------------------------------------------------------
```

Intercept: In contrast coding, the intercept is the grand mean of sleep hygiene across all the three treatment groups.

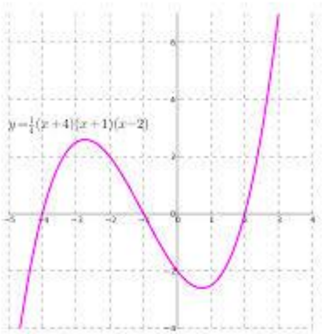
Contrast_1v2: The difference between condition 1 and condition 2. It is statistically significant.

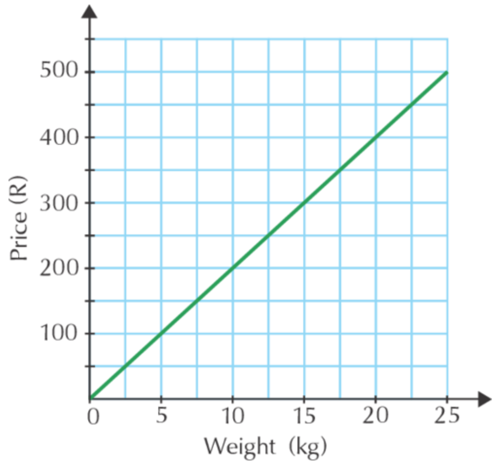
Contrast_3v12: The difference between the average of conditions 1 & 2 to condition 3. It is statistically significant.

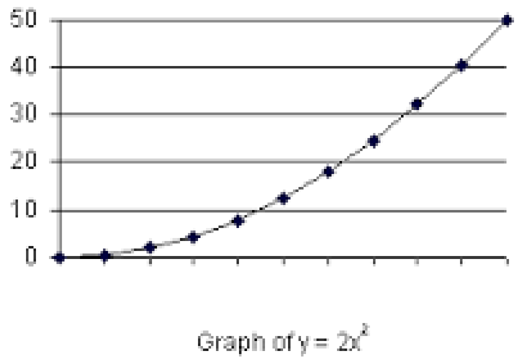
ADDITIONAL RESOURCES ON CATEGORICAL VARIABLE CODING SYSTEMS:

- https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/
- http://www.jds-online.com/files/JDS-563.pdf

PART 2: NONLINEAR REGRESSION







NONLINEAR REGRESSION

- Nonlinear regression is used to assess models in which there is *not* a linear trend
- We can see quadratic, cubic, or even quartic effects. Other types of nonlinear trends are log transformed trends.

CREATE A NEW R-PROJECT AND R-NOTEBOOK!

Download the “cogtest.csv” file from Canvas and save it into your R-project file

LOAD LIBRARIES

````
## Load libraries
```{r}
library(psych)
library(tidyverse)
library(olsrr)
```
````

NEW DATASET DESCRIPTION

Researchers were interested in the effect of time spent in practice on the performance of a visual discrimination task. Subjects were randomly assigned to different levels of practice, following which a test of visual discrimination is administered, and the number of correct responses is recorded for each subject. 40 subjects were randomly assigned to practice 0 minutes, 2 minutes, 4 minutes, 6 minutes, 8 minutes, 10 minutes, 12 minutes, or 14 minutes.

There are three variables:

Subject = subject ID

practice = minutes spent practicing, this was assigned by the experimenter

score = the number of correct answers on the test

````
```{r}
cog <- read_csv("cogtest.csv")
```
````

```
Parsed with column specification:
cols(
  subject = col_double(),
  practice = col_double(),
  score = col_double()
)
```

DESCRIBE DATA

```
# describe data
```{r}
describe(cog)
```
```

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
subject	1	40	20.50	11.69	20.50	20.50	14.83	1.00	40.00
practice	2	40	7.00	4.64	7.00	7.00	5.93	0.00	14.00
score	3	40	16.38	7.67	19.04	17.18	6.69	1.06	25.54

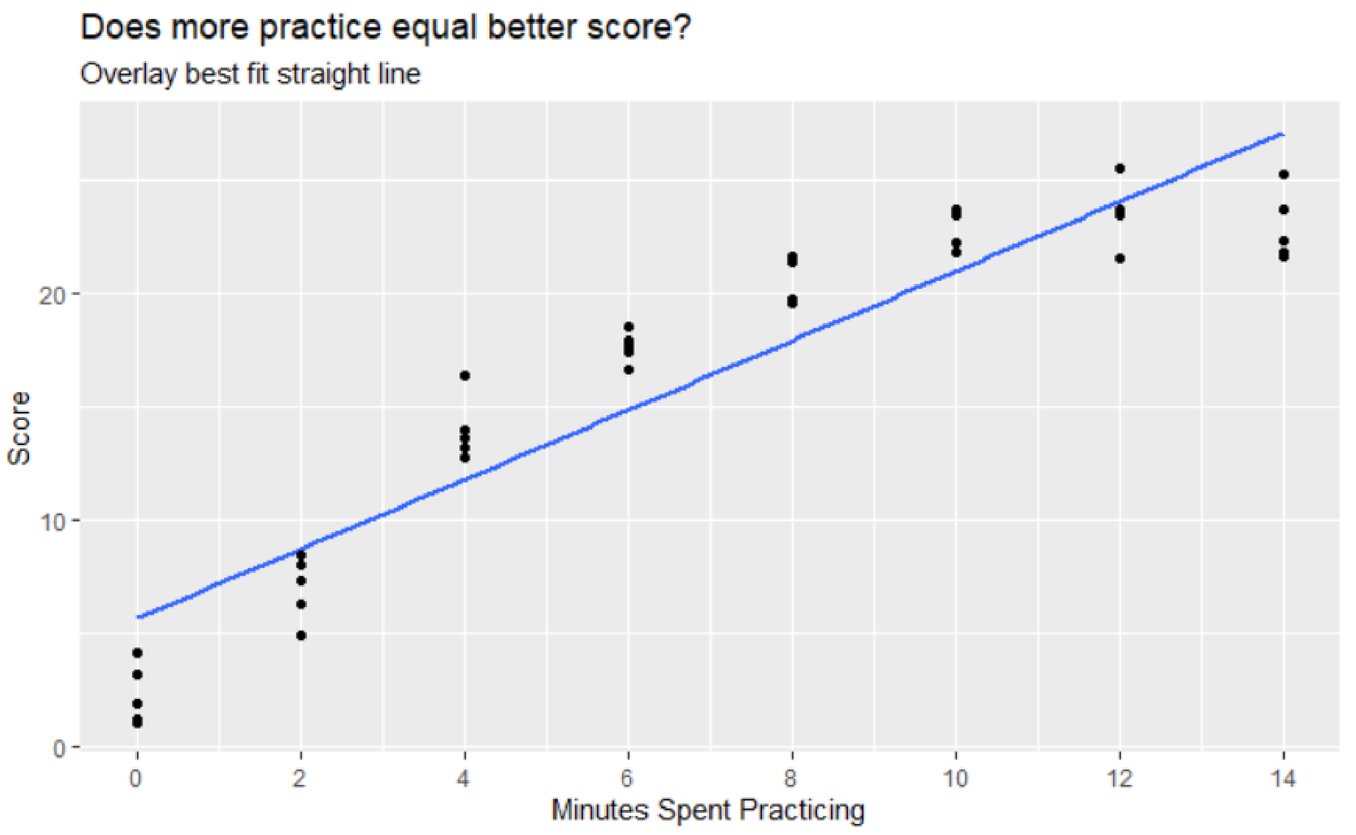
3 rows | 1-10 of 13 columns

PLOT A LINEAR RELATIONSHIP

```
## As a linear relationship
```{r}
ggplot(cog, aes(x = practice, y = score)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point() +
  scale_x_continuous(limits=c(0,14), breaks = seq(0, 14, by = 2)) +
  labs(title = "Does more practice equal better score?", subtitle = "Overlay best fit straight line",
       x = "Minutes Spent Practicing", y = "Score")

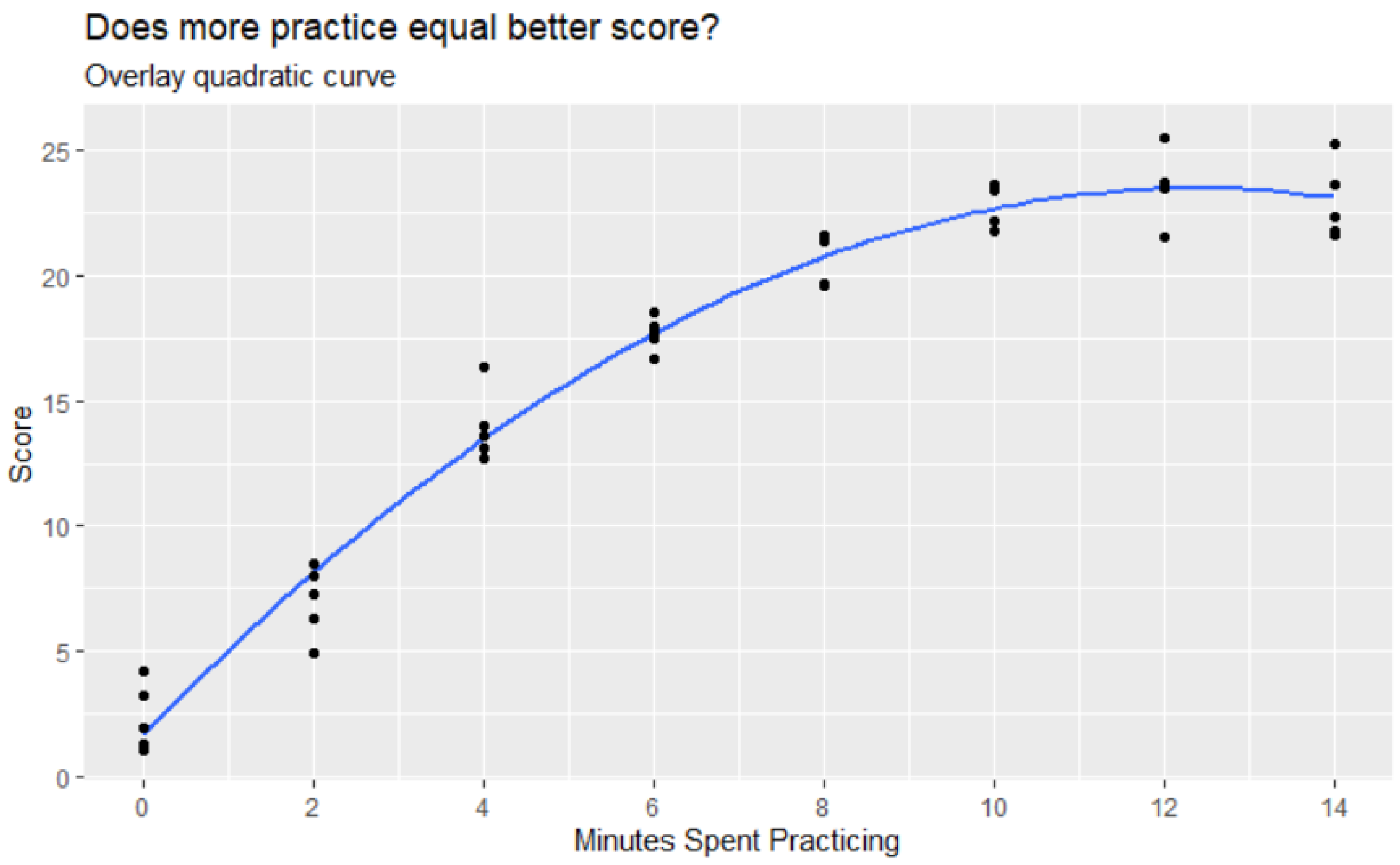

```
```

`geom_smooth()` using formula 'y ~ x'



PLOT A QUADRATIC RELATIONSHIP

```
## Plot data with a quadratic function
```{r}
# plot data with quadratic function
ggplot(cog, aes(x = practice, y = score)) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
  geom_point() +
  scale_x_continuous(limits=c(0,14), breaks = seq(0, 14, by = 2)) +
  labs(title = "Does more practice equal better score?", subtitle = "Overlay quadratic curve",
       x = "Minutes Spent Practicing", y = "Score")
```
```



MUTATE THE PRACTICE VARIABLE TO QUADRATIC AND CUBIC

```
# Mutate variables for polynomial regression: Get quadratic and cubic variables
```{r}
cog <- mutate(cog,
                 practice2 = practice^2,
                 practice3 = practice^3)
```
```

```r
# Run linear, quadratic and cubic models
## Linear model
```{r}
mod_lin <- lm(score ~ practice, data = cog)
ols_regress(mod_lin)
```
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.925       RMSE                 2.952
R-Squared               0.856       Coef. Var           18.017
Adj. R-Squared          0.852       MSE                  8.713
Pred R-Squared          0.837       MAE                  2.512
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
----------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square       F          Sig.
----------------------------------------------------------------------
Regression    1964.353         1       1964.353    225.462     0.0000
Residual       331.078        38          8.713
Total         2295.431        39
----------------------------------------------------------------------

                                  Parameter Estimates
--------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
--------------------------------------------------------------------------------------
(Intercept)     5.678         0.852                 6.664    0.000    3.953    7.403
   practice     1.529         0.102        0.925   15.015    0.000    1.323    1.735
--------------------------------------------------------------------------------------
```

The model testing the linear effect between practice and score explained 85.6% of the variance in score, and the linear trend was statistically significant at p<0.001.

This model fits the data pretty well, but since we observed a potential curved relationship when we plotted the data, there could be a better way to examine this relationship.

Quadratic model

```{r}
mod_quad <- lm(score ~ practice + practice2, data = cog)
ols_regress(mod_quad)
```

```
                        Model Summary
-------------------------------------------------------------
R                       0.987       RMSE                1.276
R-Squared               0.974       Coef. Var           7.787
Adj. R-Squared          0.972       MSE                 1.627
Pred R-Squared          0.969       MAE                 0.945
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
---------------------------------------------------------------------
               Sum of
              Squares        DF     Mean Square       F          Sig.
---------------------------------------------------------------------
Regression    2235.214        2        1117.607    686.706     0.0000
Residual        60.217       37           1.627
Total         2295.431       39
---------------------------------------------------------------------

                                  Parameter Estimates
------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta        t        Sig     lower     upper
------------------------------------------------------------------------------------------
(Intercept)       1.703         0.480                 3.547      0.001     0.730     2.676
   practice       3.517         0.160        2.127   21.949      0.000     3.192     3.841
  practice2      -0.142         0.011       -1.250  -12.901      0.000    -0.164    -0.120
------------------------------------------------------------------------------------------
```

The model testing the linear and quadratic effects between practice and score explained 97.4% of the variance in score, which is 11.8% higher than the model that only tested the linear relation.

The quadratic term is statistically significant, indicating that there is a substantial curve to the relation between practice and score (i.e., it's not linear). We need to maintain the quadratic term in the model.

```r
## Cubic model
```{r}
mod_cub <- lm(score ~ practice + practice2 + practice3, data = cog)
ols_regress(mod_cub)
```
```

```
                        Model Summary
--------------------------------------------------------------
R                       0.987       RMSE                1.270
R-Squared               0.975       Coef. Var           7.750
Adj. R-Squared          0.973       MSE                 1.612
Pred R-Squared          0.968       MAE                 0.922
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
----------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square       F        Sig.
----------------------------------------------------------------------
Regression    2237.395         3        745.798    462.622    0.0000
Residual        58.036        36          1.612
Total         2295.431        39
----------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
-----------------------------------------------------------------------------------------
(Intercept)     1.988         0.537                 3.703    0.001     0.899    3.077
   practice     3.144         0.358        1.902    8.786    0.000     2.418    3.870
  practice2    -0.071         0.062       -0.624   -1.140    0.262    -0.197    0.055
  practice3    -0.003         0.003       -0.416   -1.163    0.252    -0.009    0.003
-----------------------------------------------------------------------------------------
```

We tested the cubic term to determine if there is a second bend to the relationship between practice and score.

The cubic term is not significant, indicating that there is not a second bend to the relationship.

Therefore, the quadratic model is the best fit for these data.