PS303: Week 9

pp.457-469

Recap

- Categorical data contrasts estimate whether group *proportions* vary significantly and meaningfully between/within categories
- Continuous data contrasts estimate whether group *means* (parametric tests) or *medians* (non-parametric tests) vary between/within categories
- *Significance* is informed by p-values ($p < .05$), which tells us how likely the observed data would be *assuming* H_O is correct
- Effect sizes (e.g., Cohen's d, Cramer's V, η_p^2) tell us whether a statistically significant difference is practically meaningful
- For $k \leq 2$ groups, we run independent or pairwise contrasts to estimate differences
- For $k > 2$ groups, we run Analyses of Variance (ANOVAs) to estimate for "overall" effects.
- Significant ANOVAs can be followed by post-hoc (after the fact) tests, which identify whether any groups notably vary across each other

Our strategies so far involved estimating *differences* between conditions. This is fine for experimental/quasi-experimental designs, where specific conditions are identified beforehand. In many real-world scenarios however, experimentation is not possible. In such cases, generating valid **predictions** may be more feasible.

Linear Regression Models

```
require(kableExtra)
require(tidyverse)

# Prepare data
set.seed(23)
physical <- c(sample(c(100:150),10,replace = T),sample(c(70:110),10,replace = T))
depress  <- c(sample(c(40:50),10,replace = T),sample(c(45:65),10,replace = T))
id <- seq(1:20)
df <- cbind.data.frame(id,physical,depress)
df$id <- as.factor(df$id)
```

Let's assume you've collected data for your research report from $n = 20$ participants. You have collected two sets of scores that refer to **Physical Activity** and **Depression**.

ID	Physical Activity (minutes)	Depression scores
1	128	48
2	127	45
3	150	46
4	107	41
5	142	50
6	138	50
7	144	41
8	133	44
9	147	48
10	116	50
11	90	64
12	86	63
13	109	48
14	105	57
15	75	51
16	75	58
17	100	60
18	91	61
19	82	62
20	79	49

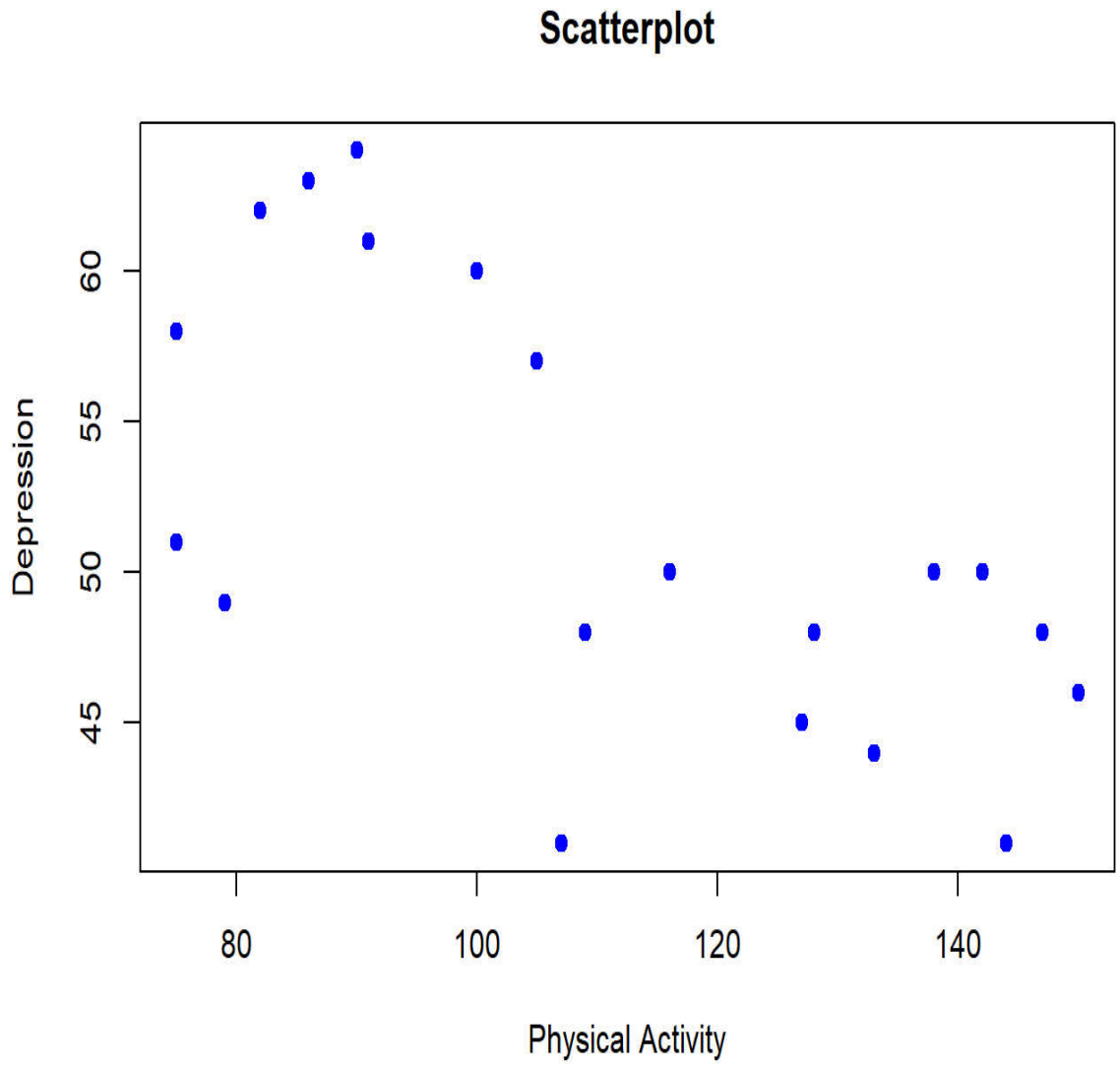
The two scales measure different qualities, and cannot be directly contrasted with one other.

Instead of looking at the differences *between* measures, can we assess whether one variable may *predict* another?

Drawing a scatterplot

We can begin by plotting the data to explore for any apparent linear trends using scatterplots.

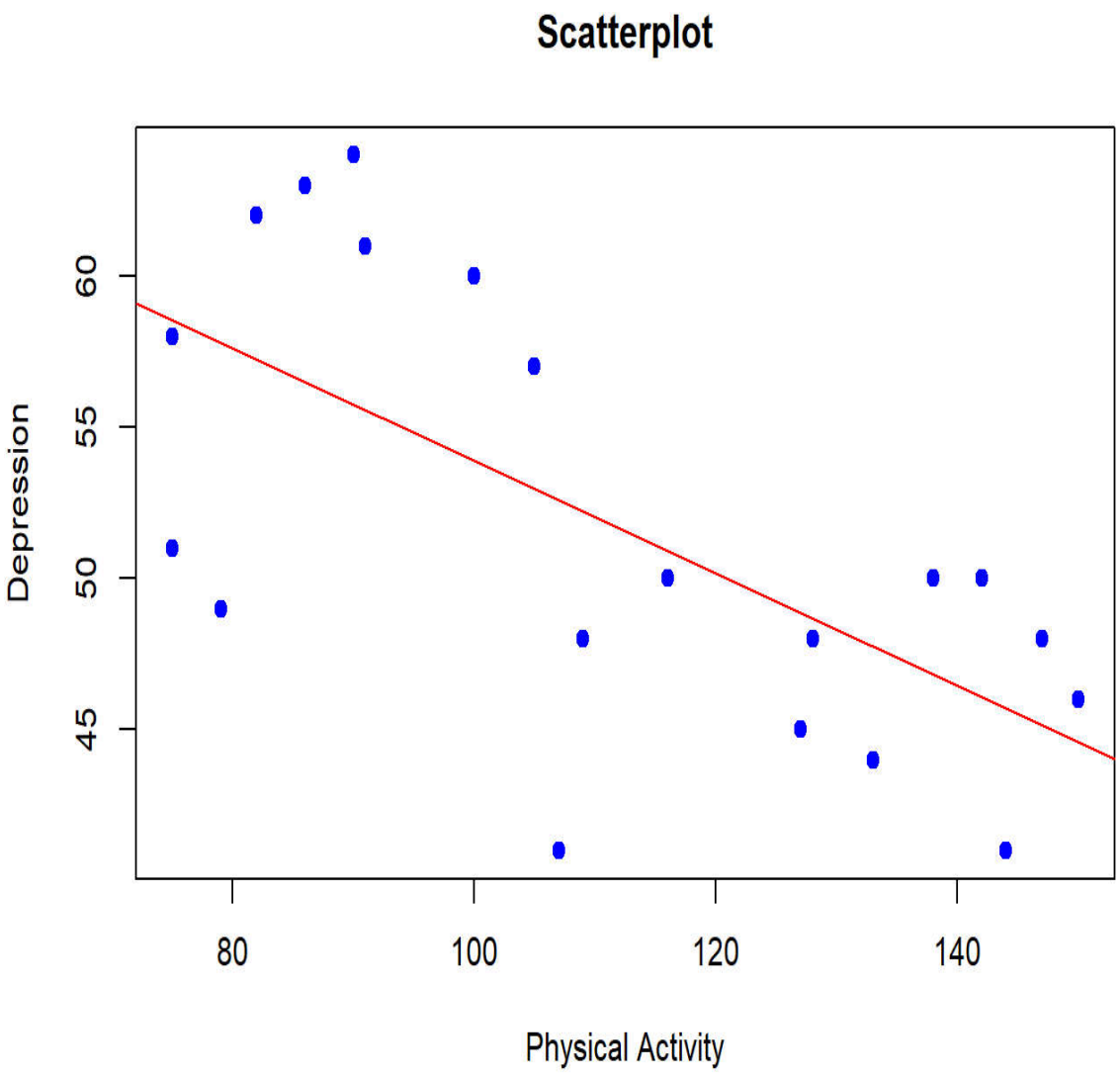
```
# Create scatterplot
attach(df)
plot(physical,depress,main="Scatterplot",
     xlab = "Physical Activity", ylab = "Depression",pch=19,
     col = "blue")
```



The data appears to be **negatively** associated - *more* time spent on physical activity corresponds with *lower* depression scores.

We can illuminate the relationship by drawing a straight line through the center of the plot

```
# Create scatterplot
attach(df)
plot(physical,depress,main="Scatterplot",
     xlab = "Physical Activity", ylab = "Depression",pch=19,
     col = "blue")
abline(lm(depress~physical),col="red")
```



The formula for drawing a straight line can be expressed as $y = mx + c$ where y and x represent the two *variables*, and m and c are the two *coefficients* (multiplier quantity). Specifically, m represents the *slope* of the line (how much will y change when x is incremented by a single unit) and c represents the estimated value of the dependent variable (y) when $x = 0$, otherwise known as the y-intercept.

The straight line going through the center of the data is a **regression line**. This can be estimated using the formula described earlier (with some modifications)

$$\hat{Y}_i = b_1 X_i + b_0$$

Y_i and X_i indicates i-th data points of the dependent and independent measures respectively. $\hat{Y}_i$* indicates the *estimated* data at the i-th point. The latter is the **predicted** data point, which we compare with the **observed** data point. The remaining values (b_0, b_1) indicate the *slope* and *intercept* of the regression line.

* The difference between the *predicted* and *observed* values is called *epsilon* (ϵ). For each i-th datapoint, we can estimate the residual, $\epsilon_i = Y_i - \hat{Y}_i$. We can apply this to the above formula to complete the linear regression model

$$\hat{Y}_i = b_1 X_i + b_0$$

$$Y_i - \epsilon_i = b_1 X_i + b_0$$

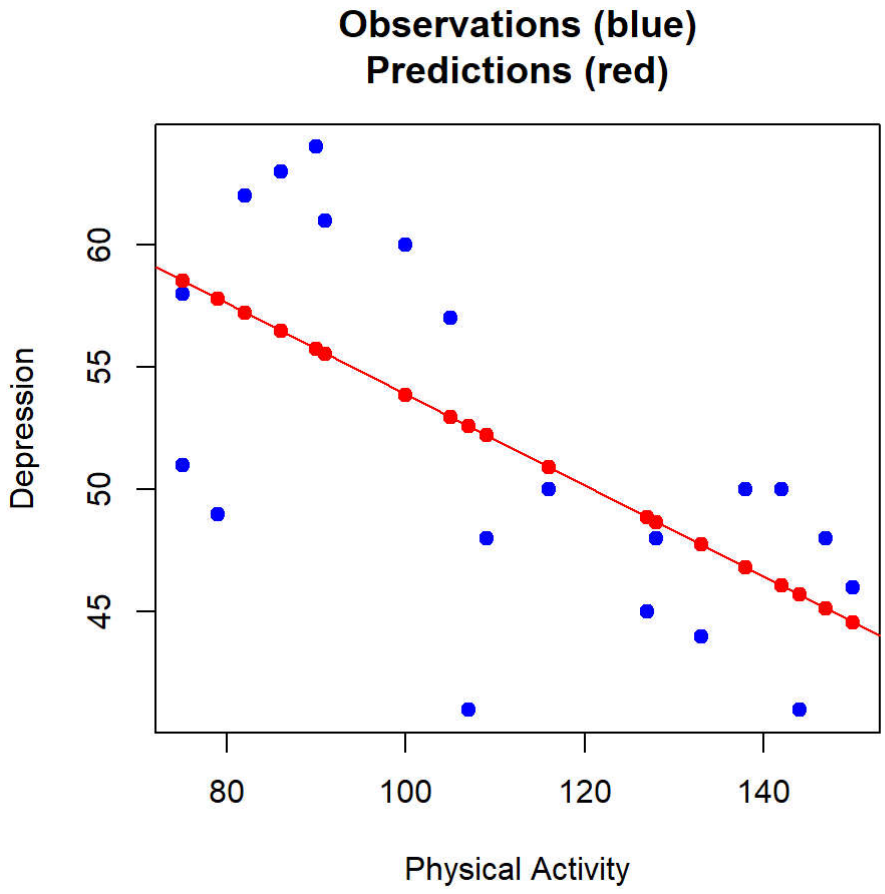
$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

The difference between *predicted* ($\hat{Y}$) and *observed* values (Y) are the *residuals* (ϵ). The closer the 'fit' between predicted and observed values (ie the smaller the residuals), the better the model.

```
lmod <- lm(df$depress~df$physical) # Regression model

cof1<- lmod$coefficients[1] # Intercept Coefficient
cof2<- lmod$coefficients[2] # Predictor Coefficient

# Compute predicted value from regression model coefficients
df$depress_pred <- cof1 + cof2*df$physical

# Compute residuals
df$depress_res <- df$depress - df$depress_pred

# Create new data table
df2 <- df

# Create table
df2.tab <- kableExtra::kbl(df2,
  booktabs = TRUE,
  col.names = c("ID", "Physical Activity (minutes)", "Depression scores","Depression Predicted",
"Depression Residuals"),
  align = c("l", "c", "c"))

df2.tab
```

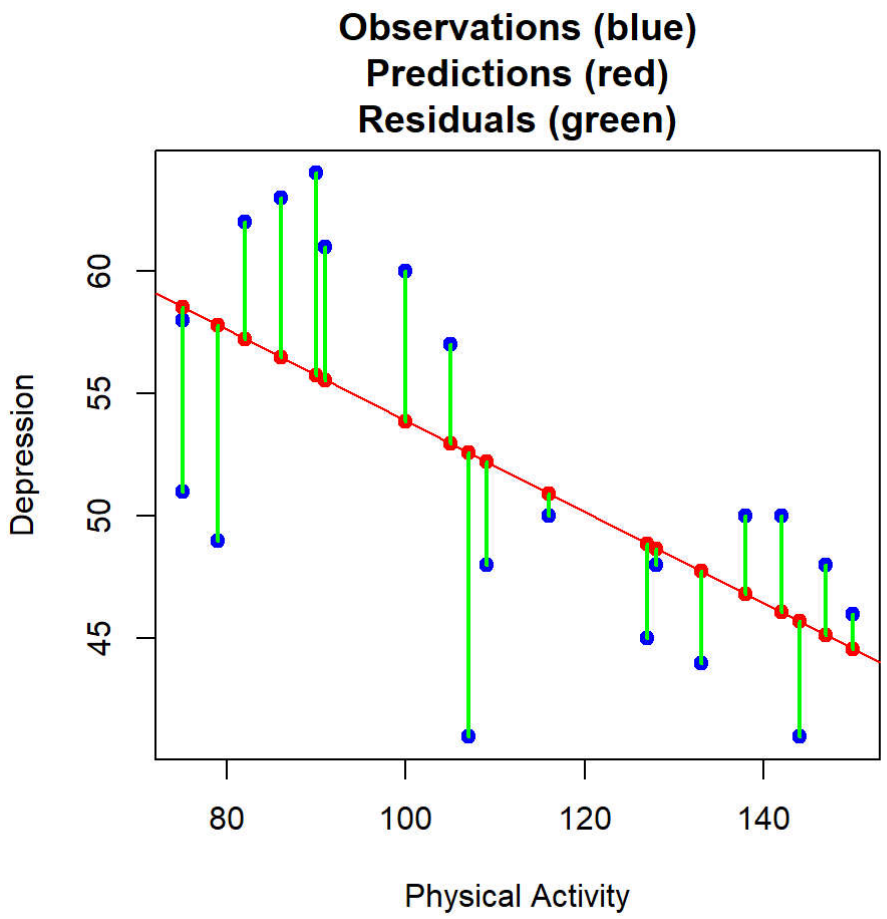
ID	Physical Activity (minutes)	Depression scores	Depression Predicted	Depression Residuals
1	128	48	48.67803	-0.6780347
2	127	45	48.86387	-3.8638660
3	150	46	44.58975	1.4102532
4	107	41	52.58049	-11.5804913
5	142	50	46.07640	3.9236030
6	138	50	46.81972	3.1802780
7	144	41	45.70473	-4.7047344
8	133	44	47.74888	-3.7488784
9	147	48	45.14724	2.8527594
10	116	50	50.90801	-0.9080099
11	90	64	55.73962	8.2603771
12	86	63	56.48295	6.5170521
13	109	48	52.20883	-4.2088288
14	105	57	52.95215	4.0478461
15	75	51	58.52709	-7.5270919
16	75	58	58.52709	-0.5270919
17	100	60	53.88131	6.1186898
18	91	61	55.55379	5.4462084
19	82	62	57.22627	4.7737270
20	79	49	57.78377	-8.7837668

```
# Create scatterplot with residual values only
attach(df)
plot(physical,depress,main="Observations (blue)\nPredictions (red)",
     xlab = "Physical Activity", ylab = "Depression",pch=19,
     col = "blue")
abline(lm(depress~physical),col="red")
points(physical,depress_pred,pch=19,col="red")  # Predicted points
```




```
detach()
```

```
# Create scatterplot with residual lines
attach(df)
plot(physical,depress,main="Observations (blue)\nPredictions (red)\nResiduals (green)",
     xlab = "Physical Activity", ylab = "Depression",pch=19,
     col = "blue")
abline(lm(depress~physical),col="red")
points(physical,depress_pred,pch=19,col="red")  # Predicted points
for (i in 1:20) { # Add residual lines
  lines(c(physical[i],physical[i]),
        c(depress[i],depress_pred[i]),
        lwd=2,col="green")
}
```




```
detach()
```

Entering `Physical Activity` and `Depression` as our X and Y variables into the earlier formula, we can estimate the *predicted* depression ($\hat{Y}_i = b_1 X_i + b_0$) scores and their corresponding residuals ($\epsilon_i = Y_i - \hat{Y}_i$).

How do we estimate the coefficients (ie $b_0, b_1 \ldots b_i$)?

Running LRMs

The goal for a LRM is to estimate coefficients that produce the smallest residuals possible (least discrepancy between predictions and observations). Formally, this would be the smallest sum of the *squared residuals*.

$$(Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + (Y_3 - \hat{Y}_3)^2 \ldots (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i \epsilon_i^2$$

The estimation process is known as **Ordinary Least Squares (OLS)** regression. We can estimate our coefficients using the `lm()` function (*lm = linear model*).

Lets create our data frame

```
# Individual IDs
id <- c(  1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
# Physical activity scores
physical <- c(
128, 127, 150, 107, 142, 138, 144, 133, 147, 116, 90, 86, 109, 105, 75, 75, 100, 91, 82, 79)
# Depression scores
depress <- c(  48, 45, 46, 41, 50, 50, 41, 44, 48, 50, 64, 63, 48, 57, 51, 58, 60, 61, 62, 49)
# Bind columns as data frame
df <- cbind.data.frame(id,physical,depress)
# Ensure that ID variable is stored as a factor
df$id <- as.factor(df$id)
```

We can estimate our coefficients

```
mod1 <- lm(data = df,formula=depress~physical)
mod1
```

```
##
## Call:
## lm(formula = depress ~ physical, data = df)
##
## Coefficients:
## (Intercept)     physical
##     72.4644      -0.1858
```

Our coefficients are 72.46 (*y*-intercept) and -0.19 (line slope). We can enter these values into the formula for a straight line to estimate predicted changes in *y*.

$$y = mx + c$$

becomes

$$y = -.186x + 72.464$$

Similar to ANOVAs, if we want to extract *p*-values, we call the `summary()` function.

```
summary(mod1)
```

```
##
## Call:
## lm(formula = depress ~ physical, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5805  -3.9501   0.4416   4.2293   8.2604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.46444    5.78642  12.523 2.53e-10 ***
## physical    -0.18583    0.05074  -3.662  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.73 on 18 degrees of freedom
## Multiple R-squared:  0.427,  Adjusted R-squared:  0.3951
## F-statistic: 13.41 on 1 and 18 DF,  p-value: 0.001783
```

Multiple predictors

A key feature of regression models is that we are not restricted to single predictor-outcome relationships. We can include multiple predictors to estimate how much of the total variance in output are explained by each IV. While we may in principle add "as many predictors" as we want, increasing parameters renders interpretation of a model increasingly complex. It's best to add predictors that are theoretically coherent with the predictions being made.

> - Suppose we find out after our initial analyses that the ages of our $n = 20$ varied between 18 and 45 years. You want to know whether people become less depressed as they get older. Specifically, you want to know whether **age** predicts variances in depression scores.

We can add participant age to our earlier data frame and run a linear model

```
df$age <- c(  23, 18, 25, 19, 23, 21, 21, 20, 18, 22, 35, 27, 40, 44, 40, 32, 44, 25, 30, 42  )
```

```
mod2 <- lm(data=df, depress~physical+age)
mod2
```

```
##
## Call:
## lm(formula = depress ~ physical + age, data = df)
##
## Coefficients:
## (Intercept)     physical          age
##    67.83818     -0.16622      0.08596
```

We can extract R^2 and p-values using the `summary()` function as before

```
summary(mod2)
```

```
##
## Call:
## lm(formula = depress ~ physical + age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6858  -3.7645   0.4118   4.0915   8.1131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.83818   11.84449   5.727 2.47e-05 ***
## physical    -0.16622    0.06772  -2.455   0.0252 *
## age          0.08596    0.19062   0.451   0.6577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.861 on 17 degrees of freedom
## Multiple R-squared:  0.4337, Adjusted R-squared:  0.3671
## F-statistic:  6.51 on 2 and 17 DF,  p-value: 0.007957
```

R^2 describes the proportion of variance that can be accounted for by the predictor variable(s). Our regression model estimated 43% of the variance noted across depression scores, with physical activity being the only *significant* predictor. Formally, we expand our earlier regression model by adding another coefficient to represent age.

$$\hat{Y_i} = b2X_{i2} + b1X_{i1} + b_0$$

$$\hat{Y_i} = b2_{Physical} + b1_{Age} + b_0$$

$$\hat{Y_i} = -.17_{Physical} + .09_{Age} + 67.84$$

We can estimate the variance in depression scores as a function of our predictors.

Hypothesis tests

Similar to earlier procedures, we can test the utility of our model using hypothesis tests.

- Does the model perform better than a *null* model?
  - H_0: There is no relationship between predictor and outcome variables
  - H_A: There is a relationship (in the manner our model predicts)
- Is a particular coefficient better than 0?
  - H_0: The target coefficient (b) is equivalent to the null ($b = 0$)
  - H_A: The target coefficient (b) is equivalent to the null ($b \neq 0$)

> - Fortunately, there are no new tests to run. The F-ratio and t-statistics outputs produced by the `summary()` of our linear model respectively inform us whether our models and/or coefficients vary significantly from null estimates.

Reporting outcomes

A multiple OLS regression was calculated to predict depression scores based on participants' physical activity level and age. A significant regression equation was found, $F(2, 17) = 6.51, p = .008$, with an R^2 of $.43$. Participant depression scores were equal to $67.84 - .17_{PHYSICAL} + .09_{AGE}$, with both predictors measured as continuous variables. Depression scores fell by $.17$ for each minute engaged in physical activity; depression scores increased by $.09$ for each additional year of age. Only physical activity significantly predicted depression, $t = 2.45, p = .025$.

We will continue discussing regression models the following week. **There are no labs for this week**.