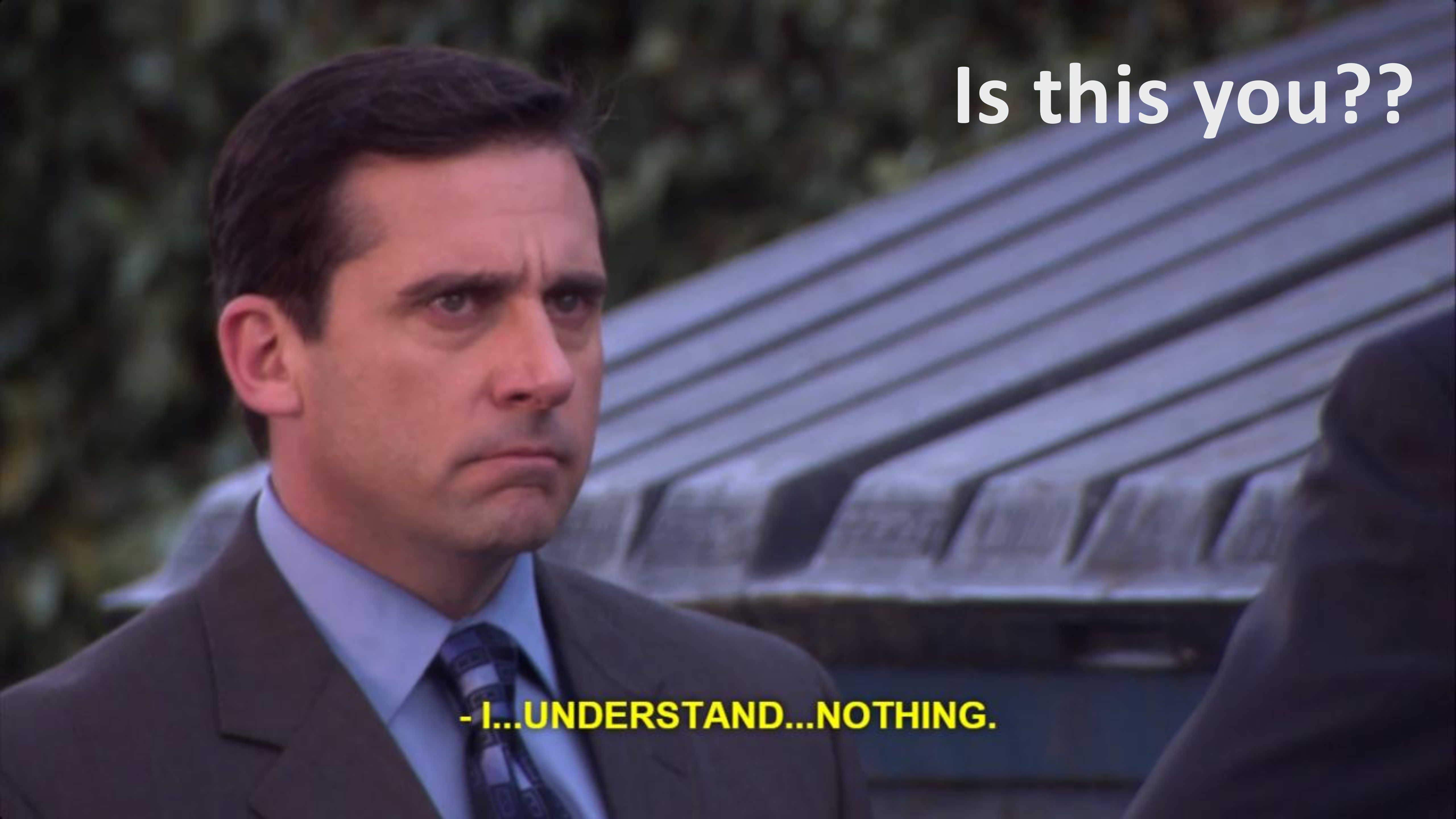

Is this you??
- I...UNDERSTAND...NOTHING.

EDUC 7610

Chapter 8

The Importance of Predictors

Fall 2018

Tyson S. Barrett, PhD

What do we mean by "important"?

1. Substantive Terms

2. Statistical Terms

What do we mean by "important"?

1 Substantive Terms

Guided by theory and subjective judgement

Not of focus in this chapter but still very important to any research project

What do we mean by “important”?

Several Ways to Quantify It:

1. How much of the variability in the outcome it explains (most common)
2. Amount that including a predictor in a model lowers the error in estimating outcome
3. The amount the outcome changes for a small change in the predictor

All are considered different ways to talk about the “effect size”

2 Statistical Terms

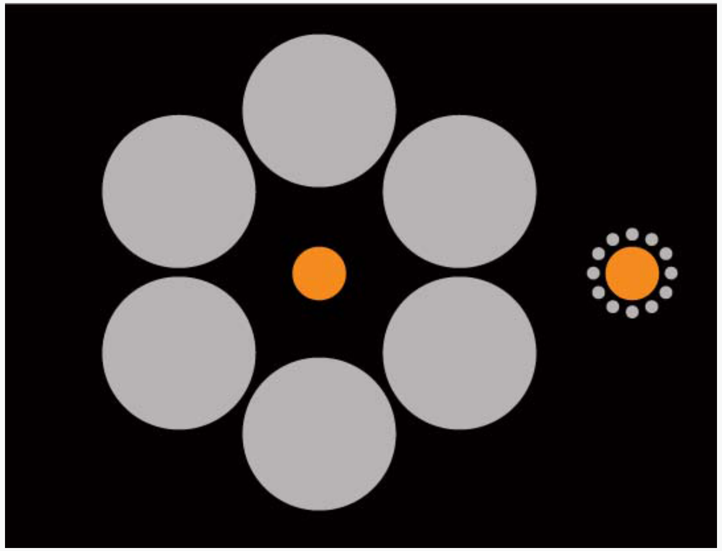
What is an "effect size"?

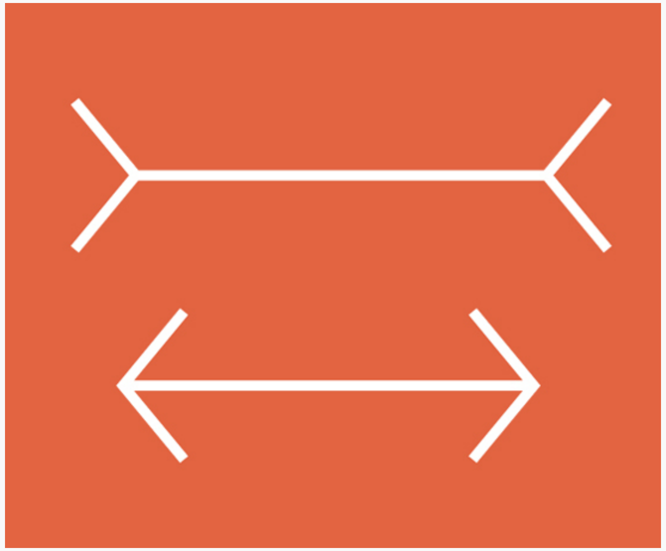
It is any quantification of the size of an effect

- Too often it is just associated with Cohen's d

Regression gives us several effect sizes:

- Regression coefficients (unstandardized and standardized)
- Multiple R (and R^2)
- Partial correlations (and partial r^2)





Three main ways to quantify size

1. How much of the variability in the outcome it explains (most common)
2. Amount that including a predictor in a model lowers the error in estimating outcome
3. The amount the outcome changes for a small change in the predictor

1. R^2 (Squared Correlations)

"It is almost taken as gospel that the square of a correlation is a measure of the importance of the relationship... But doing so can be misleading..." (pg. 213)

R^2 is the popular choice for discussing the effect size

- *The proportion of the variance in the outcome explained by the predictors*
- There are several important ways that we can quantify the size of the relationship in addition to R^2

1. R² (Squared Correlations)

Instead of squaring the correlations, we can use the correlations to understand how large a regression coefficient is compared to its maximum possible value *(in simple regression)*

$$b_1 = r_{XY}\frac{s_Y}{s_X}$$

Let's say we found the following from our regression

$$5 = .5\frac{10}{1}$$

We can replace our correlation with 1 (the max possible)

$$b_{1\,Max} = 1\frac{10}{1}$$

$$b_{1\,Max} = 10$$

1. R^2 (Squared Correlations)

Instead of squaring the correlations, we can use the correlations to understand how large a regression coefficient is compared to its maximum possible value *(in multiple regression)*

The ratio between b_j and its maximum possible value (with the same values of the other statistics)

$$pr_j = b_j s_{x_i} \sqrt{\frac{Tol_j}{V(Y.C)}}$$

The proportion of the variance in X_j not explained by the covariates

Variance in Y not explained by the covariates

Partial correlation for X_j

The observed regression coefficient

Standard deviation of X_j

1. R^2 (Squared Correlations)

Instead of squaring the correlations, we can use the correlations to understand how large a regression coefficient is compared to its maximum possible value *(in multiple regression)*

The ratio between b_j and its maximum possible value (with the same values of the other statistics)

$$pr_j = b_j s_{x_i} \sqrt{\frac{Tol_j}{V(Y.C)}}$$

If pr_j is .5 that suggests that b_j is half of its maximum possible value (given the other values remain constant)

2. Amount that including a predictor in a model lowers the error in estimating outcome

If prediction is particularly important for our analysis, then this is the most important measure of effect size

$$RMSE = \sqrt{MS_{residual}}$$

Smaller is better

3. The amount the outcome changes for a small change in the predictor

If both X and Y are in meaningful units:

- Use unstandardized b_j

If not:

- Use standardized b_j

This can often be the most important aspect of the effect size

Relative Importance of Predictors in Same Model

We may want to compare two coefficients in the same model

Create 2 new variables based on the 2 old variables

$$X^{+} = \frac{X_1 + X_2}{2} \qquad X^{-} = \frac{X_1 - X_2}{2}$$

Use X^{+} and X^{-} in place of X_1 and X_2 in the regression model

- The estimate on X^{-} is equal to $X_1 - X_2$
- This gives us the comparison of the two (complete with a standard error, CI, and p-value)

Relative Importance of Predictors in Same Model

We may want to compare two coefficients in the same model

Original Model

```
gss %>%
  lm(income06 ~ natheal + nateduc,
     data = .) %>%
  summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    58090       1341  43.328   <2e-16 ***
natheal         1470       1188   1.238    0.216
nateduc        -2220       1409  -1.575    0.115
```

Comparing natheal and nateduc

```
gss %>%
  mutate(diff1 = (natheal + nateduc)/2,
         diff2 = (natheal - nateduc)/2) %>%
  lm(income06 ~ diff1 + diff2,
     data = .) %>%
  summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    58090       1341  43.328   <2e-16 ***
diff1           -749       1187  -0.631    0.528
diff2           3690       2320   1.591    0.112
```

Relative Importance of Predictors in Same Model

We may want to compare two coefficients in the same model

Original Model

```
gss %>%
  lm(income06 ~ natheal + nateduc,
     data = .) %>%
  summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    58090       1341  43.328   <2e-16 ***
natheal         1470       1188   1.238    0.216
nateduc        -2220       1409  -1.575    0.115
```

Comparing natheal and nateduc

```
gss %>%
  lm(income06 ~ natheal + nateduc,
     data = .) %>%
  car::linearHypothesis("natheal = nateduc")

Hypothesis:
natheal - nateduc = 0

Model 1: restricted model
Model 2: income06 ~ natheal + nateduc

  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1   1772 3.4557e+12
2   1771 3.4507e+12  1 4929102092 2.5297  0.1119
```

Relative Importance of Predictors in Same Model

Dominance Analysis

Which variable explains more of Y in light of all combinations of the covariates?

Complete vs. Partial