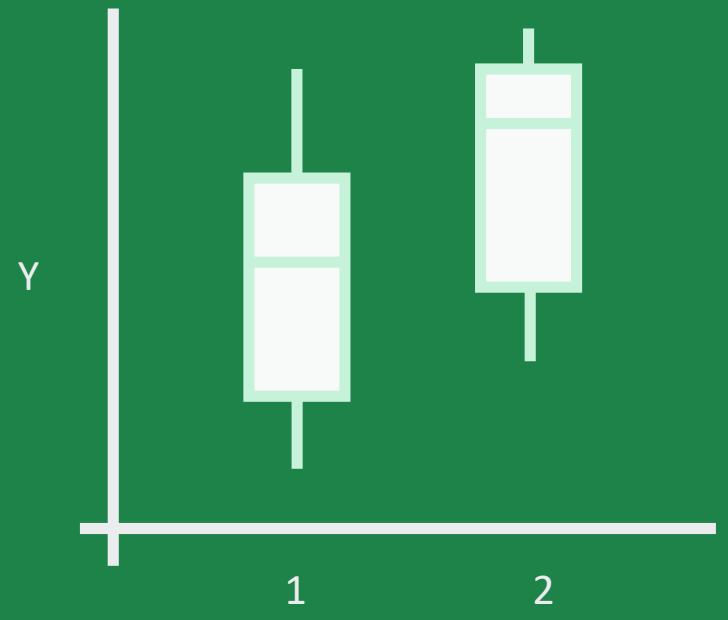
These are accurate. Very accurate.

EDUC 7610

Chapter 5

Extending Regression



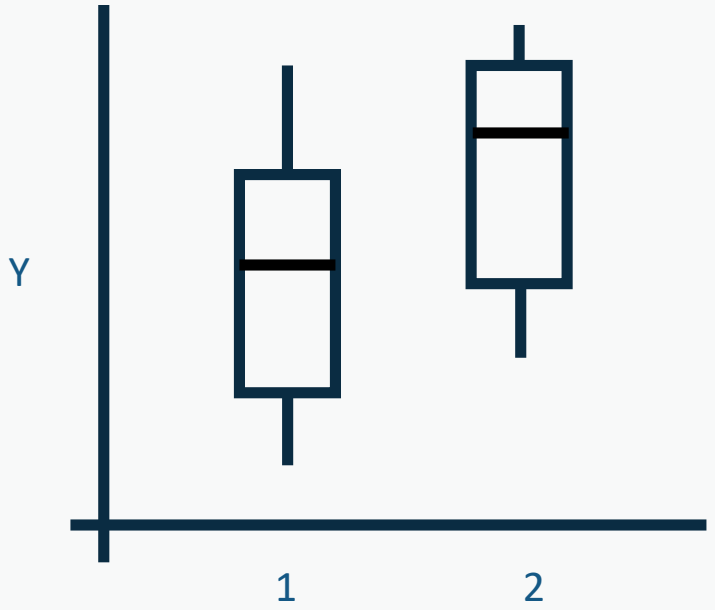
Fall 2018

Tyson S. Barrett, PhD

Categorical Predictors

Categorical predictors (e.g., disability status) can be integrated into a regression model with very little change in how it works

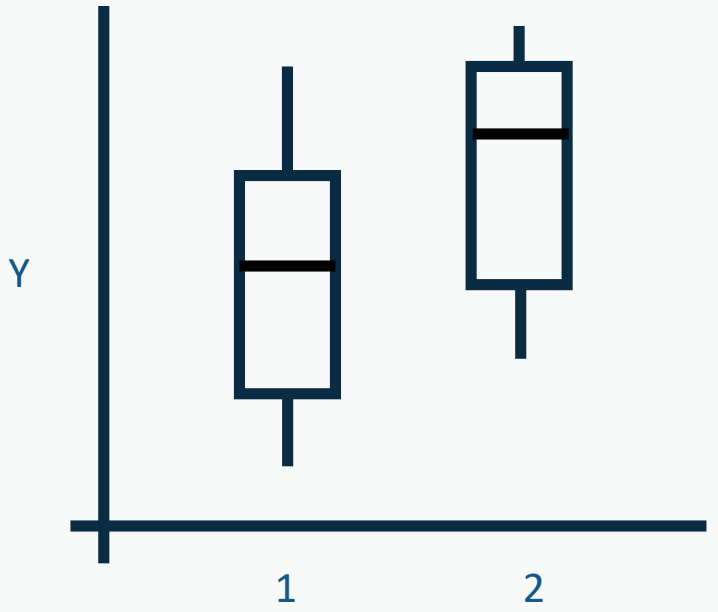
- *Now we are estimating a difference between means instead of a slope*



Categorical Predictors

Other names for it:

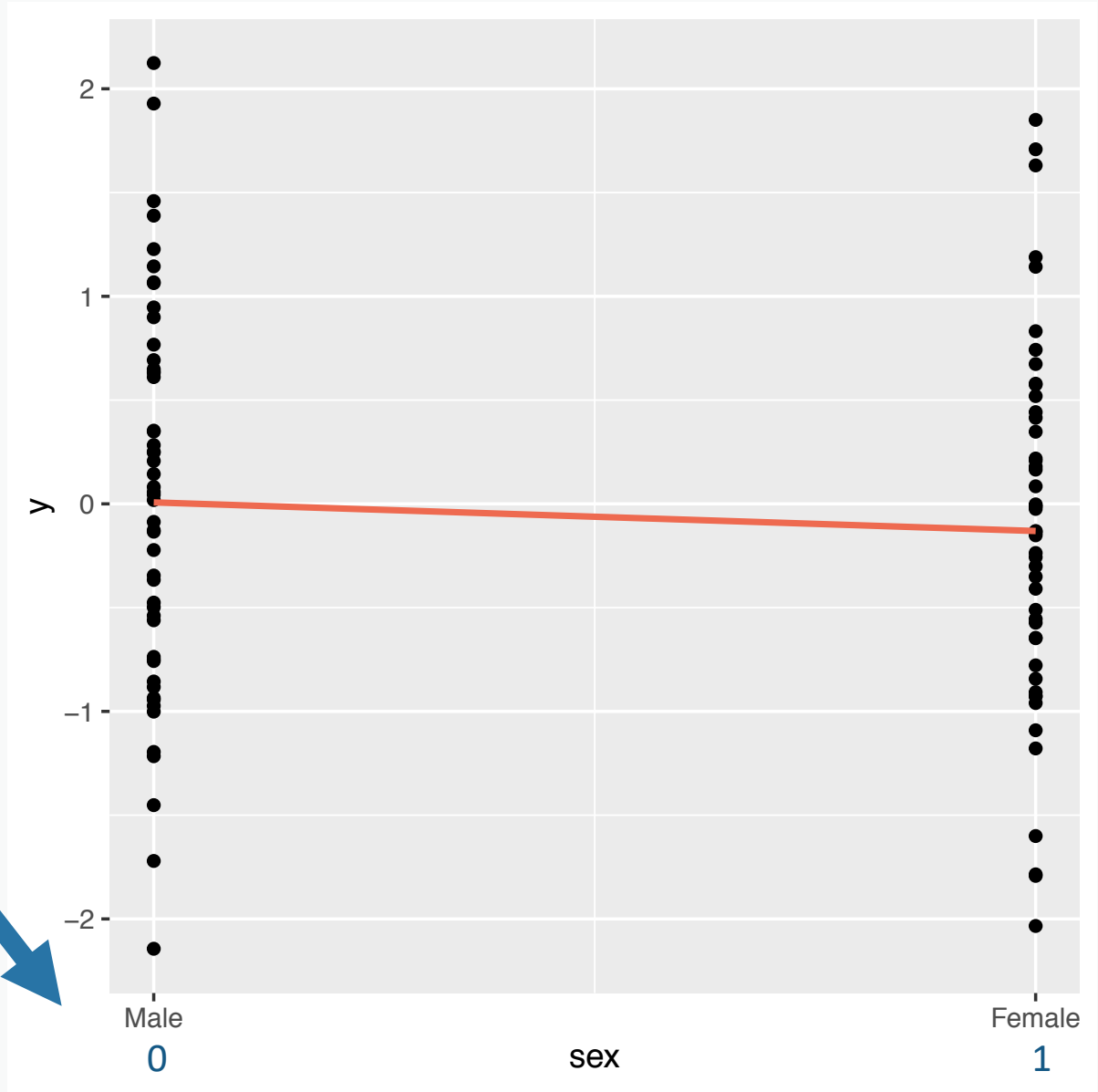
- *ANCOVA (when covariates are included)*
- *t-test (if only a single, two-level variable)*
- *ANOVA (if only a single, 2+ level variable)*



For example, we estimate the difference in height between males and females. Could use a t-test, ANOVA, or ANCOVA. Ultimately, all are from the same “regression” family.

Categorical Predictors

- The line still goes through the conditional mean just like before but now that conditional mean is at a discrete point
- We use **DUMMY CODING** to create an indicator variable (or dummy variable)
  - Use 0's and 1's usually
  - Interpretation is basically the same because we are talking about the effect of a one unit change in the predictor (i.e., a change from male to female in this case)



Interpretation of Dummy Variables

- When we dummy code, there is always a **reference group**
- In R, this is the group that isn't mentioned (in output below, female is mentioned but not male so male is the reference group)
- The estimates of the dummy variable are the mean differences from that group
  - Why might we need to have a reference group?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006949   0.126594   0.055    0.956
sexFemale   -0.137779   0.180849  -0.762    0.448
```

How do we interpret the sex variable estimate here?

Interpretation of Dummy Variables

- Let's look at this from a predicted value perspective:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006949   0.126594   0.055    0.956
sexFemale   -0.137779   0.180849  -0.762    0.448
```

$$\hat{Y} = 0.007 - 0.138 * sex$$

Male

When the person is male (dummy coded as 0) we get:

$$\hat{Y} = 0.007 - 0.138 * 0 = 0.007$$

Female

When the person is female (dummy coded as 1) we get:

$$\hat{Y} = 0.007 - 0.138 * 1 = -0.131$$

Interpretation of Dummy Variables

In other words:

- Males' and Females' means differ by the estimate of -0.138
- The predicted value of males and females differ by the estimate of -0.138
- Females, on average, are 0.138 units of y lower than males

Male

When the person is male (dummy coded as 0) we get:

$$\hat{Y} = 0.007 - 0.138 * 0 = 0.007$$

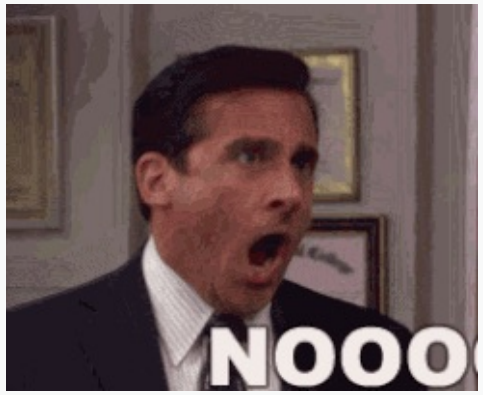
Female

When the person is female (dummy coded as 1) we get:

$$\hat{Y} = 0.007 - 0.138 * 1 = -0.131$$

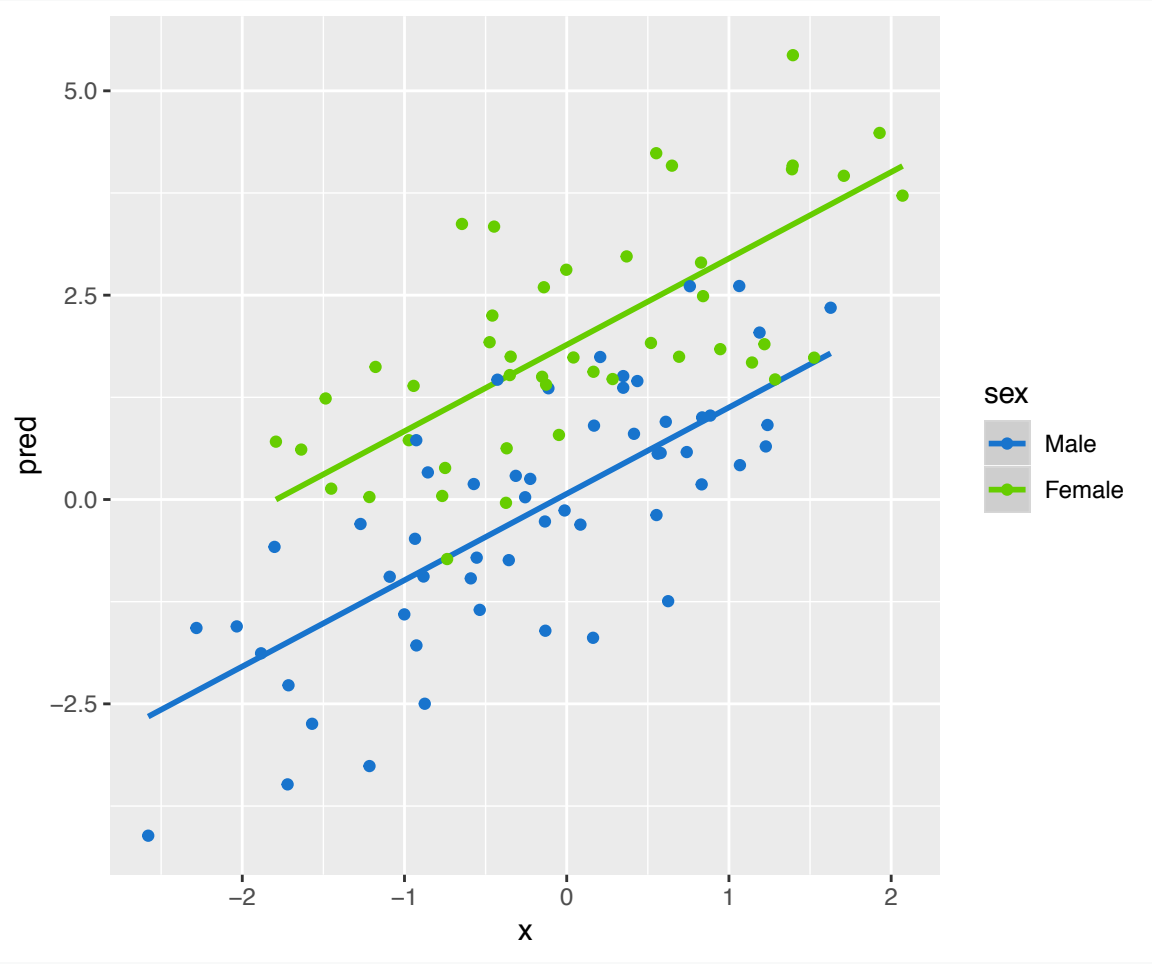
Quick note about dummy coding

- If we have a variable with two levels we have a single dummy coded variable
- We never make two variables
  - e.g., one indicator for females and another for males
  - Due to COLLINEARITY!



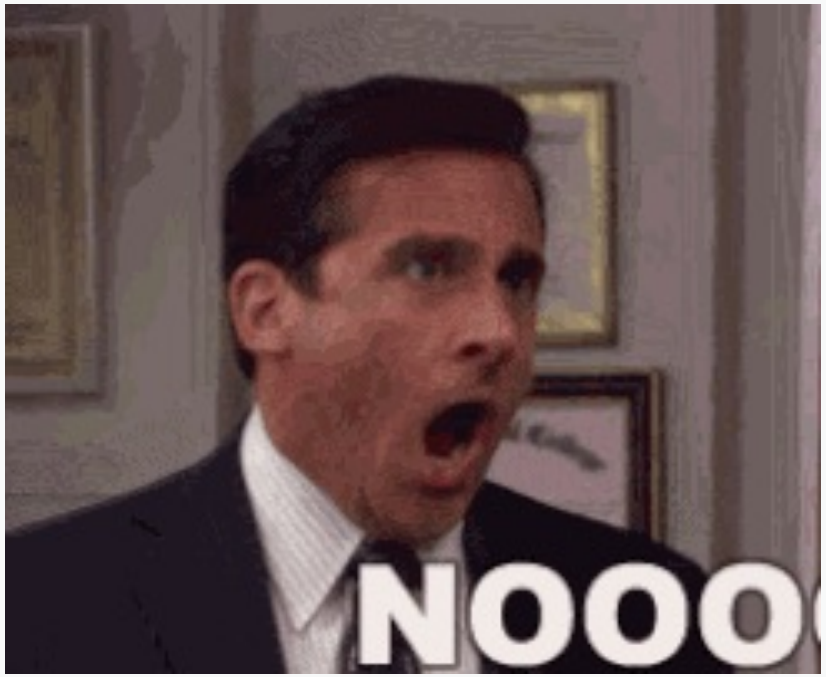
Multiple Regression with Dummy Variables

- Can have dummy variables with other dummy variables and with other continuous variables
- Essentially the dummy variable changes the intercept (i.e., the line is higher for females than males but the slope is the same)
  - When done with 2+ covariates, the line turns into a plane (see page 130)



Don’t Standardize Categorical Predictors

- The book discusses this in depth (pages 130 – 132)
- It just isn’t a great idea because it hurts the interpretability in several ways



Have you ever …?

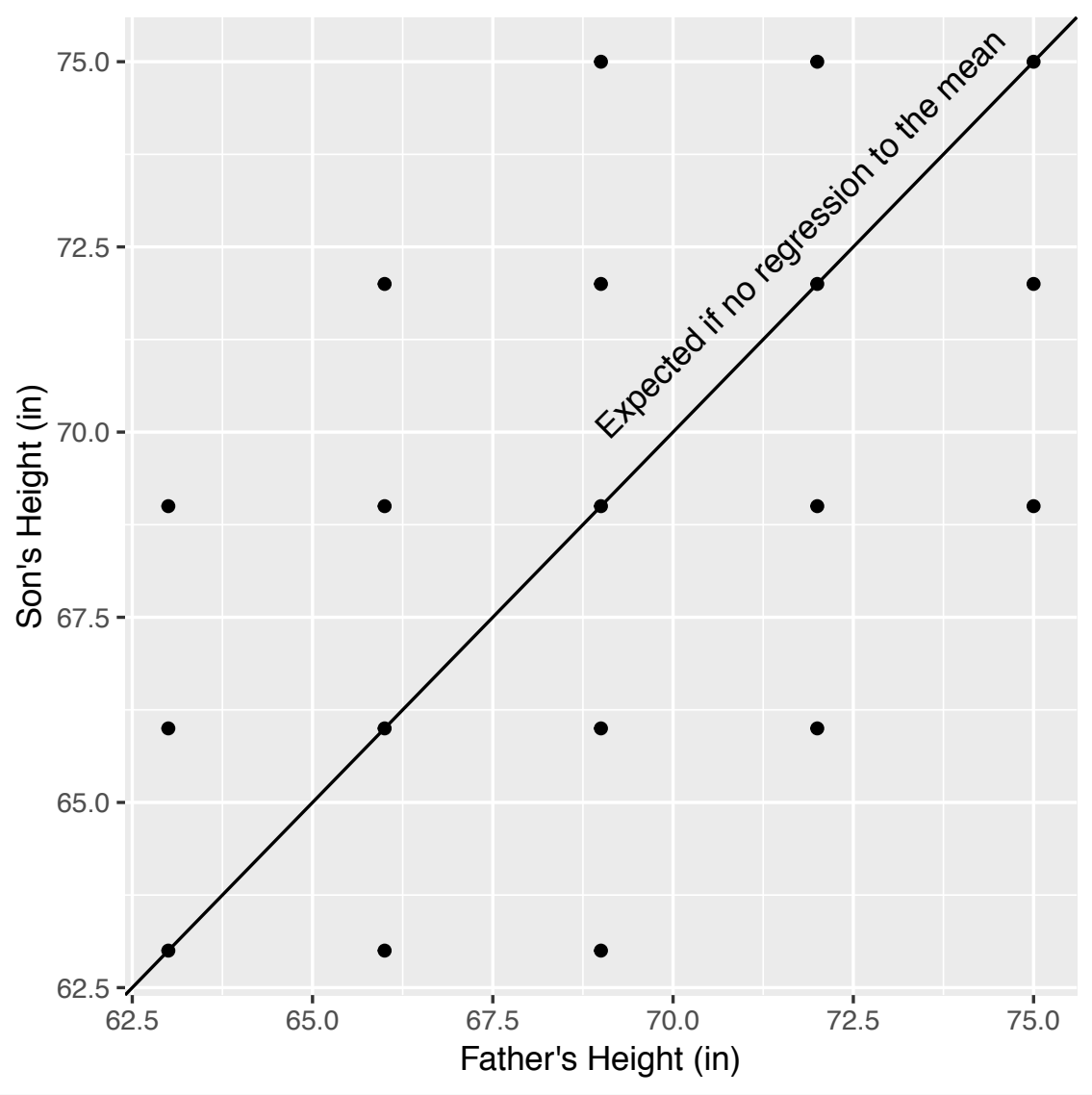
- Tried to make a numeric variable categorical so you could use ANOVA?

1. Less powerful
2. Throws away information
3. Treats similar people as though they were different
4. Increases measurement error

What about clinical depression cut off points?

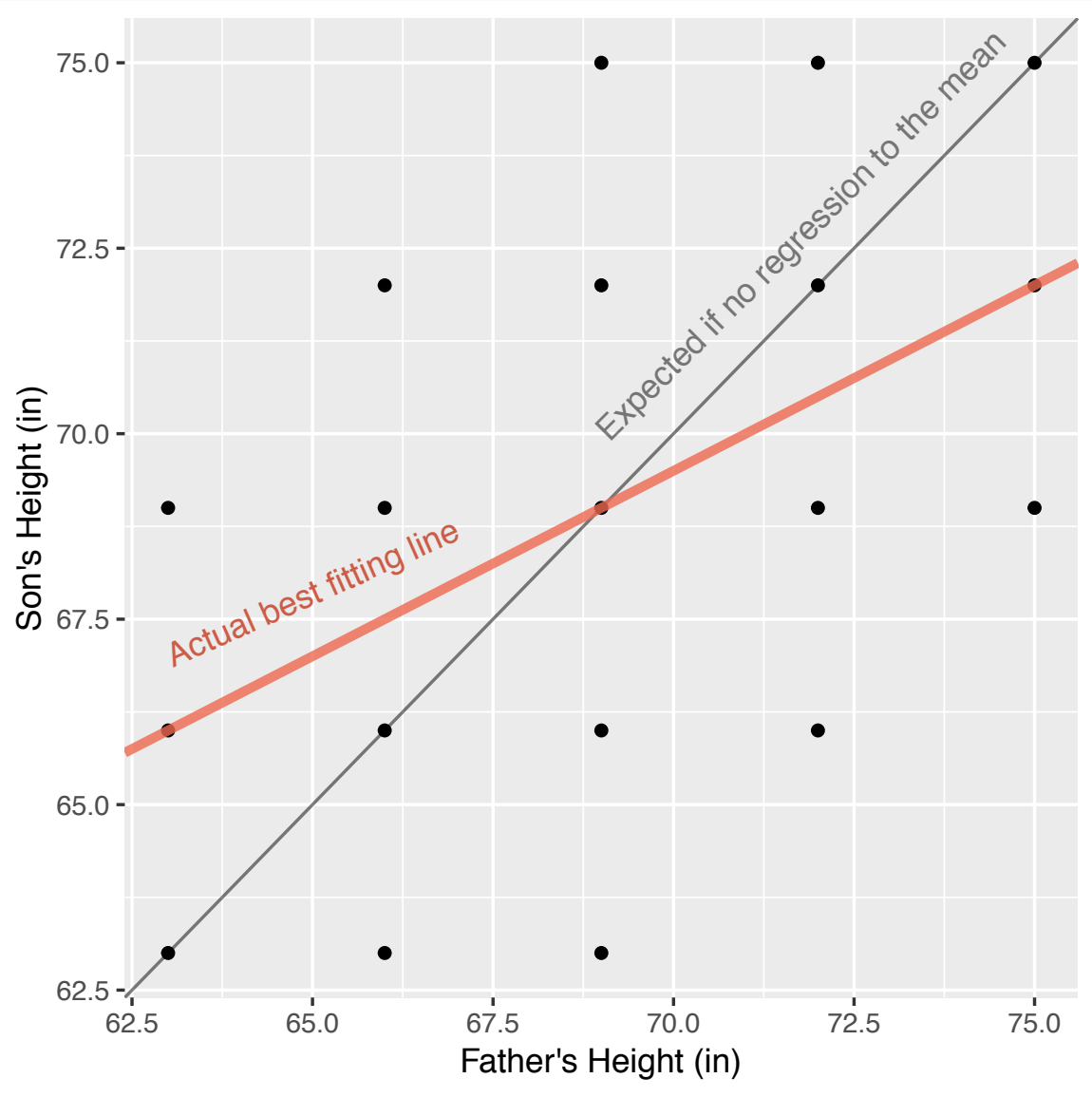
Regression to the Mean

- Regression got its name from this phenomenon



Regression to the Mean

- Regression got its name from this phenomenon
- **Extreme values in one variable are often associated with values closer to the mean in another variable**

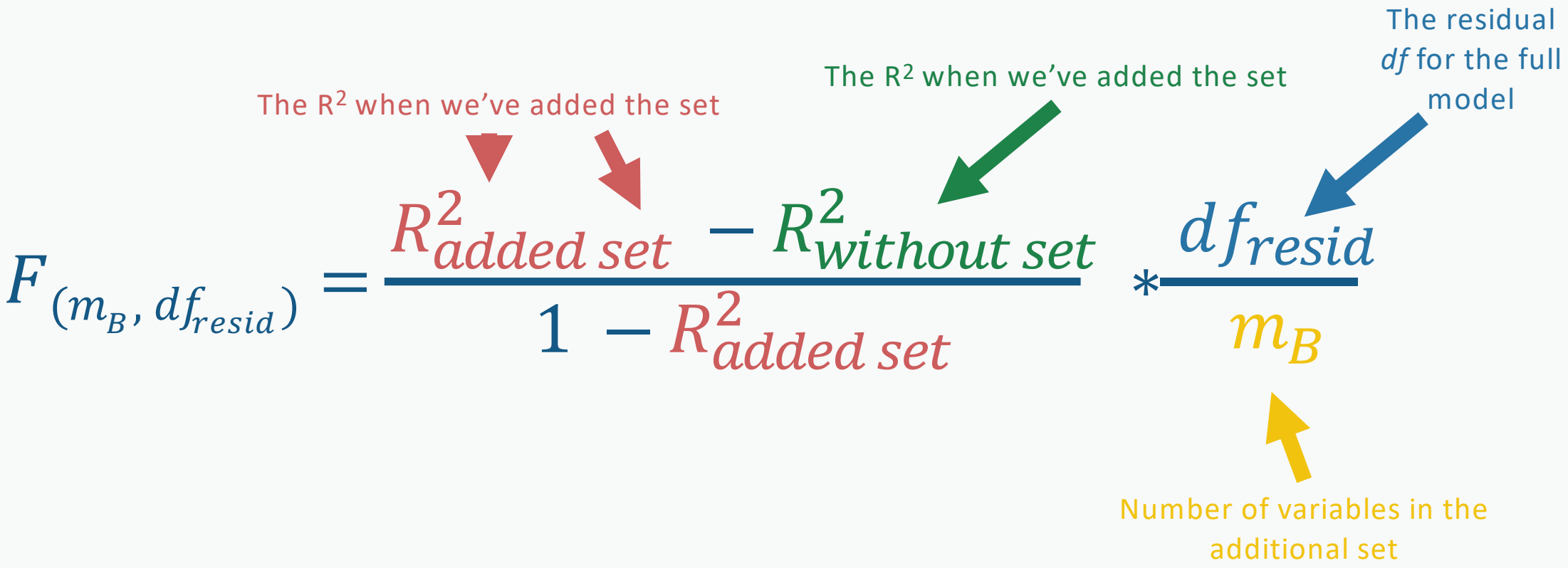


Why talk about Regression to the Mean?

- Implications for difference scores
  - Lesson: need to control for initial score as covariate (pg. 143)
- Regression, when used right, naturally accounts for regression to the mean
- Interpret results where extreme values are associated with more average scores with caution
- Aside: Looking at a difference score (with first score as a covariate) or using the second as dependent and first as a covariate are identical in meaning

Multidimensional Sets

We can infer regarding a set of variables (e.g., SES variables, demographic variables)



Multidimensional Sets

We can infer regarding a set of variables (e.g., SES variables, demographic variables)

This allows us to test:

1. the relationship between *a set of collinear variables and the outcome*
2. many variables with *similar interpretations simultaneously* without having to do them individually

Final Thoughts

A somewhat extensive list of things to be aware of when you use a regression (many of these we'll talk about later in the course)

1. Under-control
2. Over-control
3. Singularity
4. Nonlinearity
5. Interaction (moderation)
6. Heteroscedasticity
7. Non-normality of errors
8. Outliers
9. Leverage points
10. Influential outliers
11. Non-interval scaling
12. Missing data
13. Measurement error