Interactions

EDUC 643: Unit 5 Part I

David D. Liebowitz

Roadmap

Foundations

1. Introduction to regression
- The General Linear Model (GLM)
- Review of bivariate regression
- Coefficient- and model-level inference
- Correlation…and causality

2. Assumptions & Diagnostics
- Measurement error, normality, linearity, homoscedasticity, and independence
- Residuals: raw, studentized & standardized
- Outliers
- Diagnostics and solutions

Adding more and different predictors

3. Multiple regression
- Statistical adjustments ("controls")
- Statistical inference
- Multi-collinearity

4. Categorical predictors
- Two-sample t-tests
- Regression with dummy variables
- ANOVA
- ANCOVA
- Variance decomposition

5. Interactions & Non-linearity
- Interactions in MR models
- Categorical * continuous
- Continuous * continuous
- Transformations to achieve linearity

Putting it together!

6. Applied regression modeling

Goals of the unit

- Describe the main effects assumption and how this assumption can be relaxed using the statistical interaction model
- Describe in writing and verbally the concept of statistical interaction
- Estimate and interpret regression models with interactions between categorical and continuous predictors
- Estimate and interpret regression models with interactions between categorical and continuous predictors
- Estimate and interpret regression models with interactions between continuous predictors
- Visualize interaction effects graphically
- Describe statistical power and Type II error challenges resulting from interactions

- Describe in writing and verbally the assumptions we violate when we fit a non-linear relationship with a linear model
- Transform non-linear relationships into linear ones by using logarithmic scales
- Estimate regression models using logarithmic scales and interpret the results
- Estimate models with quadratic and higher-order polynomial terms (special kinds of interactions)
- Select between transformation options

Our motivating question

A team of researchers based at the University of Oregon aimed to understand the effects of the COVID-19 pandemic on students' early literacy skills.[1]

Ann Swindells Professor in Special Education Gina Biancarosa, former UO doctoral students David Fainstein, Chris Ives, and Dave Furjanic, along with CTL Research Manager Patrick Kennedy, used data from assessments of 471,456 students across 1,684 schools on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) to analyze the extent to which students' Oral Reading Fluency (ORF) scores differed across four waves of DIBELS assessment prior-to and during the pandemic.

Their study is published in *The Elementary School Journal*.

[1] For various reasons, the pandemic is a "lousy natural experiment" for examining the effects of a particular policy response (e.g, virtual schooling). However, it is quite possible to seek to understand its global effects via just the type of analysis Furjanic et al. conducted.

Our answer to-date

	(1)	(2)
(Intercept)	62.344***	16.075***
	(0.449)	(0.326)
Winter 2020	25.562***	25.562***
	(0.696)	(0.332)
Fall 2020	−2.791***	−2.791***
	(0.631)	(0.311)
Winter 2021	19.454***	19.454***
	(0.700)	(0.330)
Covariates?	No	Yes
Num.Obs.	21584	21584
R2	0.102	0.795

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Cells report coefficients and heteroscedastic-robust standard errors in parentheses. Each observation is a school-grade-test value. Covariates include grade-level and total school enrollment.

Might differ by context?

	(1)	(2)
(Intercept)	62.344***	16.075***
	(0.449)	(0.326)
Winter 2020	25.562***	25.562***
	(0.696)	(0.332)
Fall 2020	-2.791***	-2.791***
	(0.631)	(0.311)
Winter 2021	19.454***	19.454***
	(0.700)	(0.330)
Covariates?	No	Yes
Num.Obs.	21584	21584
R2	0.102	0.795

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Cells report coefficients and heteroscedastic-robust standard errors in parentheses. Each observation is a school-grade-test value. Covariates include grade-level and total school enrollment.

Some definitions

Statistical interactions

- When the relationship between one predictor variable and the outcome differs by the level of another predictor
- A statistical term

Moderation effects

- A substantive interpretation of the statistical interaction
- In practice, used interchangeably with interactions

Mediation effects

- A third variable that "explains" why or how one predictor variable is related to the outcome
- Question predictor related to the mediator variable, which in turn is related to the outcome
- Requires some strong assumptions to interpret as "mediation" or "mechanism" that are highly dependent on research design
- We'll return to this in our final unit (*not what we're talking about now*)

Statistical interactions abound

Teacher–child interaction quality moderates social risks of problem behavior[1]

> "We found a negative interaction between early peer context problems and classroom instructional support in the prediction of disconnected play. In classrooms with low instructional quality, children who displayed early problem behavior in peer contexts displayed higher disconnected play in the spring. However, in classrooms with higher instructional support, this association was weakened" (pg. 9)

Low Instructional Support
Median Instructional Support
High Instructional Support
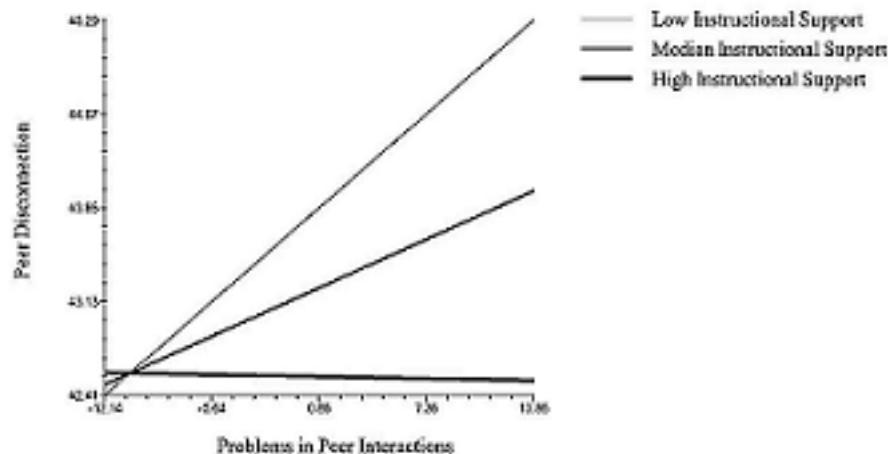
Fig. 3. Instructional support moderates the association between problem behavior in peer contexts and disconnected play.

Note. Problem behavior in peer contexts is centered around the group mean. Disconnected Play is on a standardized T-score metric ($M = 50$, $SD = 10$). Instructional Support is displayed as the 25th percentile (thin line), 50th percentile (median line) and the 75th percentile (thickest line).

[1] Bulotsky–Hearer, R., Fernandez, V., Bichay–Awadalla, K., Bailey, J. Futterer, J. & Qi, C. (2020). Teacher–child interaction quality moderates social risks associated with problem behavior in preschool classroom contexts. *Journal of Applied Developmental Psychology, 67*, 101103.

Main effects model
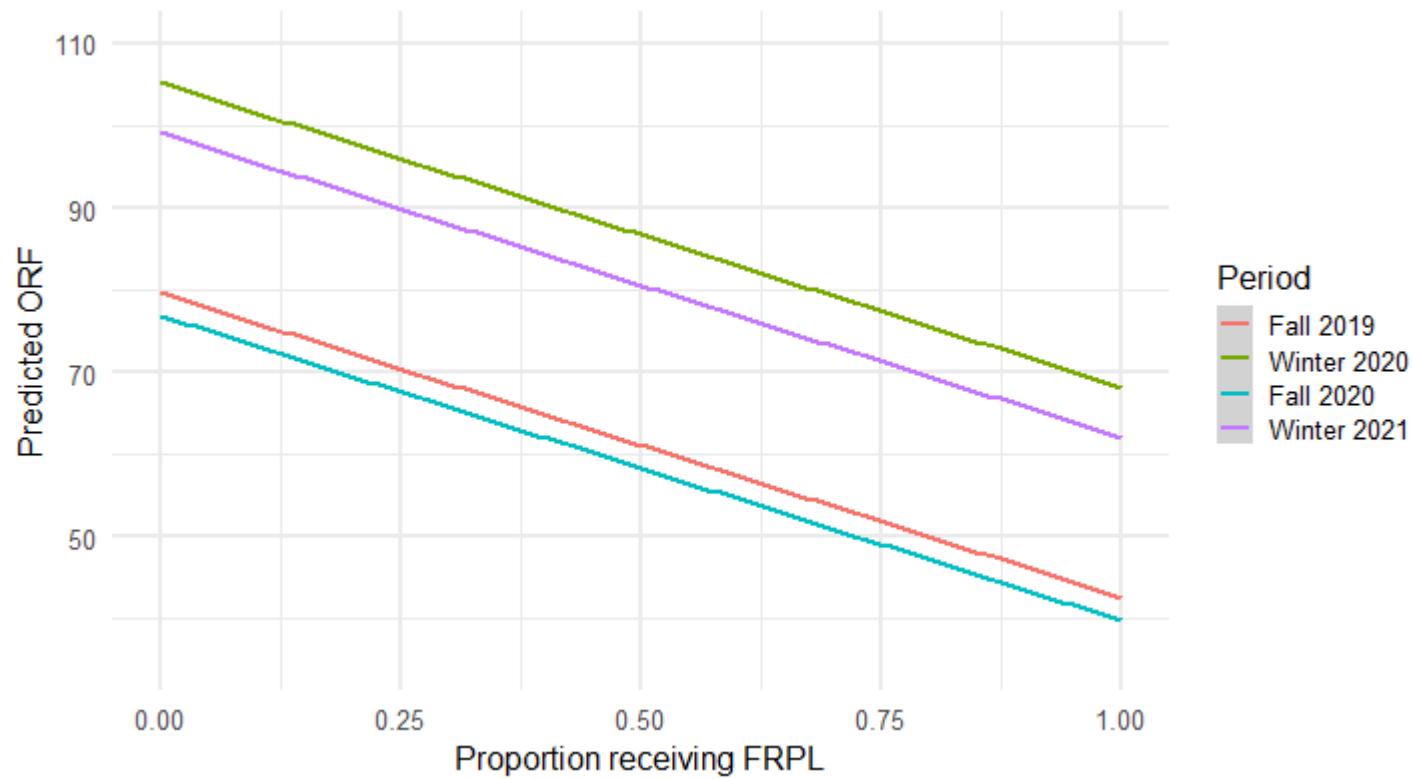
Parallel lines: Fitted lines are parallel because the main effects model assumes that the effect of each predictor is identical and independent of the values of all other other predictors in the model.

Interaction model
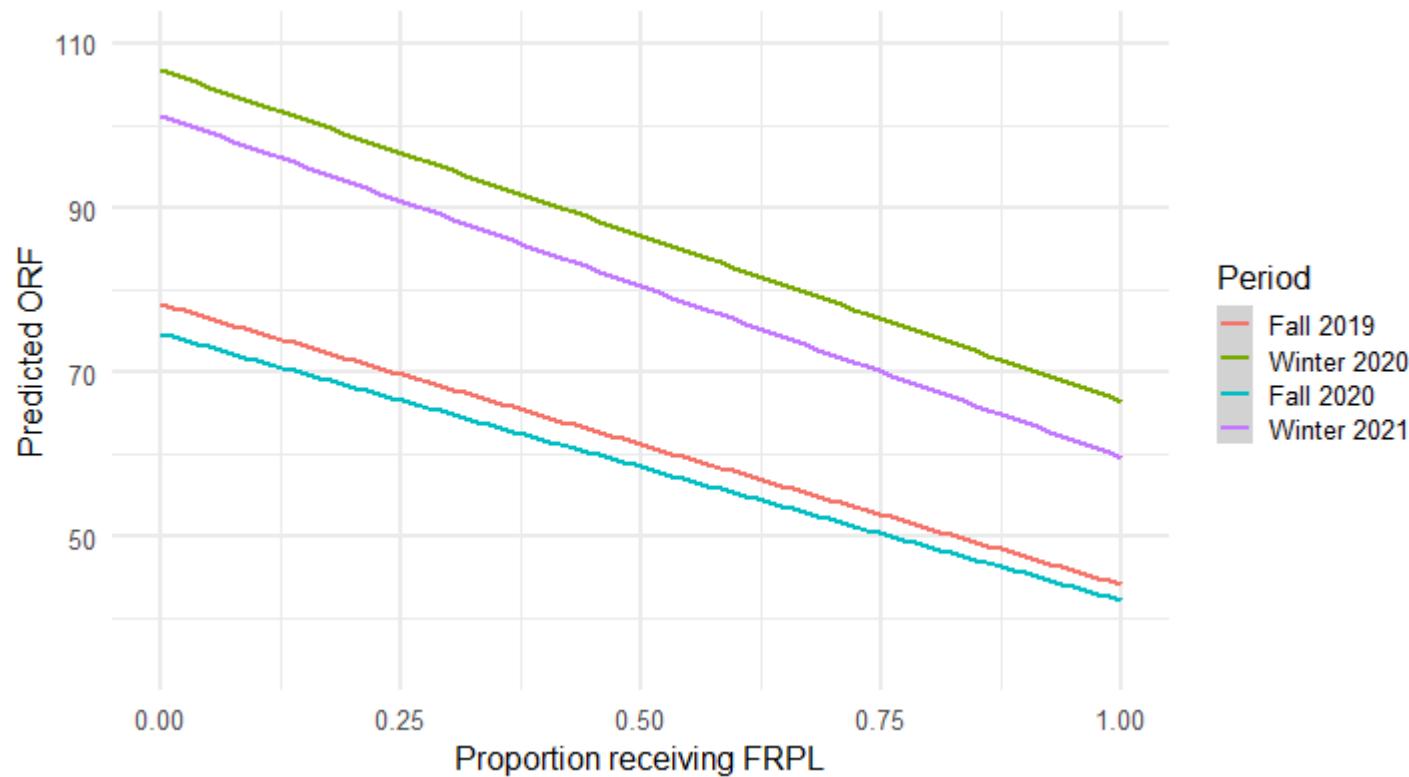
Unparallel lines: Fitted lines are **not** assumed parallel because the interaction model allows the effect of each predictor to differ by the values of other predictor(s) in the model.

Some distinctions

Ordinal interaction

- Direction of predictor's effect consistent across moderator's levels, but **magnitude** differs
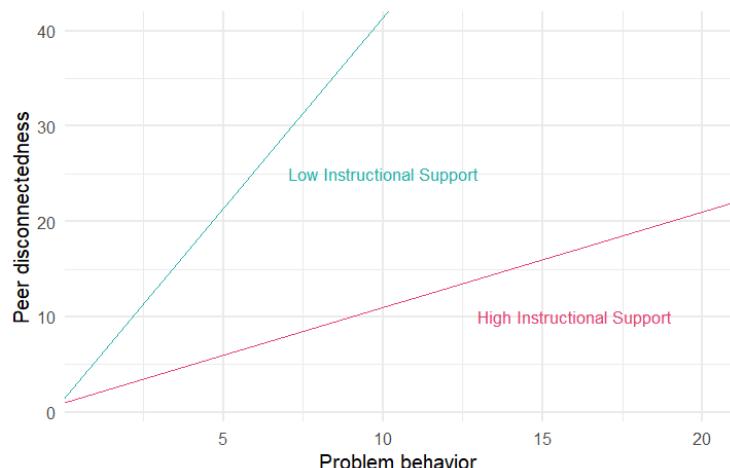
Disordinal interaction

- **Direction** of predictor's relationship with outcome differs across moderator's levels
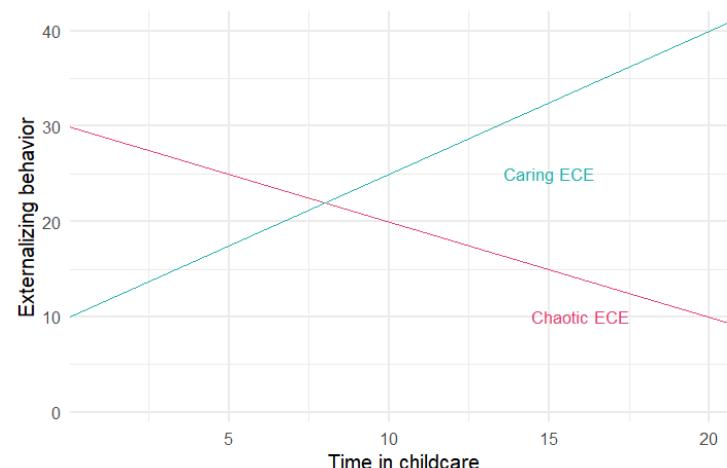
These are somewhat arbitrary distinctions, as all non-parallel lines eventually intersect. However, the key is to constrain your interpretation **within the range of your data**.

Other interactions terms you may encounter: *synergistic interaction*, *buffering interaction*. These are subject- and discipline-specific.

Additional research questions

1. How did students' Oral Reading Fluency (ORF) trajectories differ pre- and post-pandemic-onset?
   - Main effects model
2. To what extent did differences in students' ORF trajectories pre- and post-pandemic-onset differ by the proportion of students receiving free- or reduced-price lunch (FRPL) in their schools?
   - Interaction model: **categorical X continuous**
3. To what extent did differences in students' ORF trajectories pre- and post-pandemic-onset differ by the Title I status of their schools?
   - Interaction model: **categorical x categorical**
4. To what extent do students' average ORF scores differ by the rate of FRPL-receipt and enrollment in their schools?
   - Interaction model: **continuous X continuous**

Interactions: categorical and continuous

Period as predictor

```
summary(lm(mean_orf ~ period, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ period, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.178 -30.719   2.705  28.803 122.202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.3436     0.4921 126.701  < 2e-16 ***
## periody1_moy  25.5620     0.6959  36.734  < 2e-16 ***
## periody2_boy  -2.7914     0.6959  -4.011 6.06e-05 ***
## periody2_moy  19.4541     0.6959  27.957  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.15 on 21580 degrees of freedom
## Multiple R-squared:  0.1021,    Adjusted R-squared:  0.1019
## F-statistic: 817.7 on 3 and 21580 DF,  p-value: < 2.2e-16
```

FRPL_PROP as predictor

```
summary(lm(mean_orf ~ frpl_prop, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ frpl_prop, data = dibels_long)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -87.926 -30.046   1.351  27.450 130.115
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.2822     0.4410  204.70   <2e-16 ***
## frpl_prop   -37.1145     0.7802  -47.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.29 on 21582 degrees of freedom
## Multiple R-squared:  0.09491,    Adjusted R-squared:  0.09486
## F-statistic:  2263 on 1 and 21582 DF,  p-value: < 2.2e-16
```

Both

```
summary(lm(mean_orf ~ period + frpl_prop, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ period + frpl_prop, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.916 -27.204   1.903  27.194 121.217
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.7260     0.5788 137.743  < 2e-16 ***
## periody1_moy  25.5620     0.6581  38.843  < 2e-16 ***
## periody2_boy  -2.7914     0.6581  -4.242 2.23e-05 ***
## periody2_moy  19.4541     0.6581  29.562  < 2e-16 ***
## frpl_prop    -37.1145     0.7349 -50.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.18 on 21579 degrees of freedom
## Multiple R-squared:  0.197,    Adjusted R-squared:  0.1968
## F-statistic: 1323 on 4 and 21579 DF,  p-value: < 2.2e-16
```

Testing for an interaction

1. **Create a cross-product term**, which is the product (the interaction) of the two predictors whose interaction you want to test

- We can create this term by hand or we can ask R to do this for us within our estimating equation

2. **Include the cross-product in a multiple-regression model** *alongside the constituent main effects* of the variables that make up the cross-product term

$$MEAN\_ORF_j = \beta_0 + \beta_1 FRPL_j + \beta_2 PERIOD_j + \textcolor{red}{\beta_3 FRPL \times PERIOD_j} + \varepsilon_j$$

- R will automatically do this if we specify our interaction term appropriately
- For now, assume you should **always** include the main effects (e.g., $FRPL_j$ and $PERIOD_j$). Way down the line, there are cases where it might make sense to not do this.

3. Test H_0: $\beta_{\text{cross-product}} = 0$

What will parameters mean?

$$MEAN\_ORF_j = \beta_0 + \beta_1 FRPL_j + \beta_2 PERIOD_j + \beta_3 FRPL \times PERIOD_j + \varepsilon_j$$

β_2 tells us the difference in the **intercept** when FRPL=0. I.e., how did ORF scores differ at different time periods for schools with no students receiving FRPL.

β_3 tells us the difference in the **slope**. I.e., how did ORF scores differ for different levels of FRPL–receipt in a school, at different time periods across the different waves of the assessments

Estimating in R

```
summary(lm(mean_orf ~ period * frpl_prop, dibels_long))
```

```
...
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            78.3362     0.8304  94.331  < 2e-16 ***
## periody1_moy           28.4229     1.1744  24.202  < 2e-16 ***
## periody2_boy           -3.5915     1.1744  -3.058 0.002230 **
## periody2_moy           22.9528     1.1744  19.544  < 2e-16 ***
## frpl_prop             -34.1469     1.4690 -23.245  < 2e-16 ***
## periody1_moy:frpl_prop  -6.1084     2.0775  -2.940 0.003282 **
## periody2_boy:frpl_prop   1.7084     2.0775   0.822 0.410903
## periody2_moy:frpl_prop  -7.4703     2.0775  -3.596 0.000324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.16 on 21576 degrees of freedom
## Multiple R-squared:  0.198,    Adjusted R-squared:  0.1978
## F-statistic: 761.1 on 7 and 21576 DF,  p-value: < 2.2e-16
...
```

Estimating in R

```
summary(lm(mean_orf ~ period * frpl_prop, dibels_long))
```

```
...
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             78.3362     0.8304  94.331  < 2e-16 ***
## periody1_moy            28.4229     1.1744  24.202  < 2e-16 ***
## periody2_boy            -3.5915     1.1744  -3.058 0.002230 **
## periody2_moy            22.9528     1.1744  19.544  < 2e-16 ***
## frpl_prop              -34.1469     1.4690 -23.245  < 2e-16 ***
## periody1_moy:frpl_prop  -6.1084     2.0775  -2.940 0.003282 **
## periody2_boy:frpl_prop   1.7084     2.0775   0.822 0.410903
## periody2_moy:frpl_prop  -7.4703     2.0775  -3.596 0.000324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.16 on 21576 degrees of freedom
## Multiple R-squared:  0.198,    Adjusted R-squared:  0.1978
## F-statistic: 761.1 on 7 and 21576 DF,  p-value: < 2.2e-16
...
```

Estimating in R

```
summary(lm(mean_orf ~ period * frpl_prop, dibels_long))
```

```
...
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               78.3362     0.8304  94.331  < 2e-16 ***
## periody1_moy              28.4229     1.1744  24.202  < 2e-16 ***
## periody2_boy              -3.5915     1.1744  -3.058 0.002230 **
## periody2_moy              22.9528     1.1744  19.544  < 2e-16 ***
## frpl_prop                -34.1469     1.4690 -23.245  < 2e-16 ***
## periody1_moy:frpl_prop    -6.1084     2.0775  -2.940 0.003282 **
## periody2_boy:frpl_prop     1.7084     2.0775   0.822 0.410903
## periody2_moy:frpl_prop    -7.4703     2.0775  -3.596 0.000324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.16 on 21576 degrees of freedom
## Multiple R-squared:  0.198,    Adjusted R-squared:  0.1978
## F-statistic: 761.1 on 7 and 21576 DF,  p-value: < 2.2e-16
...
```

Estimating in R

```
summary(lm(mean_orf ~ period * frpl_prop, dibels_long))
```

```
...
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             78.3362     0.8304  94.331  < 2e-16 ***
## periody1_moy            28.4229     1.1744  24.202  < 2e-16 ***
## periody2_boy            -3.5915     1.1744  -3.058 0.002230 **
## periody2_moy            22.9528     1.1744  19.544  < 2e-16 ***
## frpl_prop              -34.1469     1.4690 -23.245  < 2e-16 ***
## periody1_moy:frpl_prop  -6.1084     2.0775  -2.940 0.003282 **
## periody2_boy:frpl_prop   1.7084     2.0775   0.822 0.410903
## periody2_moy:frpl_prop  -7.4703     2.0775  -3.596 0.000324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.16 on 21576 degrees of freedom
## Multiple R-squared:  0.198,    Adjusted R-squared:  0.1978
## F-statistic: 761.1 on 7 and 21576 DF,  p-value: < 2.2e-16
...
```

Do not interpret main effects or interactions by themselves!

For Fall 2019:

$$
\begin{aligned}
MEA\hat{N}\_ORF_j =\ &78.3 + (-34.1) * FRPL + 28.4(0) + (-3.6) * (0) + 23.0 * (0) +\\
&(-6.1) * FRPL \times (0) + 1.7 * FRPL \times (0) + (-7.5) * FRPL \times (0)\\
=\ &78.3 - 34.1 * FRPL
\end{aligned}
$$

For Winter 2020:

$$
\begin{aligned}
MEA\hat{N}\_ORF_j =\ &78.3 + (-34.1) * FRPL + 28.4(1) + (-3.6) * (0) + 23.0 * (0) +\\
&(-6.1) * FRPL \times (1) + 1.7 * FRPL \times (0) + (-7.5) * FRPL \times (0)\\
=\ &106.7 - 40.2 * FRPL
\end{aligned}
$$

For Fall 2020:

$$
\begin{aligned}
MEA\hat{N}\_ORF_j =\ &78.3 + (-34.1) * FRPL + 28.4(0) + (-3.6) * (1) + 23.0 * (0) +\\
&(-6.1) * FRPL \times (0) + 1.7 * FRPL \times (1) + (-7.5) * FRPL \times (0)\\
=\ &74.7 - 32.4 * FRPL
\end{aligned}
$$

For Winter 2021:

$$
\begin{aligned}
MEA\hat{N}\_ORF_j =\ &78.3 + (-34.1) * FRPL + 28.4(0) + (-3.6) * (0) + 23.0 * (1) +\\
&(-6.1) * FRPL \times (0) + 1.7 * FRPL \times (0) + (-7.5) * FRPL \times (1)\\
=\ &101.3 - 41.6 * FRPL
\end{aligned}
$$

Until we do the math, we only know whether interactions are statistically different, but not by how much!

Show results

```r
fit5 <- lm(mean_orf ~ period * frpl_prop, data=dibels_long)
df5 <- margins::margins(fit5,
            at = list(period=c("y1_boy", "y1_moy",
                               "y2_boy", "y2_moy")))

# Use prototypical values in resulting dataset to show results
proto2 <- ggplot(data=df5, aes(x=frpl_prop, y=fitted, color=period)) +
          geom_smooth(method='lm') +
          xlab("Proportion receiving FRPL") + ylab("Predicted ORF") +
          ylim(35, 110) +
          scale_color_discrete(name = "Period",
                               breaks=c("y1_boy", "y1_moy",
                                        "y2_boy", "y2_moy"),
                               labels=c("Fall 2019","Winter 2020",
                                        "Fall 2020", "Winter 2021")) +
          theme_minimal(base_size=16)
```

Show results
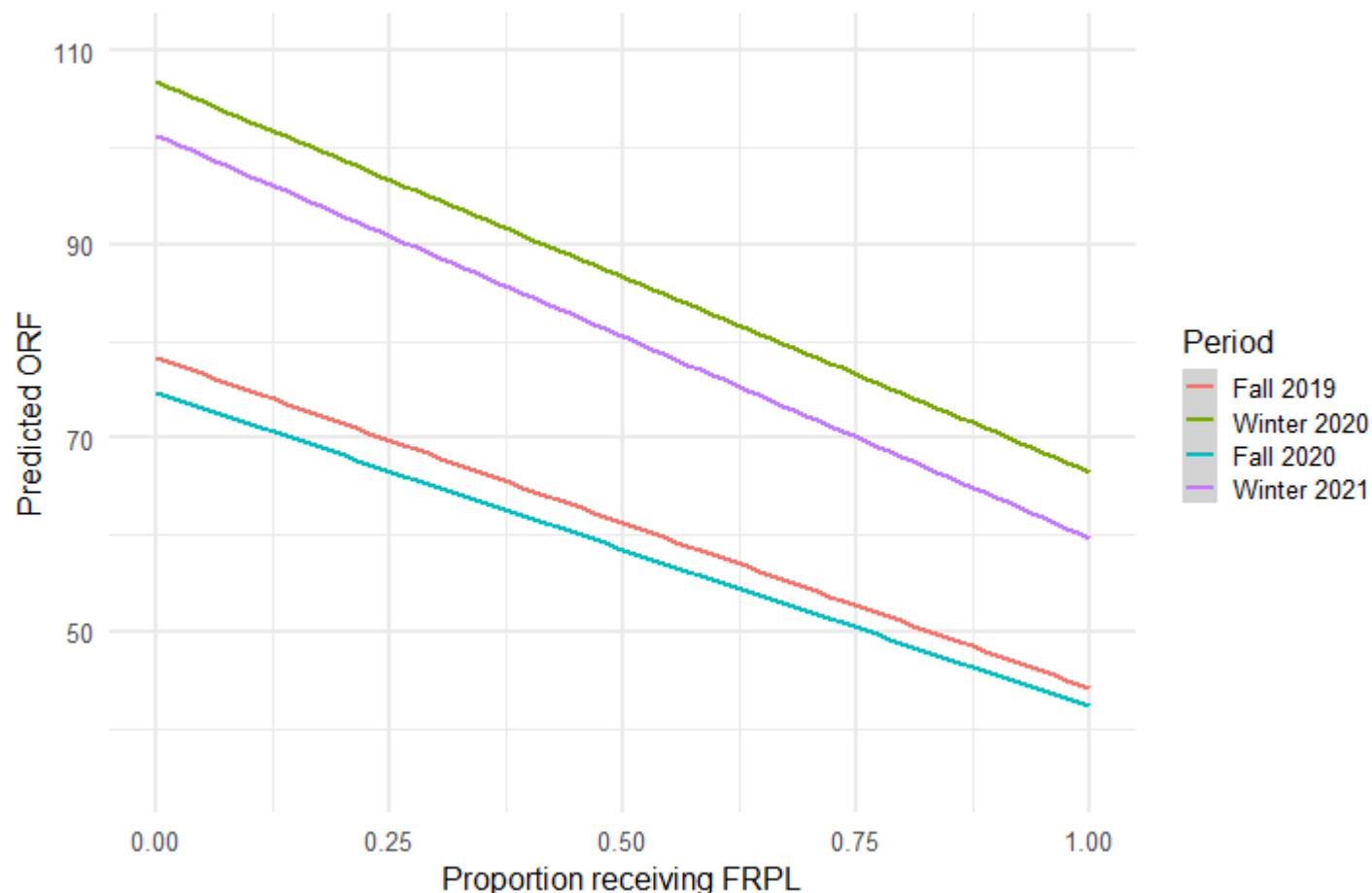
Another way?

```
## 
## Call:
## lm(formula = mean_orf ~ frpl_prop, data = subset(dibels_long,
##     period == "y1_boy"))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.154 -25.613  -0.189  25.572  80.831
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.3362     0.7582  103.31   <2e-16 ***
## frpl_prop   -34.1469     1.3413  -25.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 31.19 on 5394 degrees of freedom
## Multiple R-squared:  0.1073,    Adjusted R-squared:  0.1071
## F-statistic: 648.1 on 1 and 5394 DF,  p-value: < 2.2e-16
```

Another way?

```
## 
## Call:
## lm(formula = mean_orf ~ frpl_prop, data = subset(dibels_long,
##     period == "y2_moy"))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.378 -31.799   3.262  29.915 121.097
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.2890     0.9037  112.08   <2e-16 ***
## frpl_prop   -41.6172     1.5986  -26.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 37.18 on 5394 degrees of freedom
## Multiple R-squared:  0.1116,    Adjusted R-squared:  0.1115
## F-statistic: 677.7 on 1 and 5394 DF,  p-value: < 2.2e-16
```

Value (and perils) of interaction

1. As we'll see momentarily, interaction models can be fit whether one (or more) of the predictors are dichotomous or continuous
2. Interaction models provide an easy statistical test of whether the slopes differ across groups (or across levels of a continuous predictor)
3. Interaction models keep the sample intact (you don't need to break it down into many different groups)
   - Has some implementation advantages, but the purported statistical power advantages have been historically misunderstood and overstated
   - For more on this, check out Andrew Gelman's, "You need 16 times the sample size to estimate an interaction than to estimate a main effect"
4. Can be quite hard to clearly interpret
   - Visualization critical to communicate

Interactions: categorical and categorical

What are our categories?

```
table(dibels_long$title1, exclude=NULL)
```

```
##
##         Not Title I Title I schoolwide   Title I targeted                 Missing
##                2752               15956               2124                     752
```

```
dibels_long %>% group_by(title1) %>% summarize(mean= mean(mean_orf))
```

```
## # A tibble: 4 x 2
##   title1               mean
##   <fct>               <dbl>
## 1 Not Title I          89.0
## 2 Title I schoolwide   70.3
## 3 Title I targeted     78.6
## 4 Missing              52.9
```

Title I as predictor

```
summary(lm(mean_orf ~ title1, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ title1, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.215 -32.520   2.159  28.532 133.689
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              89.0149     0.7132 124.814   <2e-16 ***
## title1Title I schoolwide -18.7034     0.7722 -24.220   <2e-16 ***
## title1Title I targeted   -10.4537     1.0806  -9.674   <2e-16 ***
## title1Missing            -36.1616     1.5395 -23.490   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.41 on 21580 degrees of freedom
## Multiple R-squared:  0.03796,    Adjusted R-squared:  0.03783
## F-statistic: 283.8 on 3 and 21580 DF,  p-value: < 2.2e-16
```

Both

```
summary(lm(mean_orf ~ period + title1, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ period + title1, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.169 -30.793   3.011  28.610 124.791
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              78.4587     0.7929  98.954  < 2e-16 ***
## periody1_moy             25.5620     0.6810  37.533  < 2e-16 ***
## periody2_boy             -2.7914     0.6810  -4.099 4.17e-05 ***
## periody2_moy             19.4541     0.6810  28.565  < 2e-16 ***
## title1Title I schoolwide -18.7034     0.7302 -25.615  < 2e-16 ***
## title1Title I targeted   -10.4537     1.0217 -10.231  < 2e-16 ***
## title1Missing            -36.1616     1.4556 -24.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.38 on 21577 degrees of freedom
```

Interaction

```
summary(lm(mean_orf ~ period * title1, dibels_long))
```

```
...
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                 77.6809     1.3487  57.598  < 2e-16 *
## periody1_moy                                27.0794     1.9073  14.198  < 2e-16 *
## periody2_boy                                -3.6222     1.9073  -1.899   0.0576 .
## periody2_moy                                21.8790     1.9073  11.471  < 2e-16 *
## title1Title I schoolwide                   -17.8299     1.4604 -12.209  < 2e-16 *
## title1Title I targeted                     -10.3759     2.0434  -5.078 3.85e-07 *
## title1Missing                              -32.5905     2.9113 -11.195  < 2e-16 *
## periody1_moy:title1Title I schoolwide       -1.7557     2.0653  -0.850   0.3953
## periody2_boy:title1Title I schoolwide        1.0411     2.0653   0.504   0.6142
## periody2_moy:title1Title I schoolwide       -2.7795     2.0653  -1.346   0.1784
## periody1_moy:title1Title I targeted         -0.6469     2.8899  -0.224   0.8229
## periody2_boy:title1Title I targeted          0.8777     2.8899   0.304   0.7613
## periody2_moy:title1Title I targeted         -0.5420     2.8899  -0.188   0.8512
## periody1_moy:title1Missing                  -4.4702     4.1171  -1.086   0.2776
## periody2_boy:title1Missing                  -0.7234     4.1171  -0.176   0.8605
## periody2_moy:title1Missing                  -9.0909     4.1171  -2.208   0.0272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

You try!

Construct an equation of the form:

$$
\begin{aligned}
\hat{MEAN\_ORF}_j = & \beta_0 + \beta_1 P_2 + \beta_2 P_3 + \beta_3 P_4 + \\
& \beta_4 T_2 + \beta_5 T_3 + \beta_6 T_4 + \\
& \beta_7 P_2 \times T_2 + \beta_8 P_2 \times T_3 + \beta_9 P_2 \times T_4 + \\
& \beta_{10} P_3 \times T_2 + \beta_{11} P_3 \times T_3 + \beta_{12} P_3 \times T_4 + \\
& \beta_{13} P_4 \times T_2 + \beta_{14} P_4 \times T_3 + \beta_{15} P_4 \times T_4
\end{aligned}
$$

with the coefficients from the previous slide and determine the fitted equation comparing Title I schoolwide schools in Winter 2021 to Title I schoolwide schools in Fall 2019.

Visualization

```r
# Let's just focus on ORF changes in first fall of pandemic

fit6 <- lm(mean_orf ~ period * title1, dibels_long)
df6 <- margins::margins(fit6,
        at = list(period=c("y2_boy", "y2_moy"),
        title1 = c("Not Title I", "Title I schoolwide",
                    "Title I targeted", "Missing")))

# Show results for each category
categ <- ggplot(data=df6,
            aes(x=period, y=fitted,
                ymin=fitted-1.96*se.fitted, ymax=fitted+1.96*se.fitted,
                group=title1, color=title1)) +
        geom_pointrange(position=position_dodge(width=0.2)) +
        ylab("Predicted ORF") + xlab(" ") +
        scale_x_discrete(labels= c("y2_boy" = "Fall 2020",
                                    "y2_moy" = "Winter 2021")) +
        ylim(0, 110) +
        theme_minimal(base_size=16) +
        theme(legend.title = element_blank())
```

Visualization
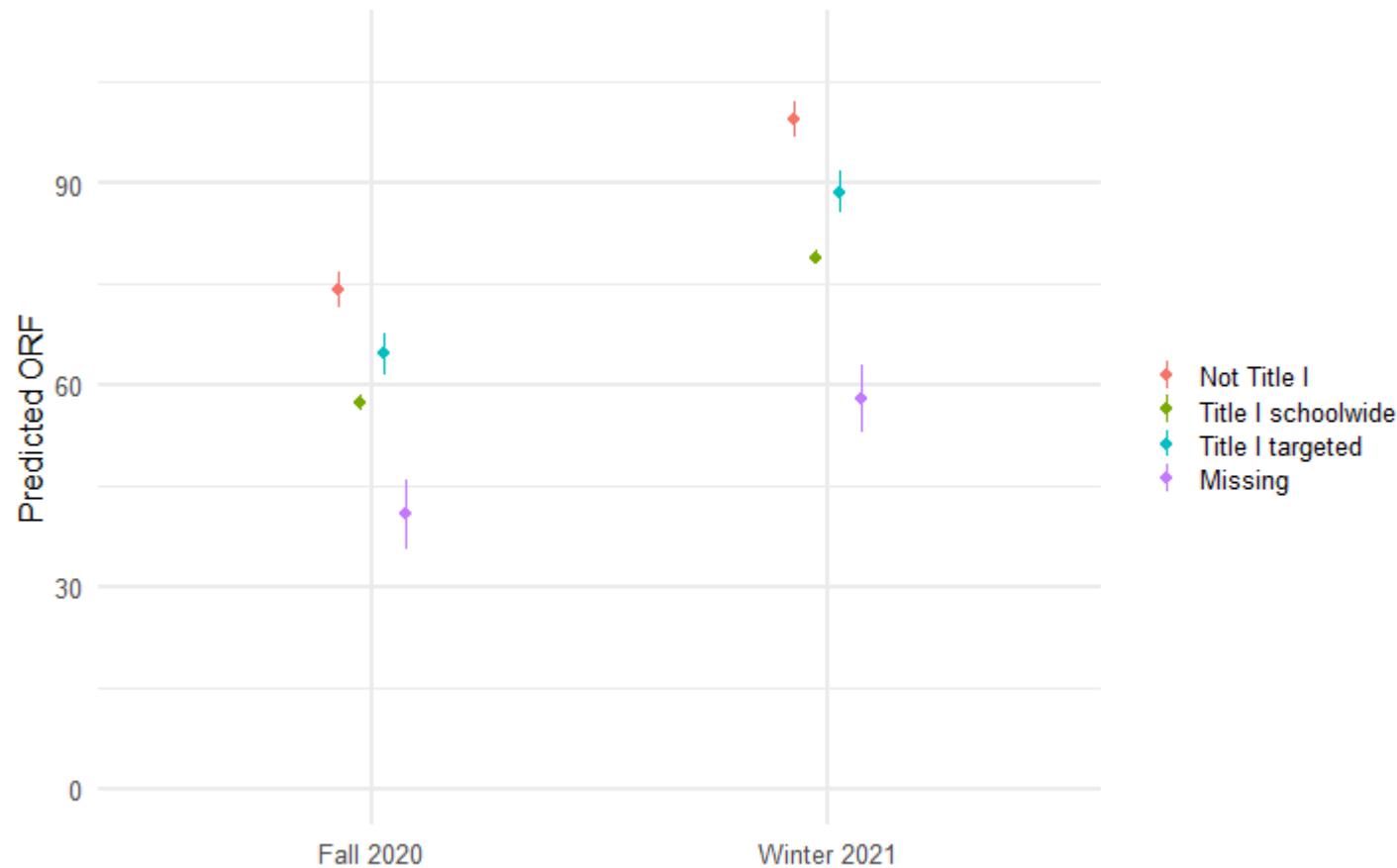
Visualization

For simplicity just focusing on primary question predictor interacted with one other variable (moderator). Could include multiple covariates and/or three-way interactions. Warning! It gets complicated (and underpowered) fast!

Interactions: continuous and continuous

School enrollment as predictor

```
summary(lm(mean_orf ~ school_enroll, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ school_enroll, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.247 -31.962   1.117  28.587 136.732
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.073893   0.568179  105.73   <2e-16 ***
## school_enroll  0.038062   0.001505   25.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.59 on 21582 degrees of freedom
## Multiple R-squared:  0.02876,    Adjusted R-squared:  0.02872
## F-statistic: 639.2 on 1 and 21582 DF,  p-value: < 2.2e-16
```

Both

```
summary(lm(mean_orf ~ frpl_prop + school_enroll, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ frpl_prop + school_enroll, data = dibels_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.355 -28.850   1.075  26.915 134.743
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    79.010988   0.684100  115.50   <2e-16 ***
## frpl_prop     -35.319032   0.776608  -45.48   <2e-16 ***
## school_enroll   0.030953   0.001447   21.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.91 on 21581 degrees of freedom
## Multiple R-squared:  0.1137,    Adjusted R-squared:  0.1136
## F-statistic:  1384 on 2 and 21581 DF,  p-value: < 2.2e-16
```

Interaction

```
summary(lm(mean_orf ~ frpl_prop * school_enroll, dibels_long))
```

```
##
## Call:
## lm(formula = mean_orf ~ frpl_prop * school_enroll, data = dibels_long)
##
## Residuals:
##      Min      1Q  Median      3Q      Max
## -92.138 -28.539   1.124  26.924 134.563
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             81.125811   1.031551  78.644  < 2e-16 ***
## frpl_prop              -39.378046   1.673133 -23.536  < 2e-16 ***
## school_enroll            0.024601   0.002733   9.000  < 2e-16 ***
## frpl_prop:school_enroll  0.012675   0.004628   2.739  0.00617 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.9 on 21580 degrees of freedom
## Multiple R-squared:  0.114,    Adjusted R-squared:  0.1139
## F-statistic: 925.7 on 3 and 21580 DF,  p-value: < 2.2e-16
```

Prototypical values?

How could we choose meaningful values to demonstrates the differing relationship?

```
quantile(dibels_long$school_enroll, probs = seq(0, 1, 0.1))
```

```
##    0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    25  120  182  233  283  326  370  414  472  564 1014
```

Maybe at roughly the 10th, median (50th), and 90th percentiles? Say 120, 326 and 600 students?

Displaying continuous interactions

```r
fit7 <- lm(mean_orf ~ frpl_prop * school_enroll, dibels_long)

df7 <- margins::margins(fit7,
          at = list(school_enroll=c(120, 326, 600)))

# Use prototypical values in resulting dataset to show results
cont <- ggplot(data=df7, aes(x=frpl_prop, y=fitted,
                             color=as.factor(school_enroll))) +
        geom_smooth(method='lm') +
        xlab("Proportion receiving FRPL") + ylab("Predicted ORF") +
        ylim(35, 100) +
        scale_color_discrete(name = "School Enrollment",
                             breaks=c(120, 326, 600),
                             label=c("~10th pctile (120 stu.)",
                                     "Median (326 stu.)",
                                     "~90th pctile (600 stu.)")) +
        theme_minimal(base_size=16)
```

Displaying continuous interactions
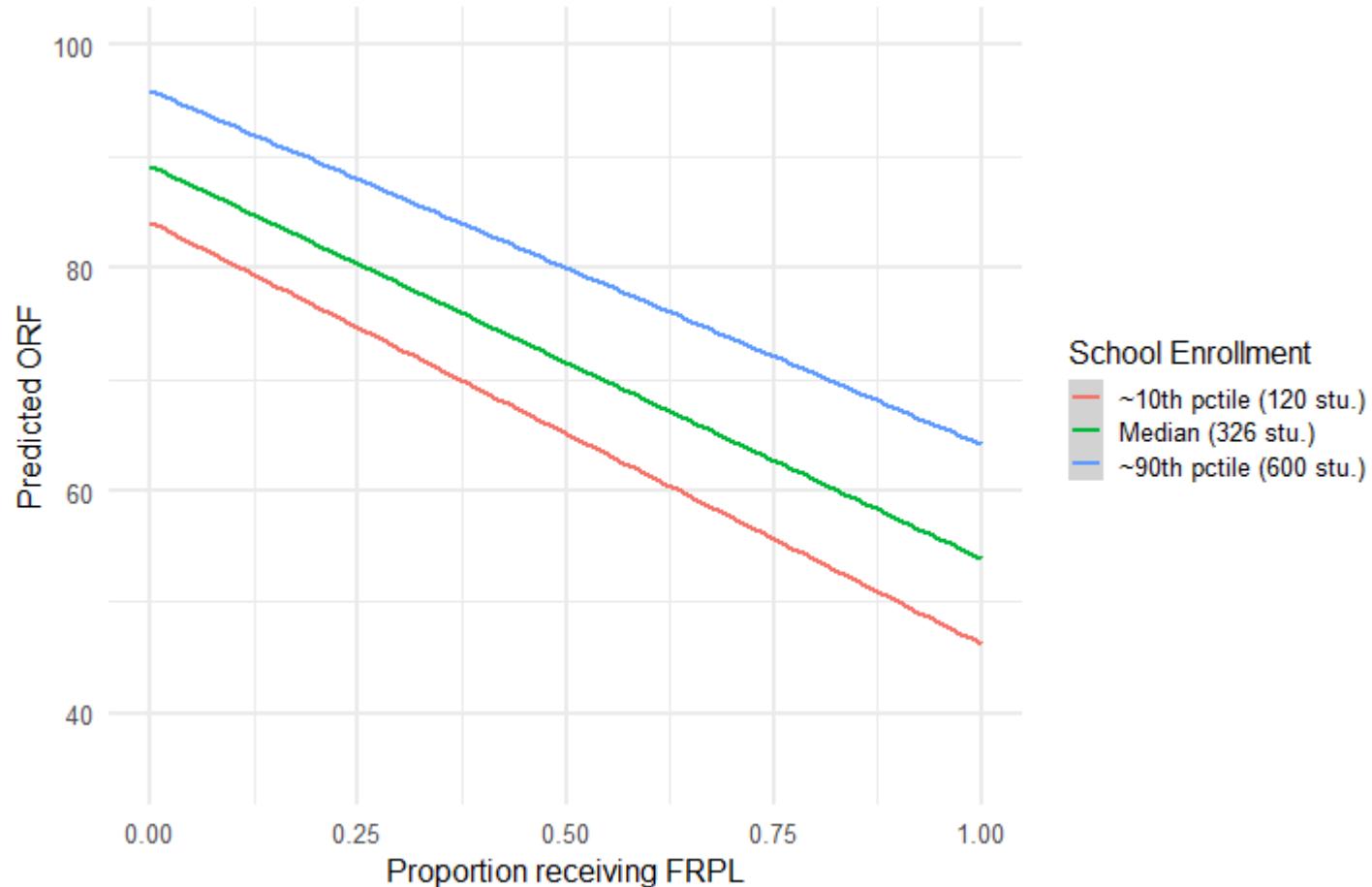
Can also add a confidence band

```
cont +
geom_ribbon(aes(ymin=fitted-1.96*se.fitted, ymax=fitted+1.96*se.fitted,
fill=as.factor(school_enroll)), alpha=0.3, linetype=0, show.legend = F)
```
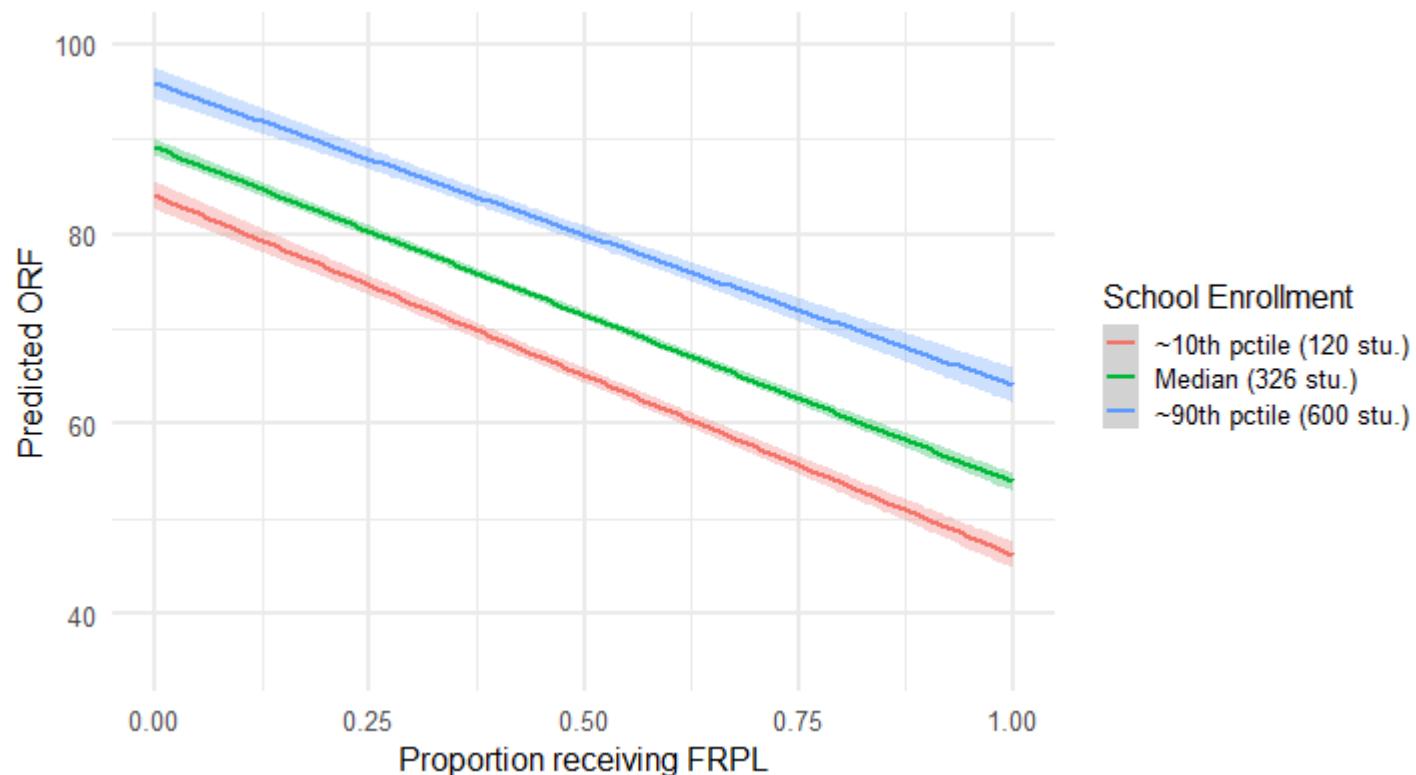
Synthesis and wrap-up

Synthesize interactions

- **Statistical interactions are ubiquitous**
    - An interaction tells us that the relationship between one predictor and the outcome differs by levels of another
    - The standard regression model, which initially assumes that there are no interactions (the main effects assumption), can be easily modified to accommodate their presence
    - Many substantive theories suggest that relationships will be interactive
- **Test for a statistical interaction by including a cross-product term**
    - The cross-product is literally the product of the two constituent variables
    - In incorporating an interaction (moderation) term, be careful about:
        - Removing the main effects
        - Statistical power
    - Graph out the fitted model to ensure correct interpretation
        - Explore the `sjPlot` package for more ways to visualize interactions
- **More learning (beyond this course)**
    - Centering and/or standardizing variables can aid with interpretation
    - The `contrast` and `emtrends` functions in the `emmeans` package can quickly test for whether fitted values at different levels of predictor and moderator are significantly different
- **Predictors can interact with themselves** (next sub-unit: non-linearities)

Goals of the unit

- Describe the main effects assumption and how this assumption can be relaxed using the statistical interaction model
- Describe in writing and verbally the concept of statistical interaction
- Estimate and interpret regression models with interactions between categorical and continuous predictors
- Estimate and interpret regression models with interactions between categorical and continuous predictors
- Estimate and interpret regression models with interactions between continuous predictors
- Visualize interaction effects graphically
- Describe statistical power and Type II error challenges resulting from interactions

- Describe in writing and verbally the assumptions we violate when we fit a non-linear relationship with a linear model
- Transform non-linear relationships into linear ones by using logarithmic scales
- Estimate regression models using logarithmic scales and interpret the results
- Estimate models with quadratic and higher-order polynomial terms (special kinds of interactions)
- Select between transformation options

To-Dos

Reading:

- **Finish by Feb. 27**: LSWR Chapter 16.2

Assignment 3:

- Due Feb. 28, 12:01pm (noon)

Assignment 4 (last one!!!):

- Just on interactions (not non-linearity)
- Due Mar. 10, 12:01pm (noon)

Quiz 4:

- **NOW!!** Due 5pm tomorrow, Feb. 26