Lesson 11: Interactions, Part 1

Nicky Wakim

2025-02-12

Learning Objectives

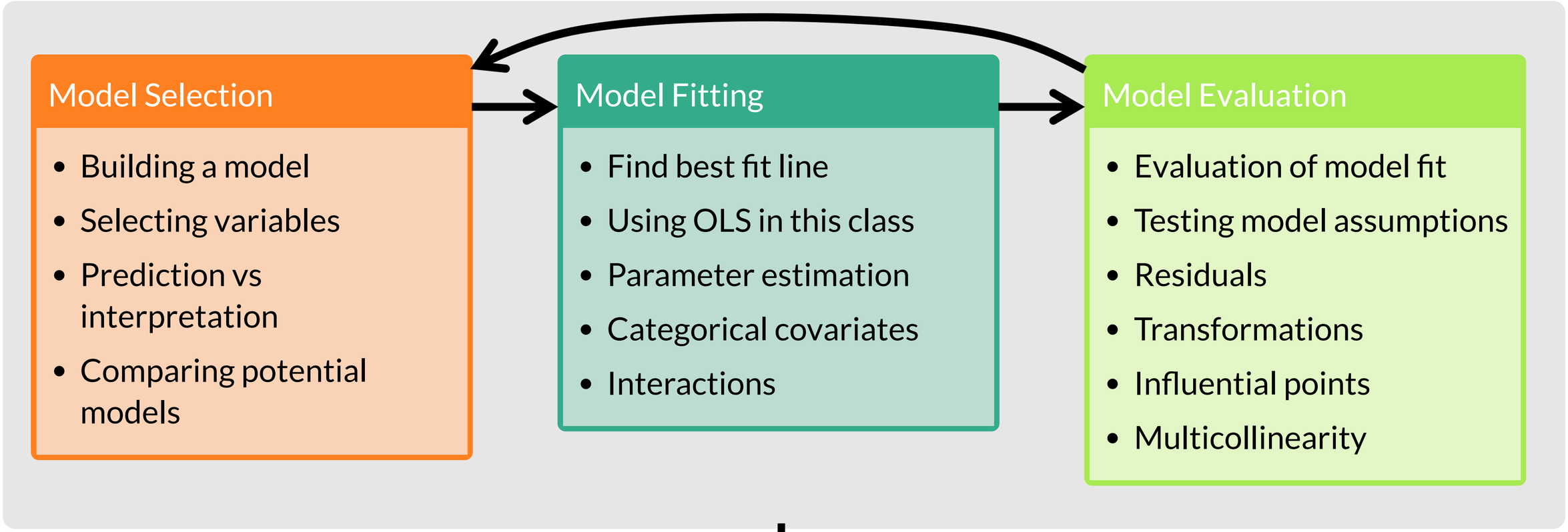
This time:

1. Define confounders and effect modifiers, and how they interact with the main relationship we model.
2. Interpret the interaction component of a model with **a binary categorical covariate and continuous covariate**, and how the main variable's effect changes.
3. Interpret the interaction component of a model with **a multi-level categorical covariate and continuous covariate**, and how the main variable's effect changes.

Next time:

4. Interpret the interaction component of a model with **two categorical covariates**, and how the main variable's effect changes.
5. Interpret the interaction component of a model with **two continuous covariates**, and how the main variable's effect changes.

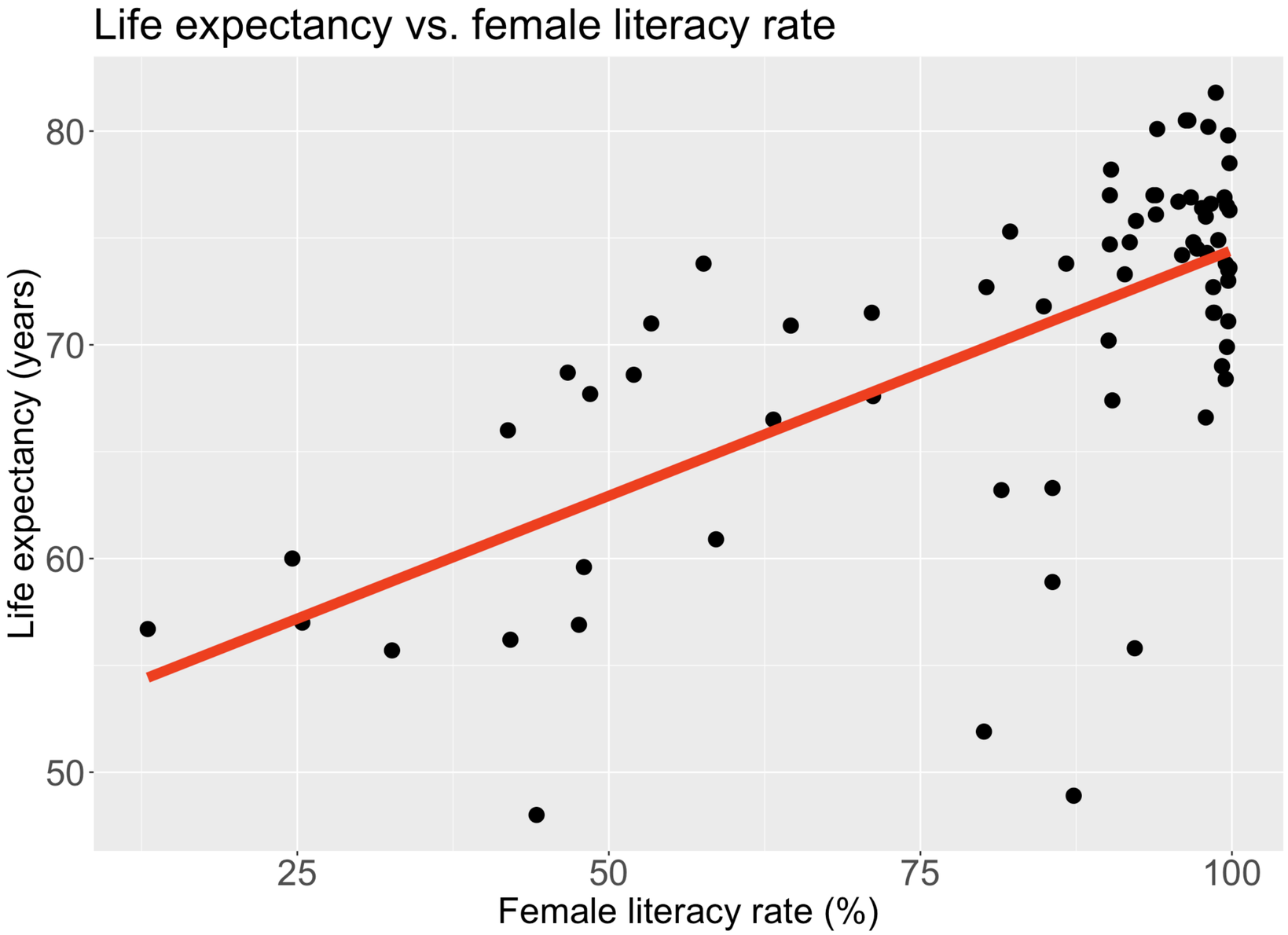
Regression analysis process



Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Recall our data and the main relationship



Learning Objectives

This time:

1. Define confounders and effect modifiers, and how they interact with the main relationship we model.
2. Interpret the interaction component of a model with **a binary categorical covariate and continuous covariate**, and how the main variable's effect changes.
3. Interpret the interaction component of a model with **a multi-level categorical covariate and continuous covariate**, and how the main variable's effect changes.

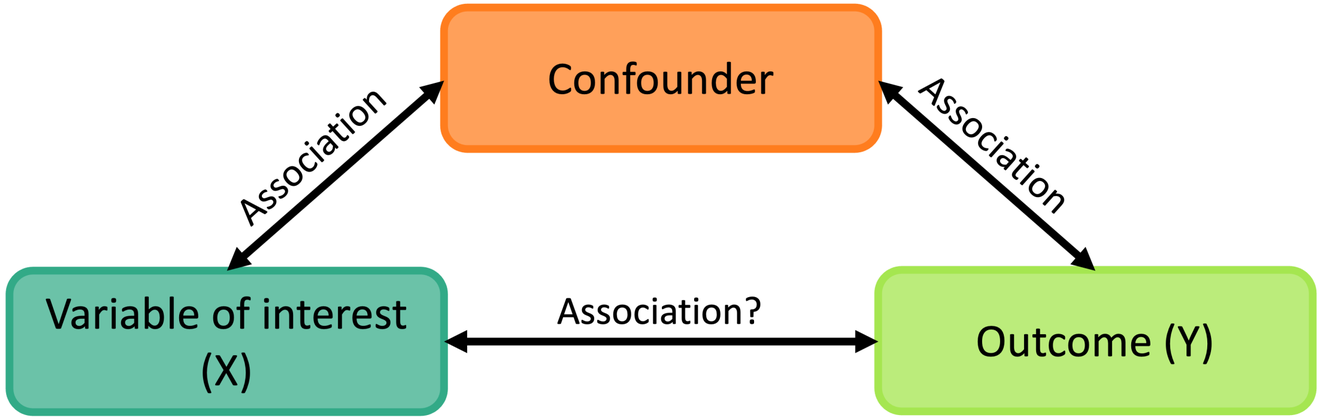
Next time:

4. Interpret the interaction component of a model with **two categorical covariates**, and how the main variable's effect changes.
5. Interpret the interaction component of a model with **two continuous covariates**, and how the main variable's effect changes.

What is a confounder?

- A **confounding variable**, or **confounder**, is a factor/variable that wholly or partially *accounts for the observed effect of the risk factor on the outcome*

- A confounder must be…
  - Related to the outcome Y, but not a consequence of Y
  - Related to the explanatory variable X, but not a consequence of X



- A classic example: We found an association between ice cream consumption and sunburn!
  - If we adjust for a potential confounder, temperature/hot weather, we may see that the association between ice and sunburn is not as large
- Another example: We found an association between socioeconomic status (SES) and lung cancer!
  - If we adjust for a potential confounder, exposure to air pollution, we may see that the association between SES and lung cancer decreases

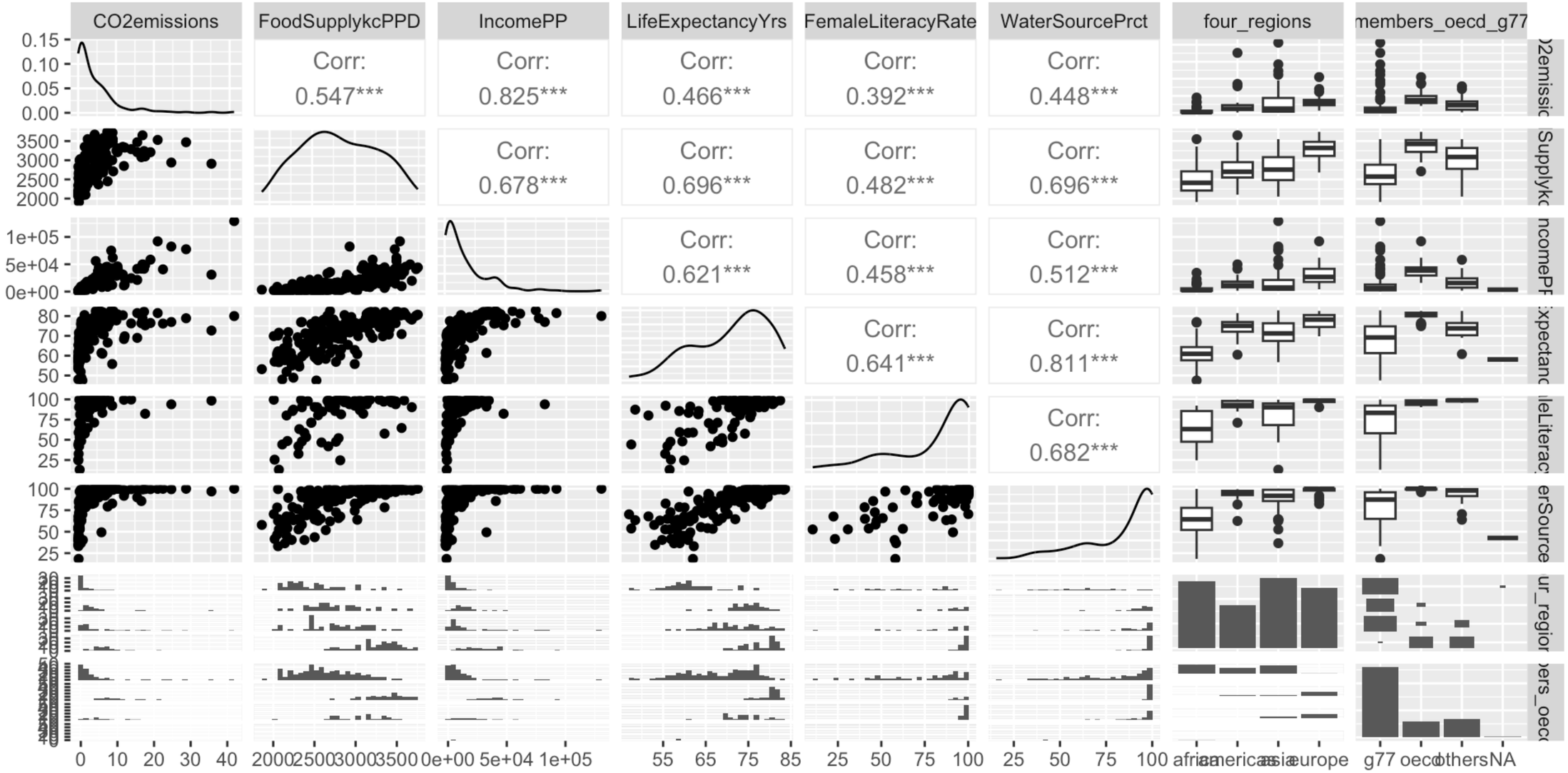
Proxies and confounders: the good and the harmful

- *This is totally my own tangent*
- A **proxy variable** is used to stand-in or represent another variable that is harder to measure
- Sometimes a confounder can be used as a proxy if it is hard to measure you explanatory variable/variable of interest
- Proxies can be helpful statistically while harmful socially OR helpful for both!

- Examples
  - Bad: BMI serving as a measurement for physical health or diet
    - Many studies show how harmful, mentally and physically, it is to equate BMI to health
  - Interesting: Using occurrence of online search queries as a proxy for public health risk perception
  - Helpful contextualization: Using race as a proxy for systemic racism, and thus a way to identify how to and who needs resources
- In our lab, I discuss using sex assigned at birth in our model

Exploratory approach to identifying confounders

```
1  gapm2 %>% ggpairs()
```



Including a confounder in the model

- In the following model we have two variables, X_1 and X_2

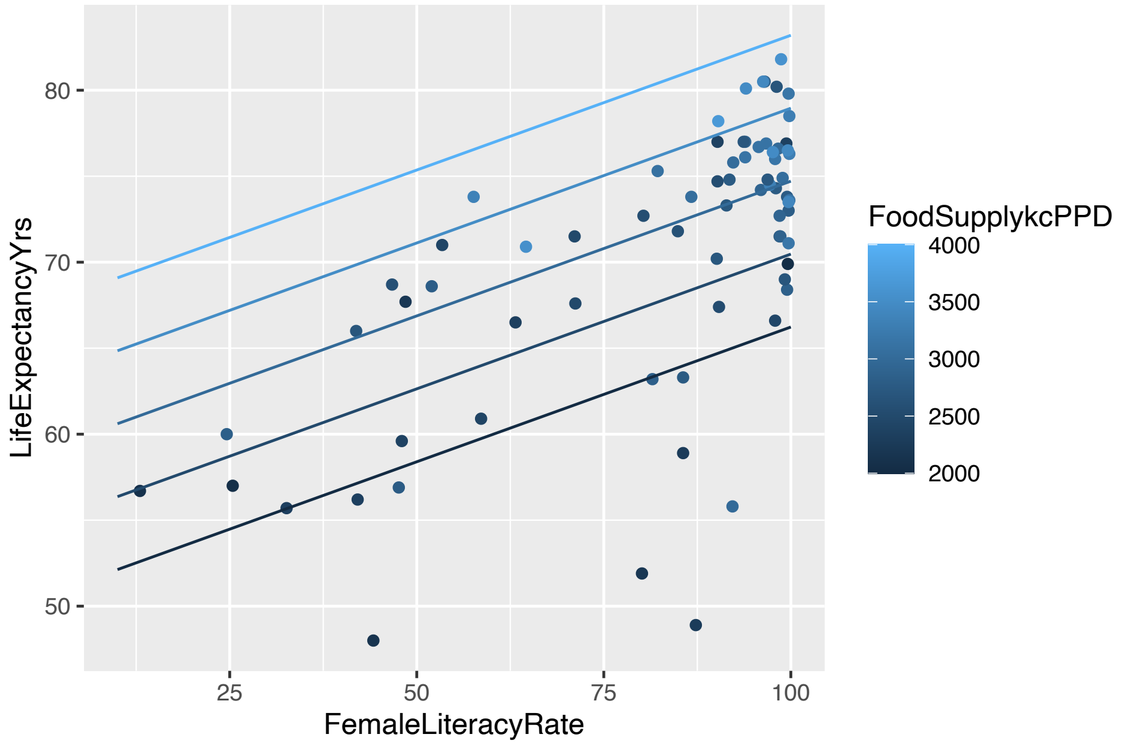
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- And we assume that every level of the confounder, there is parallel slopes
- Note: to interpret β_1, we did not specify any value of X_2; only specified that it be held constant
  - Implicit assumption: effect of X_1 is equal across all values of X_2
- The above model assumes that X_1 and X_2 do not *interact* (with respect to their effect on Y)
  - Epidemiology: no "effect modification"
  - Meaning the effect of X_1 is the same regardless of the values of X_2
  - This model is often called a **"main effects model"**

Where have we modeled a confounder before?

- We have seen a plot of Life expectancy vs. female literacy rate with different levels of food supply colored (Lesson 8)
- In our plot and the model, we treat food supply as a **confounder**
- If food supply is a confounder in the relationship between life expectancy and female literacy rate, then we only use main effects in the model:

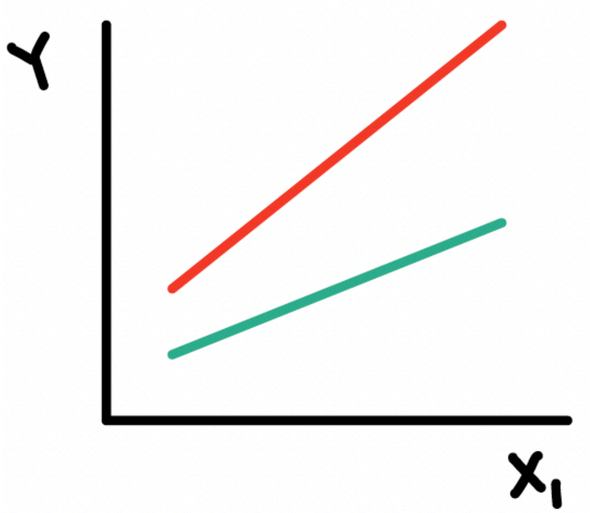
$$\mathrm{LE} = \beta_0 + \beta_1 \mathrm{FLR} + \beta_2 \mathrm{FS} + \epsilon$$

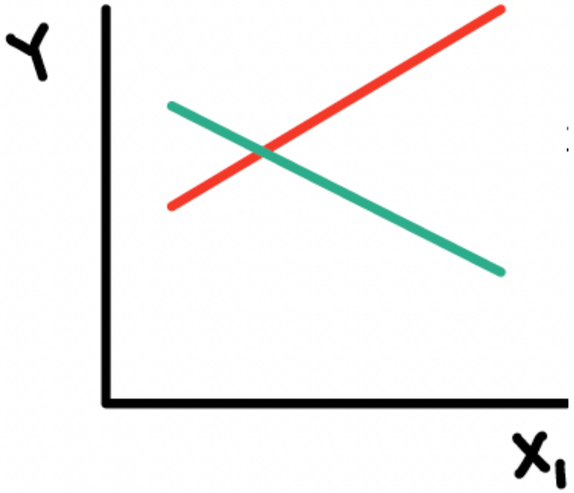


Poll everywhere question 1

What is an effect modifier?

- An additional variable in the model
  - Outside of the main relationship between Y and X_1 that we are studying
- An effect modifier will change the effect of X_1 on Y depending on its value
  - Aka: as the effect modifier's values change, so does the association between Y and X_1
  - So the coefficient estimating the relationship between Y and X_1 changes with another variable
- Example: A breast cancer education program (the exposure) that is much more effective in reducing breast cancer (outcome) in rural areas than urban areas.
  - Location (rural vs. urban) is the EMM





How do we include an effect modifier in the model?

- Interactions!!
- We can incorporate interactions into our model through product terms:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

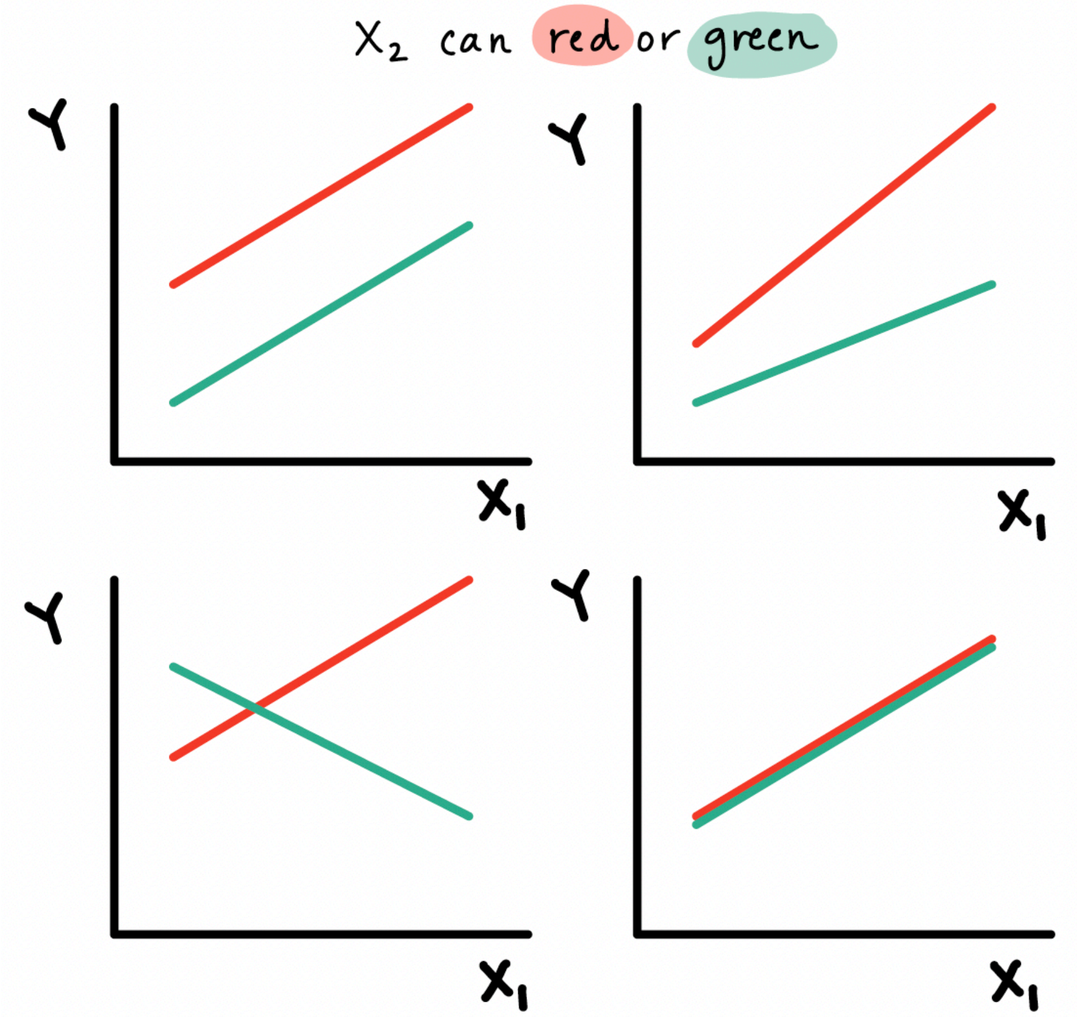
- Terminology:
  - main effect parameters: β_1, β_2
    - The main effect models estimate the *average* X_1 and X_2 effects
  - interaction parameter: β_3

Types of interactions / non-interactions

- Common types of interactions:
  - Synergism: X_2 strengthens the X_1 effect
  - Antagonism:X_2 weakens the X_1 effect

- If the interaction coefficient is not significant
  - No evidence of effect modification, i.e., the effect of X_1 does not vary with X_2

- If the main effect of X_2 is also not significant
  - No evidence that X_2 is a confounder



Learning Objectives

This time:

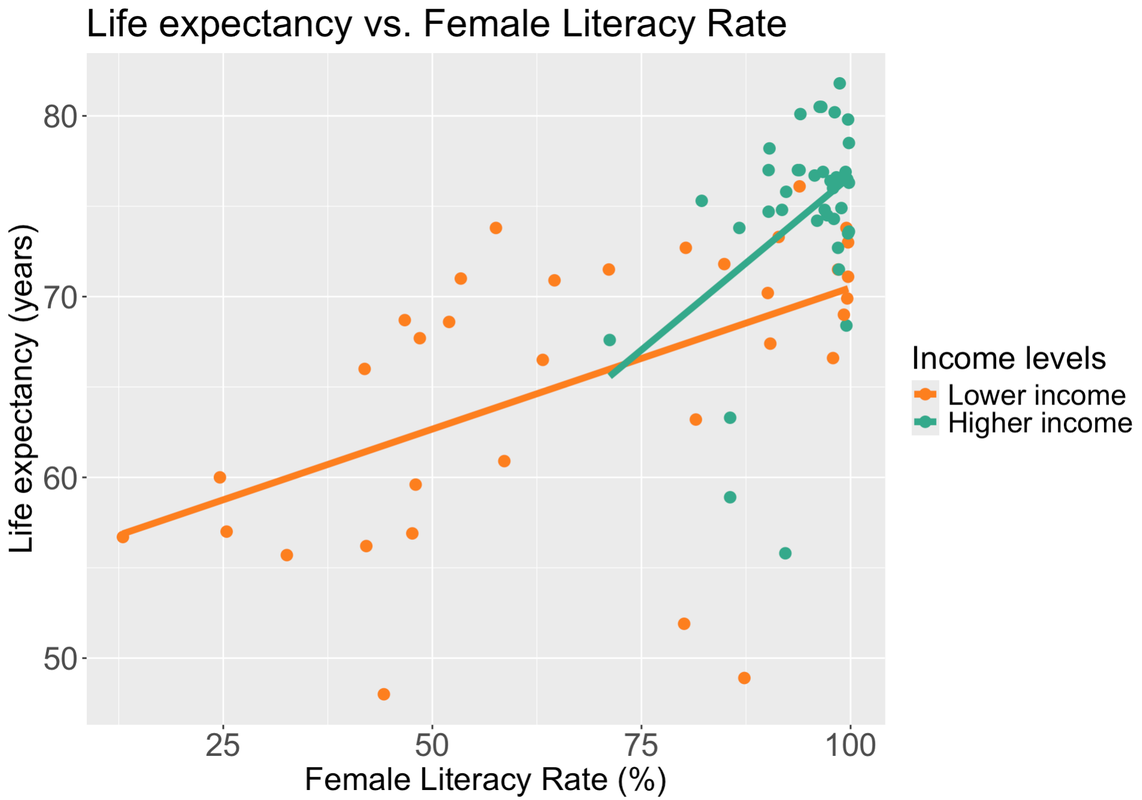
1. Define confounders and effect modifiers, and how they interact with the main relationship we model.
2. Interpret the interaction component of a model with **a binary categorical covariate and continuous covariate**, and how the main variable's effect changes.
3. Interpret the interaction component of a model with **a multi-level categorical covariate and continuous covariate**, and how the main variable's effect changes.

Next time:

4. Interpret the interaction component of a model with **two categorical covariates**, and how the main variable's effect changes.
5. Interpret the interaction component of a model with **two continuous covariates**, and how the main variable's effect changes.

Do we think income level is an effect modifier for female literacy rate?

- Let's say we only have two income groups: low income and high income
- We can start by visualizing the relationship between life expectancy and female literacy rate *by income level*
- Questions of interest: Is the effect of female literacy rate on life expectancy differ depending on income level?
  - This is the same as: Is income level is an effect modifier for female literacy rate?
  - "effect of female literacy rate" differing = different slopes between FLR and LE depending on the income group
- Let's run an interaction model to see!



Model with interaction between a *binary categorical and continuous variables*

Model we are fitting:

$$LE = \beta_0 + \beta_1 FLR + \beta_2 I(\text{high income}) + \beta_3 FLR \cdot I(\text{high income}) + \epsilon$$

- LE as outcome
- FLR as continuous variable that is our main variable of interest
- $I(\text{high income})$ as the indicator that income level is "high income" (binary categorical variable)

```
m_int_inc2 = gapm_sub %>%
  lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate + income_levels2 +
       FemaleLiteracyRate*income_levels2)
```

OR

```
m_int_inc2 = gapm_sub %>%
  lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate*income_levels2)
```

Displaying the regression table and writing fitted regression equation

```
1 tidy(m_int_inc2, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	54.849	2.846	19.270	0.000	49.169	60.529
FemaleLiteracyRate	0.156	0.039	3.990	0.000	0.078	0.235
income_levels2Higher income	−16.649	15.364	−1.084	0.282	−47.308	14.011
FemaleLiteracyRate:income_levels2Higher income	0.228	0.164	1.392	0.168	−0.099	0.555

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 I(\text{high income}) + \widehat{\beta}_3 FLR \cdot I(\text{high income})$$

$$\widehat{LE} = 54.85 + 0.156 \cdot FLR - 16.65 \cdot I(\text{high income}) + 0.228 \cdot FLR \cdot I(\text{high income})$$

Poll Everywhere Question 2

Comparing fitted regression lines for each income level

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 I(\text{high income}) + \widehat{\beta}_3 FLR \cdot I(\text{high income})$$

$$\widehat{LE} = 54.85 + 0.156 \cdot FLR - 16.65 \cdot I(\text{high income}) + 0.228 \cdot FLR \cdot I(\text{high income})$$

For lower income countries: $I(\text{high income}) = 0$

$$\begin{aligned}
\widehat{LE} &= \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 FLR \cdot 0 \\
\widehat{LE} &= 54.85 + 0.156 \cdot FLR - 16.65 \cdot 0 + \\
&\quad 0.228 \cdot FLR \cdot 0 \\
\widehat{LE} &= 54.85 + 0.156 \cdot FLR
\end{aligned}$$

For higher income countries: $I(\text{high income}) = 1$

$$\begin{aligned}
\widehat{LE} &= \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 \cdot 1 + \widehat{\beta}_3 FLR \cdot 1 \\
\widehat{LE} &= 54.85 + 0.156 \cdot FLR - 16.65 \cdot 1 + \\
&\quad 0.228 \cdot FLR \cdot 1 \\
\widehat{LE} &= (54.85 - 16.65 \cdot 1) + \\
&\quad (0.156 \cdot FLR + 0.228 \cdot FLR \cdot 1) \\
\widehat{LE} &= (54.85 - 16.65) + (0.156 + 0.228) \cdot FLR \\
\widehat{LE} &= 38.2 + 0.384 \cdot FLR
\end{aligned}$$

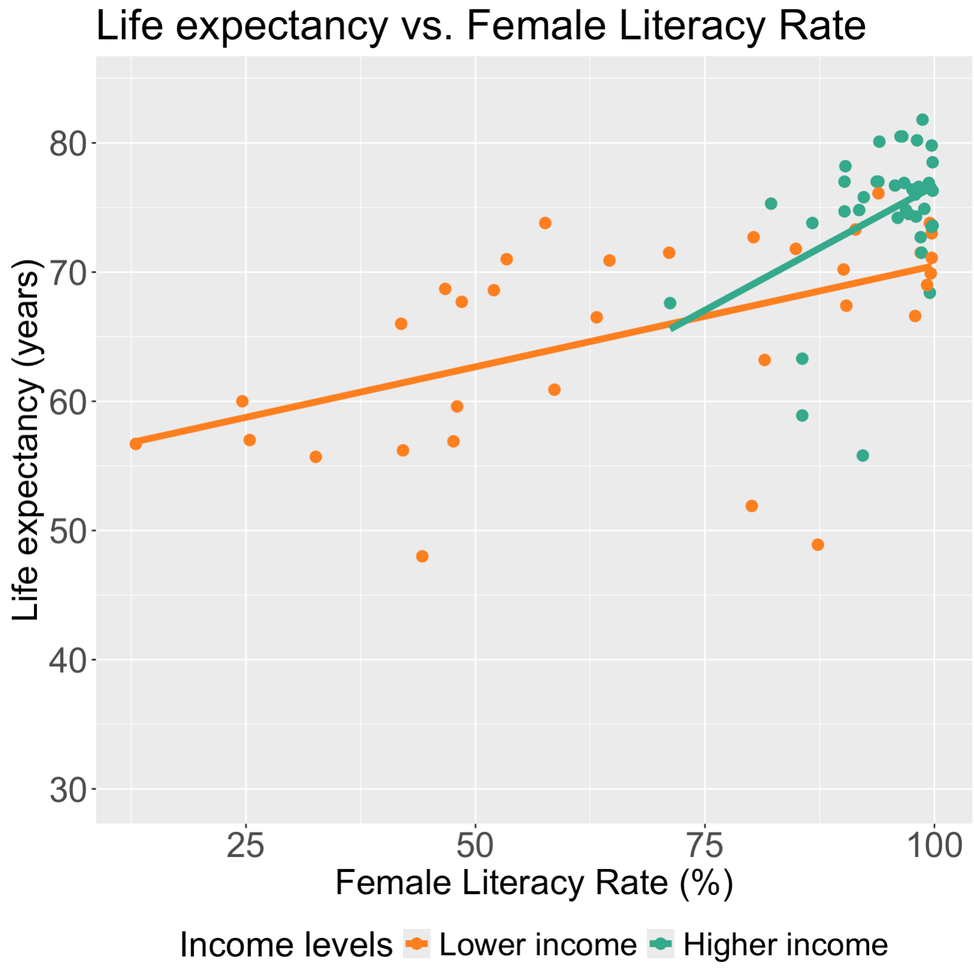
Let's take a look back at the plot

For lower income countries: $I(\text{high income}) = 0$

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 FLR$$
$$\widehat{LE} = 54.85 + 0.156 \cdot FLR$$

For higher income countries: $I(\text{high income}) = 1$

$$\widehat{LE} = (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) FLR$$
$$\widehat{LE} = (54.85 - 16.65) + (0.156 + 0.228) \cdot FLR$$
$$\widehat{LE} = 38.2 + 0.384 \cdot FLR$$



Poll Everywhere Question 3

PAUSE: Centering continuous variables when including interactions

- For the high income group, the mean life expectancy had a regression line with a small intercept

$$\begin{aligned}\widehat{LE} &= (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3)FLR \\ \widehat{LE} &= (54.85 - 16.65) + (0.156 + 0.228) \cdot FLR \\ \widehat{LE} &= 38.2 + 0.384 \cdot FLR\end{aligned}$$

- Intercept of 38.2 is misleading because
  - Makes you think some of the life expectancies for high income countries are lower than that of low income countries (depending on the FLR)
  - There are no high income countries with FLR less than ~70%
- Other online sources about when and when not to center:
  - The why and when of centering continuous predictors in regression modeling
  - When not to center a predictor variable in regression

Centering a variable

- Centering a variable means that we will subtract the mean or median (or other measurement of center) from the measured value
- Mean centered:

$$X_i^c = X_i - \overline{X}$$

- Median centered:

$$X_i^c = X_i - \text{median } X$$

- Centering the continuous variables in a model (when they are involved in interactions) helps with:
  - Interpretations of the coefficient estimates
  - Correlation between the main effect for the variable and the interaction that it is involved with
    - To be discussed in future lecture: leads to multicollinearity issues

It'll be helpful to center female literacy rate

- Centering female literacy rate:

$$FLR^c = FLR - \overline{FLR}$$

- Centering in R:

```
1  gapm_sub = gapm_sub %>%
2    mutate(FLR_c = FemaleLiteracyRate - median(FemaleLiteracyRate))
```

- I'm going to print the mean so I can use it for my interpretations

```
1  (mean_FLR = mean(gapm_sub$FemaleLiteracyRate))
```

```
[1] 82.03056
```

- Now all intercept values (in each respective world region) will be the mean life expectancy when female literacy rate is 82.03%
- We will used center FLR for the rest of the lecture

Displaying the regression table and writing fitted regression equation AGAIN

```
m_int_inc2 = gapm_sub %>%
  lm(formula = LifeExpectancyYrs ~ FLR_c*income_levels2)
```

```
tidy(m_int_inc2, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	69.281	1.387	49.964	0.000	66.514	72.047
FLR_c	0.156	0.039	3.990	0.000	0.078	0.235
income_levels2Higher income	4.405	1.725	2.554	0.013	0.963	7.848
FLR_c:income_levels2Higher income	0.228	0.164	1.392	0.168	−0.099	0.555

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 FLR^c + \widehat{\beta}_2 I(\text{high income}) + \widehat{\beta}_3 FLR^c \cdot I(\text{high income})$$

$$\widehat{LE} = 69.281 + 0.156 \cdot FLR^c + 4.405 \cdot I(\text{high income}) + 0.228 \cdot FLR^c \cdot I(\text{high income})$$

Interpretation for interaction between binary categorical and continuous variables

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 FLR^c + \widehat{\beta}_2 I(\text{high income}) + \widehat{\beta}_3 FLR^c \cdot I(\text{high income})$$

$$\widehat{LE} = \left[\widehat{\beta}_0 + \widehat{\beta}_2 \cdot I(\text{high income})\right] + \underbrace{\left[\widehat{\beta}_1 + \widehat{\beta}_3 \cdot I(\text{high income})\right]}_{\text{FLR's effect}} FLR^c$$

- Interpretation:
  - β_3 = mean change in female literacy rate's effect, comparing higher income to lower income levels
    - AKA: the change in slopes (for line between FLR and LE) comparing high income to low income
  - where the "female literacy rate effect" = change in mean life expectancy per percent increase in female literacy (slope) with income level held constant, i.e. "adjusted female literacy rate effect"
- In summary, the interaction term can be interpreted as "difference in adjusted female literacy rate effect comparing higher income to lower income levels"
- It will be helpful to test the interaction to round out this interpretation!!

Test interaction between binary categorical and continuous variables

- We run an F-test for a single coefficient (β_3) in the below model (see Lesson 10, MLR: Using the F-test)

$$LE = \beta_0 + \beta_1 FLR^c + \beta_2 I(\text{high income}) + \beta_3 FLR^c \cdot I(\text{high income}) + \epsilon$$

Null H_0

$\beta_3 = 0$

Alternative H_1

$\beta_3 \neq 0$

Null / Smaller / Reduced model

$$LE = \beta_0 + \beta_1 FLR^c + \beta_2 I(\text{high income}) + \epsilon$$

Alternative / Larger / Full model

$$LE = \beta_0 + \beta_1 FLR^c + \beta_2 I(\text{high income}) + \beta_3 FLR^c \cdot I(\text{high income}) + \epsilon$$

- I'm going to be skipping steps so please look back at Lesson 10 for full steps (required in HW 4)

Test interaction between binary categorical and continuous variables

- Fit the reduced and full model

```
m_int_inc_red = lm(LifeExpectancyYrs ~ FLR_c + income_levels2,
                   data = gapm_sub)
m_int_inc_full = lm(LifeExpectancyYrs ~ FLR_c + income_levels2 +
                    FLR_c*income_levels2, data = gapm_sub)
```

- Display the ANOVA table with F-statistic and p-value

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FLR_c + income_levels2	69.000	2,407.667	NA	NA	NA	NA
LifeExpectancyYrs ~ FLR_c + income_levels2 + FLR_c * income_levels2	68.000	2,340.948	1.000	66.719	1.938	0.168

- Conclusion: There is not a significant interaction between female literacy rate and income level (p = 0.168).
  - If significant, we say more: For higher income levels, for every one percent increase in female literacy rate, the mean life expectancy increases 0.384 years. For lower income levels, for every one percent increase in female literacy rate, the mean life expectancy increases 0.156 years. Thus, the female literacy rate almost doubles comparing high income to low income levels.

Learning Objectives

This time:

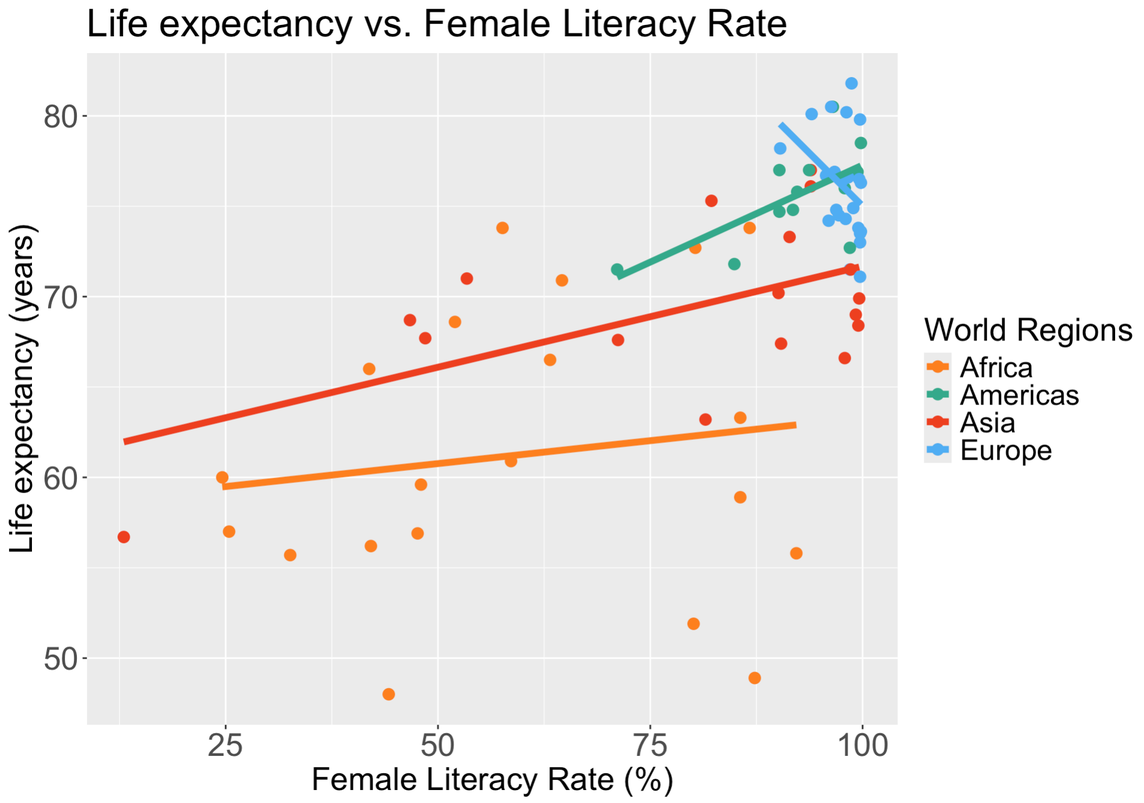
1. Define confounders and effect modifiers, and how they interact with the main relationship we model.
2. Interpret the interaction component of a model with **a binary categorical covariate and continuous covariate**, and how the main variable's effect changes.
3. Interpret the interaction component of a model with **a multi-level categorical covariate and continuous covariate**, and how the main variable's effect changes.

Next time:

4. Interpret the interaction component of a model with **two categorical covariates**, and how the main variable's effect changes.
5. Interpret the interaction component of a model with **two continuous covariates**, and how the main variable's effect changes.

Do we think world region is an effect modifier for female literacy rate?

- We can start by visualizing the relationship between life expectancy and female literacy rate *by world region*
- Questions of interest: Does the effect of female literacy rate on life expectancy differ depending on world region?
  - This is the same as: Is world region is an effect modifier for female literacy rate?
- Let's run an interaction model to see!



Model with interaction between a *multi-level categorical and continuous variables*

Model we are fitting:

$$\begin{aligned} LE = &\beta_0 + \beta_1 FLR^c + \beta_2 I(\text{Americas}) + \beta_3 I(\text{Asia}) + \beta_4 I(\text{Europe}) + \\ &\beta_5 FLR^c \cdot I(\text{Americas}) + \beta_6 FLR^c \cdot I(\text{Asia}) + \beta_7 FLR^c \cdot I(\text{Europe}) + \epsilon \end{aligned}$$

- LE as life expectancy
- FLR^c as centered female literacy rate (continuous variable)
- $I(\text{Americas}), I(\text{Asia}), I(\text{Europe})$ as the indicator for each world region

In R:

```
m_int_wr = gapm_sub %>% lm(formula = LifeExpectancyYrs ~ FLR_c + four_regions +
                     FLR_c*four_regions)
```

OR

```
m_int_wr = gapm_sub %>% lm(formula = LifeExpectancyYrs ~ FLR_c * four_regions)
```

Displaying the regression table and writing fitted regression equation

```
1 tidy(m_int_wr, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt_n
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	62.906	2.050	30.680	0.000	58.810	67.002
FLR_c	0.051	0.053	0.957	0.342	−0.055	0.157
four_regionsAmericas	12.706	2.518	5.046	0.000	7.676	17.737
four_regionsAsia	7.910	2.477	3.193	0.002	2.962	12.859
four_regionsEurope	15.732	3.485	4.514	0.000	8.770	22.694
FLR_c:four_regionsAmericas	0.164	0.197	0.830	0.410	−0.231	0.558
FLR_c:four_regionsAsia	0.061	0.073	0.830	0.410	−0.086	0.208
FLR_c:four_regionsEurope	−0.519	0.476	−1.090	0.280	−1.471	0.432

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 I(\text{Americas}) + \widehat{\beta}_3 I(\text{Asia}) + \widehat{\beta}_4 I(\text{Europe}) + \\ & \widehat{\beta}_5 FLR \cdot I(\text{Americas}) + \widehat{\beta}_6 FLR \cdot I(\text{Asia}) + \widehat{\beta}_7 FLR \cdot I(\text{Europe})\end{aligned}$$

$$\begin{aligned}\widehat{LE} =& 62.906 + 0.051 \cdot FLR + 12.706 \cdot I(\text{Americas}) + 7.91 \cdot I(\text{Asia}) + 15.732 \cdot I(\text{Europe}) + \\ & 0.164 \cdot FLR \cdot I(\text{Americas}) + 0.061 \cdot FLR \cdot I(\text{Asia}) - 0.519 \cdot FLR \cdot I(\text{Europe})\end{aligned}$$

Comparing fitted regression lines for each world region

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 I(\text{Americas}) + \widehat{\beta}_3 I(\text{Asia}) + \widehat{\beta}_4 I(\text{Europe}) + \\ & \widehat{\beta}_5 FLR \cdot I(\text{Americas}) + \widehat{\beta}_6 FLR \cdot I(\text{Asia}) + \widehat{\beta}_7 FLR \cdot I(\text{Europe}) \\ \widehat{LE} =& 62.906 + 0.051 \cdot FLR + 12.706 \cdot I(\text{Americas}) + 7.91 \cdot I(\text{Asia}) + 15.732 \cdot I(\text{Europe}) + \\ & 0.164 \cdot FLR \cdot I(\text{Americas}) + 0.061 \cdot FLR \cdot I(\text{Asia}) - 0.519 \cdot FLR \cdot I(\text{Europe})\end{aligned}$$

Africa

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \\ & \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 0 + \\ & \widehat{\beta}_4 \cdot 0 + \widehat{\beta}_5 FLR \cdot 0 + \\ & \widehat{\beta}_6 FLR \cdot 0 + \widehat{\beta}_7 FLR \cdot 0 \\ \widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR\end{aligned}$$

The Americas

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \\ & \widehat{\beta}_2 \cdot 1 + \widehat{\beta}_3 \cdot 0 + \\ & \widehat{\beta}_4 \cdot 0 + \widehat{\beta}_5 FLR \cdot 1 + \\ & \widehat{\beta}_6 FLR \cdot 0 + \widehat{\beta}_7 FLR \cdot 0 \\ \widehat{LE} =& (\widehat{\beta}_0 + \widehat{\beta}_2) + \\ & (\widehat{\beta}_1 + \widehat{\beta}_5) FLR\end{aligned}$$

Asia

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \\ & \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 1 + \\ & \widehat{\beta}_4 \cdot 0 + \widehat{\beta}_5 FLR \cdot 0 + \\ & \widehat{\beta}_6 FLR \cdot 1 + \widehat{\beta}_7 FLR \cdot 0 \\ \widehat{LE} =& (\widehat{\beta}_0 + \widehat{\beta}_3) + \\ & (\widehat{\beta}_1 + \widehat{\beta}_6) FLR\end{aligned}$$

Europe

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \\ & \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 0 + \\ & \widehat{\beta}_4 \cdot 1 + \widehat{\beta}_5 FLR \cdot 0 + \\ & \widehat{\beta}_6 FLR \cdot 0 + \widehat{\beta}_7 FLR \cdot 1 \\ \widehat{LE} =& (\widehat{\beta}_0 + \widehat{\beta}_4) + \\ & (\widehat{\beta}_1 + \widehat{\beta}_7) FLR\end{aligned}$$

Interpretation for interaction between multi-level categorical and continuous variables

$$\begin{aligned}\widehat{LE} =& \widehat{\beta}_0 + \widehat{\beta}_1 FLR + \widehat{\beta}_2 I(\text{Americas}) + \widehat{\beta}_3 I(\text{Asia}) + \widehat{\beta}_4 I(\text{Europe}) + \\ & \widehat{\beta}_5 FLR \cdot I(\text{Americas}) + \widehat{\beta}_6 FLR \cdot I(\text{Asia}) + \widehat{\beta}_7 FLR \cdot I(\text{Europe}) \\ \widehat{LE} =& \left[\widehat{\beta}_0 + \widehat{\beta}_2 I(\text{Americas}) + \widehat{\beta}_3 I(\text{Asia}) + \widehat{\beta}_4 I(\text{Europe})\right] + \\ & \underbrace{\left[\widehat{\beta}_1 + \widehat{\beta}_5 \cdot I(\text{Americas}) + \widehat{\beta}_6 \cdot I(\text{Asia}) + \widehat{\beta}_7 \cdot I(\text{Europe})\right]}_{\text{FLR's effect}} FLR\end{aligned}$$

- Interpretation:
  - β_5 = mean change in female literacy rate's effect, comparing countries in the Americas to countries in Africa
  - β_6 = mean change in female literacy rate's effect, comparing countries in Asia to countries in Africa
  - β_7 = mean change in female literacy rate's effect, comparing countries in Europe to countries in Africa
- It will be helpful to test the interaction to round out this interpretation!!

Test interaction between multi-level categorical & continuous variables

- We run an F-test for a group of coefficients ($\beta_5, \beta_6, \beta_7$) in the below model (see lesson 10)

$$\begin{aligned} LE = &\beta_0 + \beta_1 FLR + \beta_2 I(\text{Americas}) + \beta_3 I(\text{Asia}) + \beta_4 I(\text{Europe}) + \\ &\beta_5 FLR \cdot I(\text{Americas}) + \beta_6 FLR \cdot I(\text{Asia}) + \beta_7 FLR \cdot I(\text{Europe}) + \epsilon \end{aligned}$$

Null H_0

$\beta_5 = \beta_6 = \beta_7 = 0$

Alternative H_1

$\beta_5 \neq 0$ and/or $\beta_6 \neq 0$ and/or $\beta_7 \neq 0$

Null / Smaller / Reduced model

$$\begin{aligned} LE = &\beta_0 + \beta_1 FLR + \beta_2 I(\text{Americas}) + \\ &\beta_3 I(\text{Asia}) + \beta_4 I(\text{Europe}) + \epsilon \end{aligned}$$

Alternative / Larger / Full model

$$\begin{aligned} LE = &\beta_0 + \beta_1 FLR + \beta_2 I(\text{Americas}) + \beta_3 I(\text{Asia}) + \\ &\beta_4 I(\text{Europe}) + \beta_5 FLR \cdot I(\text{Americas}) + \\ &\beta_6 FLR \cdot I(\text{Asia}) + \beta_7 FLR \cdot I(\text{Europe}) + \epsilon \end{aligned}$$

Test interaction between multi-level categorical & continuous variables

- Fit the reduced and full model

```
m_int_wr_red = lm(LifeExpectancyYrs ~ FLR_c + four_regions,
                  data = gapm_sub)
m_int_wr_full = lm(LifeExpectancyYrs ~ FLR_c + four_regions+
                   FLR_c*four_regions, data = gapm_sub)
```

- Display the ANOVA table with F-statistic and p-value

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FLR_c + four_regions	67.000	1,705.881	NA	NA	NA	NA
LifeExpectancyYrs ~ FLR_c + four_regions + FLR_c * four_regions	64.000	1,641.151	3.000	64.731	0.841	0.476

- Conclusion: There is not a significant interaction between female literacy rate and world region (p = 0.478).
- World region is NOT an effect measure modifier of FLR on LE