A Regression (Re)Introduction Session 4

Danielle Wallace, *ASU*

For CSG, July 12, 2021

Questions from last session

- What do interactions contribute? Why do we do them? What do they add to analyses?
  - Let's talk about this now
- How to you interpret and write up an interaction effect?
  - After we cover all interactions

What do interactions contribute?

- Remember that interactions are when the association between one independent variable and the dependent variable differs depending on the value of the second independent variable, known as a modifier
  - Put another way, the association between X_1 and Y is conditional on X_2
  - Put another way again, specific combinations of X_1 and X_2 determine the value of Y.

What interactions contribute

- When you can't specify the relationship (direction/size) between one predictor and the outcome variable without specifying the values of a second variable
- Without the inclusion of a needed interaction, you have omitted variable bias
  - Type of estimation/model specification bias
  - If important independent variables (including interactions) are omitted from the model, estimated coefficients on other variables may be biased, leading to spurious conclusions
  - Especially true for the main effects variables when an interaction is missing
- In short, without needed interactions, you've miss specified the relationship between your DV & IV.

Miller, p. 37

An example of omitted variable bias

- How does this work?
- When you're missing a variable in a model, it simply increases error. That "noise" is not a problem, UNTIL it's correlated with other predictors.
- "Intuitively, omitted variable bias occurs when the independent variable (the X) that we have included in our model picks up the effect of some other variable that we have omitted from the model."
- Sounds a lot like what an interaction is: "the association between X_1 and Y is conditional on X_2"

https://ei.com/wp-content/uploads/2018/01/OmittedVariable.pdf

Visual Example from Economics about Car Buying

"Consider what happens to price and quantity demanded when we move from a lower-income area to a higher-income area.

Because higher incomes are positively related to prices, prices will increase. And because quantity demanded and income have a negative relationship, the quantity will decrease.

If we fail to control for income, we will mistakenly fully attribute the decline in quantity demanded to the increase in price, which in reality is partially caused by the fact that higher-income consumers are less interested in purchasing domestic cars than lower-income consumers."

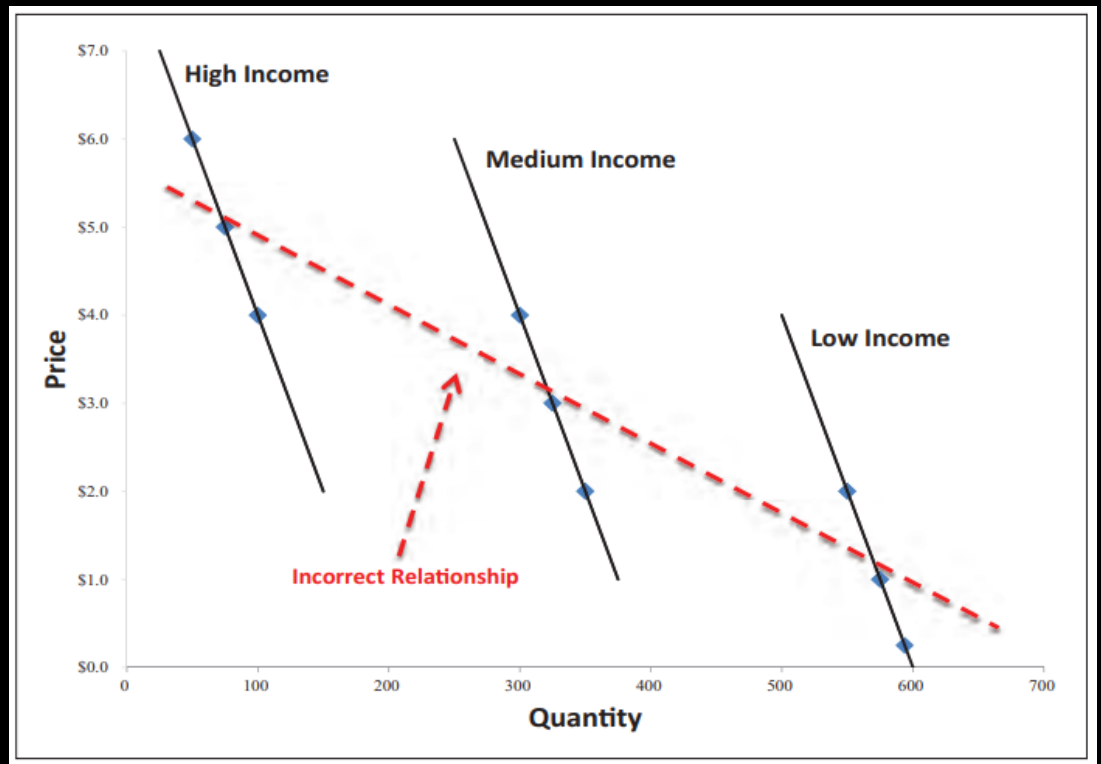


Figure 5

https://ei.com/wp-content/uploads/2018/01/OmittedVariable.pdf

How do you handle this?

- Carefully research what should be included in the model
  - Use the work of previous research to define – conceptually – what should be in the model
  - I stress conceptually: we don't want to replicate measurement/estimation problems
- Do the best you can with the data you have
  - Add outside data when needed if possible (i.e., adding census data, etc.).
- Pay attention to how well the model fits; if it looks wonky, it likely is.

How do you handle this?

- What if you done the background research and have modeled everything you think you need?
- You can't test for a variable you don't have…why omitted variable bias is so difficult to fix
- But in the context of interactions…you can test to see if an interaction contributes to model fit.
  - Let's go back to the example we used for interactions between categorical and linear variables.

Example Re-cap

- Context: The KS disparity data
- Predicting LSI score with age at admission and black
- Want to examine whether black **modifies** the relationship between age at admission and LSI score.
- We'll run two models:

$$\hat{y} = \beta_0 + \beta_{age}x_1 + \beta_{black}x_2$$

$$\hat{y} = \beta_0 + \beta_{age}x_1 + \beta_{black}x_2 + \beta_{interaction}(Black * Age)$$

- We'll save the estimates for both, then run a likelihood ratio test ("lrtest") in Stata to determine if the interaction significantly contributed to model fit
- If the likelihood ratio test is significant, then the interaction significantly contributes to model fit and needs to be included.

What is an Likelihood Ratio Test?

- Using nested models, likelihood ratio tests determines if the log-likelihood values associated with model 2 are significantly different from model 1.

$$LR = -2(L_1 - L_0)$$
$$df = d_0 - d_1$$

- Nested models: Model 2 (L_1) has the most parameters and the parameters of Model 1 (L_0) are seen in the second model
- Uses the χ^2 distribution

What are log-likelihood values?

- Remember that you have your observed values of y and your predicted or estimated values of y, or $\hat{y}$.
- The log-likelihood value of a model is essentially the sum of the deviations between each observed value of y and your estimated values of y (it's more complicated than this, but it's a good conceptual overview).
- Thus a log-likelihood value of a model is an estimate of how well you did predicting y based on the deviations of y from $\hat{y}$.

Back to the Likelihood Ratio Test…

- The likelihood ratio test effectively determines how well you did at cleaning up the deviation of y from $\hat{y}$ through the inclusion of the interaction.
- Can be used for all other types of models that use maximum likelihood estimation (logits, cox hazard models, etc.…).
- Really powerful test for knowing how well your model is doing parameter wise.

Results from Models

Model 1 – No Interaction

```
. reg totscore black age
```

Source	SS	df	MS
Model	1911.64655	2	955.823274
Residual	900150.213	14,474	62.1908396
Total	902061.859	14,476	62.3143036

Number of obs	14,477
F(2, 14474)	15.37
Prob > F	0.0000
R-squared	0.0021
Adj R-squared	0.0020
Root MSE	7.8861

totscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
black	-.0685031	.1463253	-0.47	0.640	-.3553193	.2183131
age	-.0334128	.0060347	-5.54	0.000	-.0452416	-.021584
_cons	28.28188	.137516	205.66	0.000	28.01233	28.55143

Model 2 – with Black * Age

```
. reg totscore black age i.black#c.age
```

Source	SS	df	MS
Model	7728.09928	3	2576.03309
Residual	894333.76	14,473	61.7932536
Total	902061.859	14,476	62.3143036

Number of obs	14,477
F(3, 14473)	41.69
Prob > F	0.0000
R-squared	0.0086
Adj R-squared	0.0084
Root MSE	7.8609

totscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
black	-2.419455	.2828288	-8.55	0.000	-2.973836	-1.865075
age	-.0708223	.0071451	-9.91	0.000	-.0848276	-.0568169
black#c.age						
1	.1284558	.0132402	9.70	0.000	.1025033	.1544084
_cons	28.98755	.1551775	186.80	0.000	28.68338	29.29171

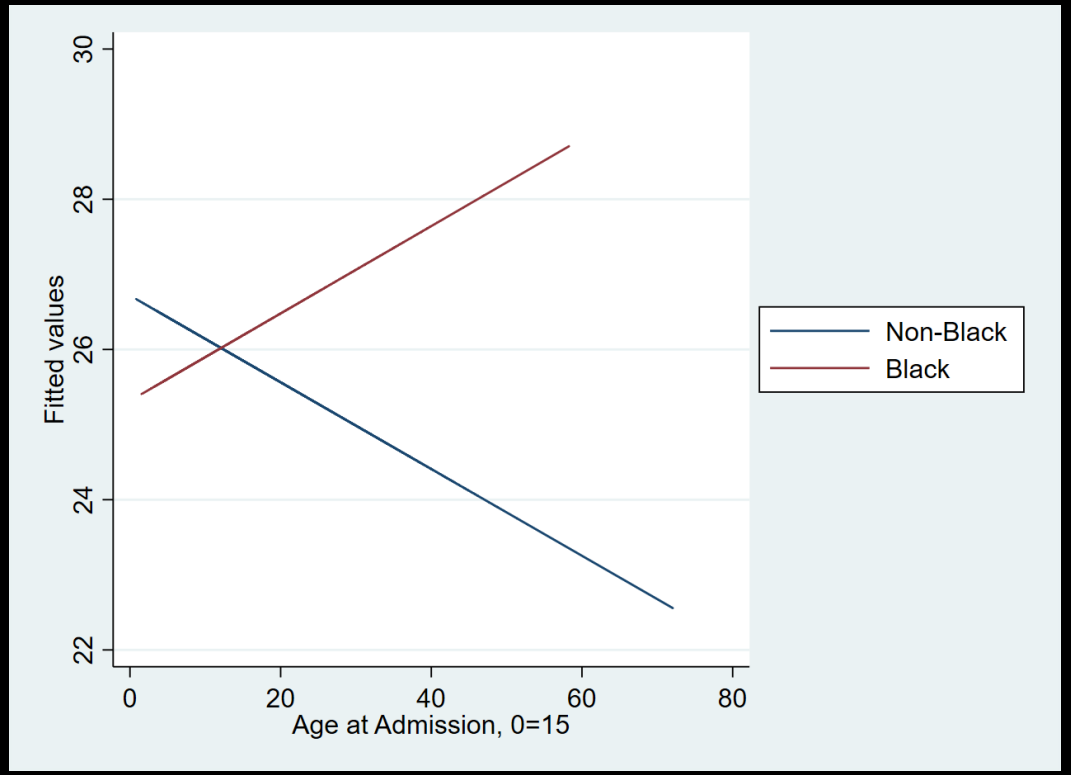
Results of “lrtest”

- Significant likelihood ratio test
- The inclusion of the interaction significantly improves model fit and needs to be included
- Helps decide when to include interactions

```
. lrtest m2 m1

Likelihood-ratio test
Assumption: m1 nested within m2

 LR chi2(1) =  93.85
Prob > chi2 = 0.0000
```



Types of Interactions

☑ Interacting two categorical variables

☑ Interacting a categorical variable with a linear variable

☒ Interacting two linear variables

Interacting Two Linear Variables

- Context: The KS disparity data
- Predicting LSI score with age at admission and length of stay in prison (los) (sorry if it doesn't make sense! Needed 2 linear vars!)
- Want to examine whether length of stay in prison **modifies** the relationship between age at admission and LSI score.
- Both length of stay in prison and age at admission are linear variables
- Statistically, constructing the interaction remains the same: the two variables are multiplied together.

Working in Do-File for Example #3

- Generate the interaction as a new variable
- Examine values of length of stay and age and compare to the interaction variable (Hint: browse…)
- Run the model with the three different ways of doing the interaction: new variable, #, and ##.
- Compare the results.
  - All methods generate similar results
  - But…what does it mean?
  - Time for a graph!

How do you interpret a linear-linear interaction?

- We can start by interpreting the main effects
- On average, as the length of one's stay increases, their LSI score decreases by 0.001
- On average, as one's age at admission increases, their LSI score decreases by 0.03
- However, taken together, length of stay modifies the relationship between age at admission and LSI
  - Let's graph this…

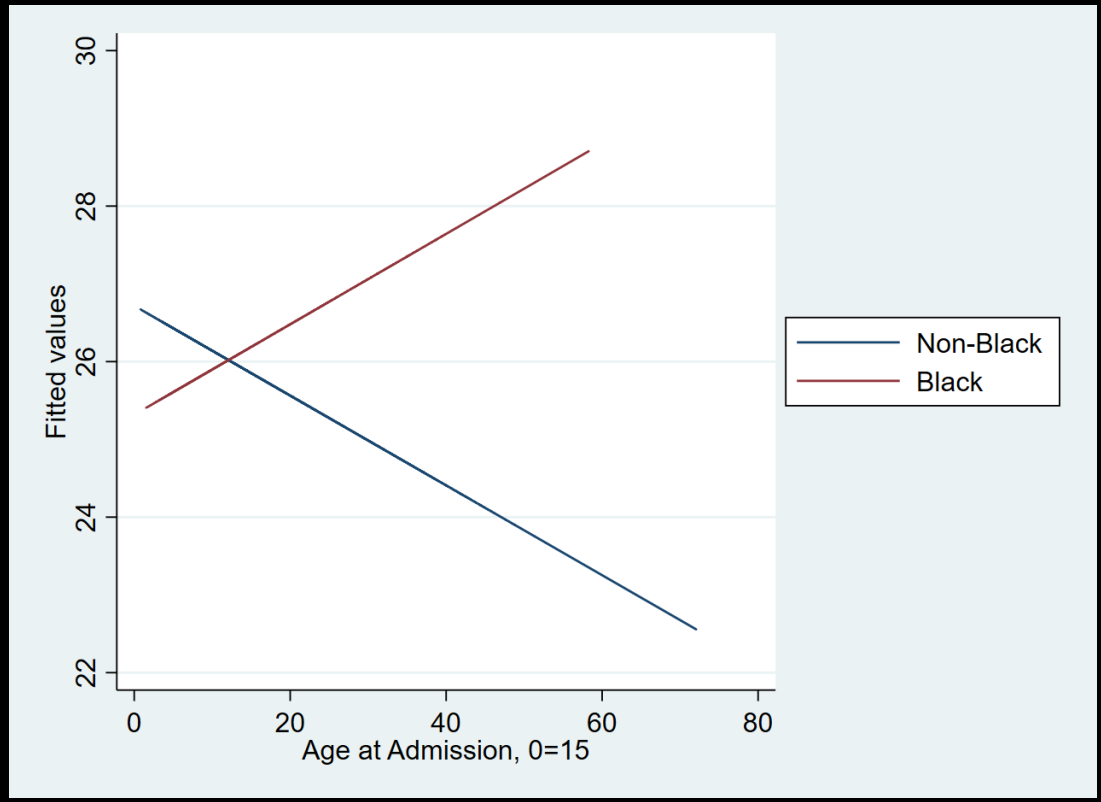
```
. reg totscore los age c.los#c.age
```

Source	SS	df	MS	Number of obs	=	14,477
				F(3, 14473)	=	388.88
Model	67290.0523	3	22430.0174	Prob > F	=	0.0000
Residual	834771.807	14,473	57.6778696	R-squared	=	0.0746
				Adj R-squared	=	0.0744
Total	902061.859	14,476	62.3143036	Root MSE	=	7.5946

totscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
los	-.0013476	.0000857	-15.72	0.000	-.0015157	-.0011795
age	-.0301434	.0071616	-4.21	0.000	-.0441811	-.0161058
c.los#c.age	-.0000122	4.49e-06	-2.71	0.007	-.000021	-3.39e-06
_cons	29.65096	.1512301	196.07	0.000	29.35453	29.94739

Graphing a linear-linear interaction

- We could generate the predicted value for each individual, but that doesn't help us summarize and understand the relationship.
  - Graphic is needed
- I typically will graph the interaction as a "categorical-linear variable" interaction, like we did in Example 2 (see the graph to the right), to help understand what's going on.
- I'll use the values of 1 SD below and above the mean, as well as the mean itself.
  - Use whatever values you think is sensible given the construction of the variable

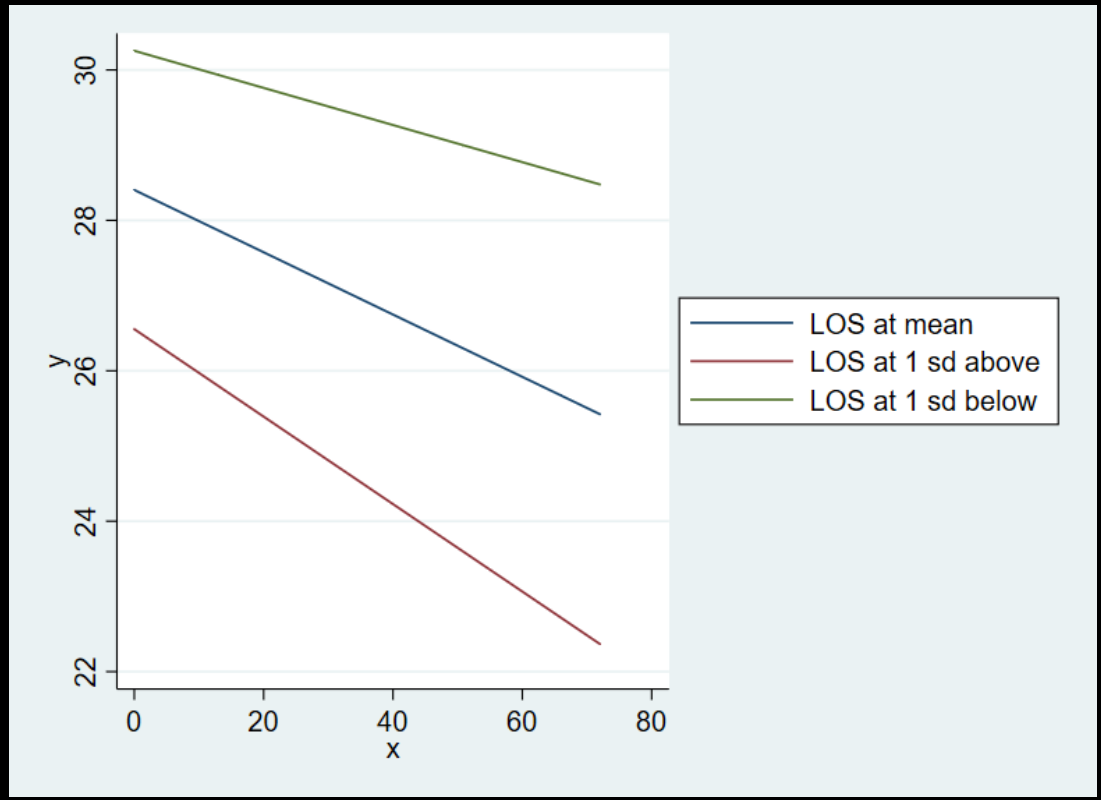


Graphing Using Do-File for Example #3

- Generate macros (or variables) for the mean and standard deviations
- Use the “ereturn list” and “matrix list” functions to grab the names of your variables (if needed).
- Using a “twoway” graphic, plot 3 lines that represent the relationship between age and LSI when length of stay is:
  - 1 SD below the mean
  - At the mean
  - 1 SD above the mean

Interpreting Graphics of linear-linear variable interactions

- When length of stay is at the 1 SD below the mean (green line), as age at admission increases, the predicted LSI score decreases.
- When length of stay is at the mean (blue line), as age at admission increases, the predicted LSI score decreases. This decrease is steeper than when the length of stay is 1 SD below the mean.
- When length of stay is 1 SD above the mean (red line), as age at admission increases, the predicted LSI score decreases. This decrease is more steep than when the length of stay is either at the mean or 1 SD below the mean.
- Interaction type: Effect Size Differences



Presenting and Interpreting the Results of an Interaction

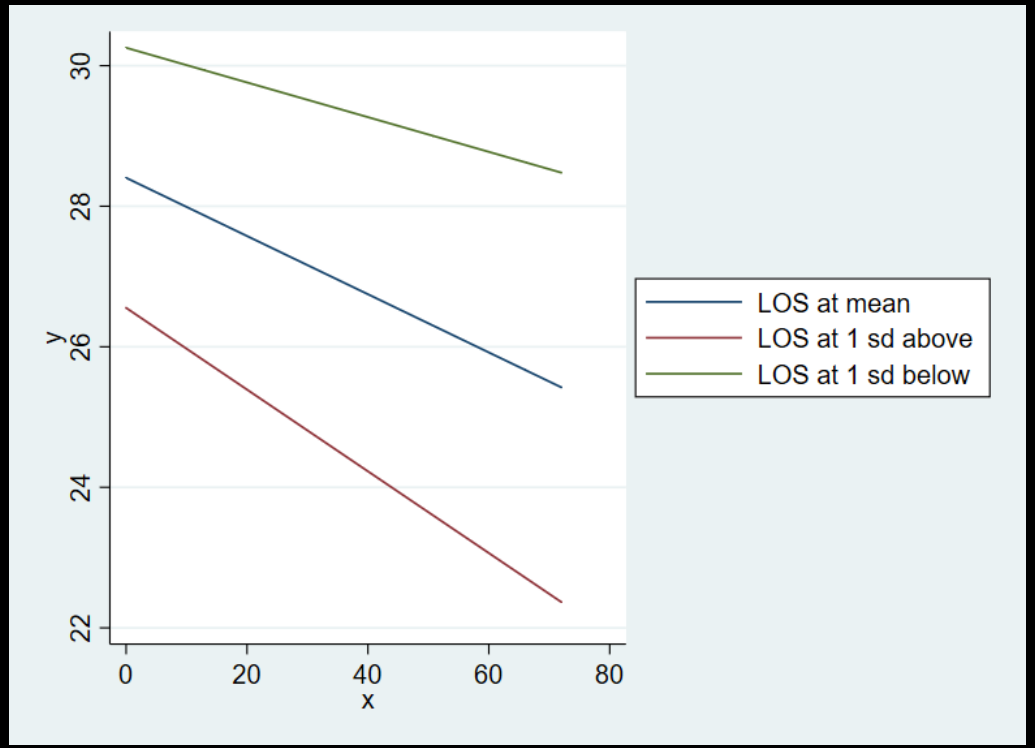
- Adopt the GEE approach to describing a pattern with three or more numbers (i.e., applies to most statistical descriptions, but perfect for interactions)
- Generalization: identify and describe the pattern in general terms
  - Come up with a description that fits the pattern “most” of the time.
- Example: give a representative example to fit that pattern
- Exceptions: explain and illustrate any exceptions

Miller, p. 30-32

GEE in Play

1. Generalize: There is a strong, negative relationship between age at admission and an individuals' LSI score. However, this relationship is contextualized by length of stay.
2. Example:
3. Exception:

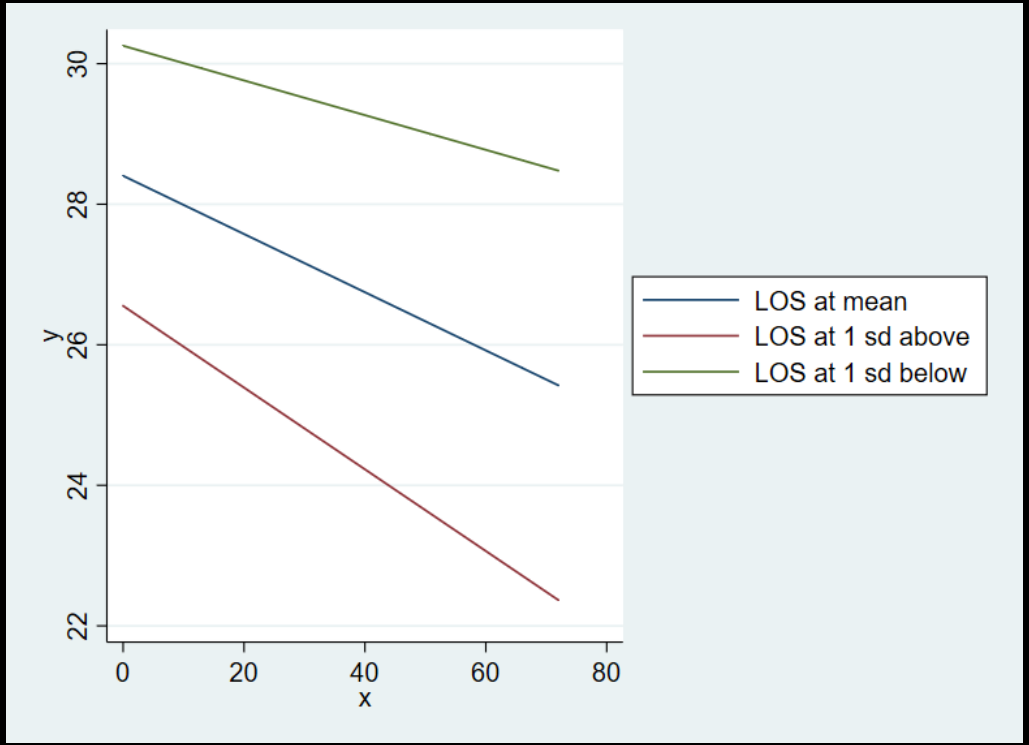
Length of stay in prison **modifies** the relationship between age at admission and LSI score.



GEE in Play

1. Generalize:
2. Example: For example, when an individual's length of stay is at or above average, there are drastic differences in LSI across younger and older individuals. For the youngest admits (age 15), when their length of stay is average, their predicted LSI is 28.4, while for the oldest admits, when their length of stay is average, their predicted LSI is 25.4.
3. Exceptions:

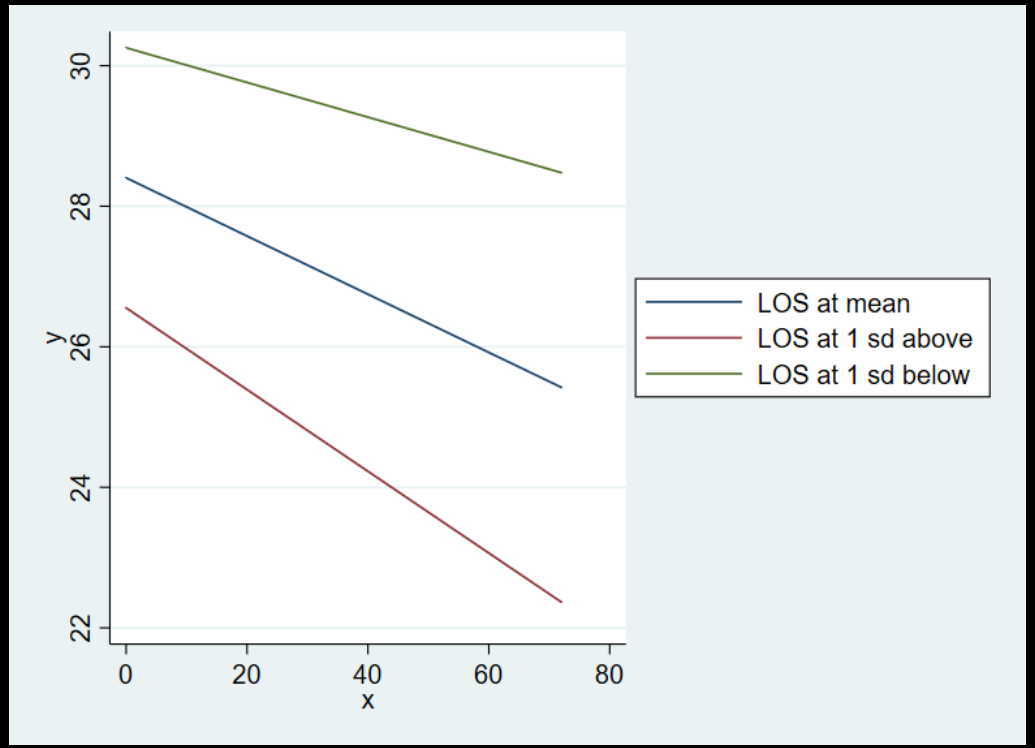
Length of stay in prison **modifies** the relationship between age at admission and LSI score.



GEE in Play

1. Generalize:
2. Example:
3. Exceptions: For individuals with the shortest length of stay, however, the negative relationship between their age at admission and their LSI is less steep than individuals with average or above average length of stays.

Length of stay in prison **modifies** the relationship between age at admission and LSI score.



All together now…

- There is a strong, negative relationship between age at admission and an individuals' LSI score. However, this relationship is contextualized by length of stay.
- For example, when an individual's length of stay is at or above average, there are drastic differences in LSI across younger and older individuals.
- For the youngest admits (age 15), when their length of stay is average, their predicted LSI is 28.4, while for the oldest admits, when their length of stay is average, their predicted LSI is 25.4.
- For individuals with the shortest length of stay, however, the negative relationship between their age at admission and their LSI is less steep than individuals with average or above average length of stays.

Let's Play with Interactions!

Using the following variables, and any we've used in lecture, create an interaction and/or graphic

Other Categorical Variables

- Newcrime: new commitment (1) or technical violation (0)
- Sped: Received Special education (1)
- Hsdip: Graduated HS or GED (1)
- Lessthan9th: received less than a 9^{th} grade education (1)
- Pchronic: Physical disability (1)
- Mchronic: mental health condition (1)

Linear Variables

- Sf12: SF-12 Physical Health Scale; z-score units
- Msf12: SF-12 Mental Health Scale; z-score units

When we re-group, tell us…

1. What type of interaction you did:
    1. Categorical- Categorical
    2. Categorical-Linear
    3. Linear-Linear
2. Your choices for the dependent variable and your two interaction variables
3. Show us your results (table/graphic)
4. Interpretation – if stuck, use GEE
5. Finally – if you get stuck at any of these points, we can chat about fixes.