Concepts you should know:

- Bivariate regression
- Multivariate regression
- Control variables
- Regression coefficient (or β)
- Null and alternative hypotheses for regression
- Regression "model"
- Adjusted R^2
- *F* ratio in regression output
- Reporting regression results

Name that Statistic: To answer the following questions, which statistical procedure would you choose?

- Does a group case management intervention result in fewer overall costs than an individual case management intervention? (_____)
- For decisions about those eligible vs. not eligible for SSI (Supplemental Social Security Income), does ethnicity of applicant matter? (_____)
- For caseworkers in a child welfare agency, is there a relationship between average caseload size and the number of families that reunify in a two-year period? (_____)
- What effect does the initial motivation for treatment have on improvement in a depression scale, controlling for age, gender, and seriousness of initial symptoms? (multivariate regression)

What is “regression”?

- It’s a form of correlation analysis. It can be either bivariate or multivariate.
- Bivariate: It predicts the value of a dependent (or outcome) variable from an observed independent (or predictor) variable
- Multivariate: It predicts the value of a dependent (or outcome) variable from an observed independent (or predictor) variable, *controlling for other variables*

What does "controlling for" mean?

- Controlling for a variable (e.g. gender) means
  1. We collect data on that variable
  2. We include that variable in the list of independent variables in our model
  3. The regression analysis separates out the effects of each attribute (male, female)
  4. You can interpret the resulting statistics for all other variables as if by saying "regardless of gender"
- Even though only some variables might be labeled control variables in the hypothesis, multiple regression analysis uses the same process on all independent variables in the model
  - So you can say about *any* variable, "controlling for the effects of all other variables…"

Bivariate example

- H_A: For children in residential care, the number of strength-based comments by staff is predictive* of the number of behavioral incidents in children.
- Null: There is no relationship between strength-based comments and behavioral incidents in children

*Note: in order to be predictive, the IV must be at least correlated with the DV

General bivariate notation (equation of a line)

$$Y = c + bX$$

Y = dependent variable
X = independent variable
b = slope of the line or the coefficient of the X variable (literally, “b times X”)
c = constant (or the point of intercept at the Y axis) (sometimes you’ll see “a” used as the symbol)

Multivariate example (more than 1 IV)

- H_A: For children in residential care, the number of strength-based comments by staff is predictive of the number of behavioral incidents in children, controlling for seriousness of diagnosis, age, gender and length of time in care. OR, YOU COULD SAY…
- H_A: …compared to other factors, the number of strength-based comments is related to behavioral incidents in children, controlling for ….
- Null: There is no relationship between strength-based comments and behavioral incidents in children

Notation for multivariate regression

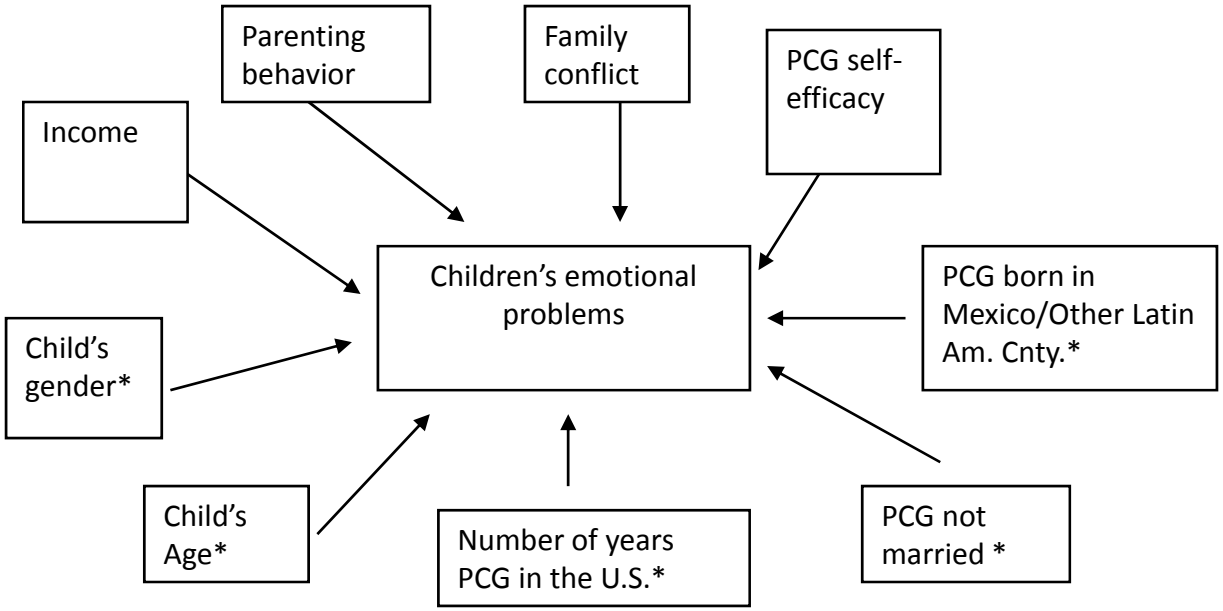
$$Y = c + b_1X_1 + b_2X_2 + b_3X_3 + \ldots b_kX_k$$

Y = dependent variable
X = independent variables
b = regression coefficient for each X
k = number of independent variables

Typical uses

- While regression can be used to predict outcomes, the procedure is most often used to:
  - Determine whether the relationship between the IV and DV is likely due to sampling error
  - Determine the strength and direction of the relationship between the primary IV and DV (as in correlation)
  - Determine the effects of other independent variables (such as control variables) in the relationship between the primary IV and DV

Analysis Model—Predictors of Children's Emotional Problems



*Control variables. All others are hypothesized predictors.

SPSS Output—Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.516[a]	.266	.240	2.696

a. Predictors: (Constant), PCG Not married, Child's Gender: Female compared to male baseline, Number of year PCG in U.S., PCG Self-efficacy score, Family Conflict, PCG Born in Mexico/Latin American Country: Baseline PCG Born in Other Country, Parenting Behavior, Child's Age, Poverty

R is the multiple correlation coefficient (extent to which the IVs, as a group, correlate with the DV). *R Square* (literally, R^2) is the percentage of variance in the DV explained by the IVs, as a group. Adjusted R^2 is an estimated R^2 of the population. ***"24% of the variance in children's emotional problems is a result of the IVs"***

Analysis of Variance Table

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	666.792	9	74.088	10.192	.000[a]
	Residual	1839.040	253	7.269		
	Total	2505.833	262			

a. Predictors: (Constant), PCG Not married, Child's Gender: Female compared to male baseline, Number of year PCG in U.S., PCG Self-efficacy score, Family Conflict, PCG Born in Mexico/Latin American Country: Baseline PCG Born in Other Country, Parenting Behavior, Child's Age, Poverty

b. Dependent Variable: internalizing bpi subscale score

Here, the *F* ratio tests the Null that the overall model results in coefficients that are equal to 0. The alternative hypothesis is that the coefficients do not = 0 (i.e. they show a linear relationship between the variables) ***F*(9,253)=10.19, *p* <.001**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.203	2.093		5.831	.000
	Poverty	-.144	.064	-.146	-2.252	.025
	Parenting Behavior	-.110	.028	-.230	-3.917	.000
	Family Conflict	-.119	.072	-.093	-1.647	.101
	PCG Self-efficacy score	-.141	.054	-.149	-2.628	.009
	Child's Gender: Female compared to male baseline	.291	.339	.047	.861	.390
	Child's Age	.023	.088	.015	.259	.796
	Number of year PCG in U.S.	-.035	.025	-.083	-1.405	.161
	PCG Born in Mexico/Latin American Country: Baseline PCG Born in Other Country	1.475	.481	.196	3.066	.002
	PCG Not married	.412	.365	.063	1.127	.261

a. Dependent Variable: internalizing bpi subscale score

Interpreting the coefficients

- *(Important to know but not necessarily included in the Results):* The regression coefficient (B) in the output indicates (all together now)

  "*the amount of change estimated in the dependent variable for a one-unit change in the independent variable, controlling for other variables in the model.*"

Interpreting the Standardized Beta

- The standardized beta (β) is in standard deviation units (like *z* scores), so that all βs can be compared to each other. **We'll use this column for reporting.**
  - Poverty: β = -.146: "As poverty increases by one unit, the children's BPI score *decreases* by .146 of a unit, controlling for other variables in the model."
  - Country of origin: β = .196: "compared to those born in any other county, children of those parents born in Latin American countries show a .196 increase in BPI scores (i.e. more likely to have emotional problems), controlling for…"
- Looking at the output, which IV seems to account for the most change in the DV?

Finally, the exciting conclusion--Hypothesis testing of the coefficients !

- The *t* statistic tests the Null that the individual B coefficient = 0
  - H_A : The individual B coefficient is not equal to 0 (meaning that the coefficient shows a linear relationship between this IV and the DV, controlling for other variables)
- Just as with other statistical tests, the tiny *p* value (equal to or lower than our alpha) means we reject the Null

Reporting Results

- Adjusted R^2 = 0.24
- *F* statistic:
  $F(df_{residual}, df_{regression})$ = *F* value, *p* = *p value*
  in our example: $F(3,26) = 10.192$, $p = <.001$
- Regression coefficient (standardized):
  β = beta value, *p* = *p value*
  in our example, for "poverty":
  $\beta = -0.146$, $p = .025$

(Although APA says to report t test results, we will not do that here)