SAN JOSÉ STATE UNIVERSITY

College of Social Work
S. W. 242
Spring 2008
Edward Cohen

Week 12, 4/11/08 &
Week 13, 4/18/08

Correlation and Multiple Linear Regression

I. What is **correlation?**

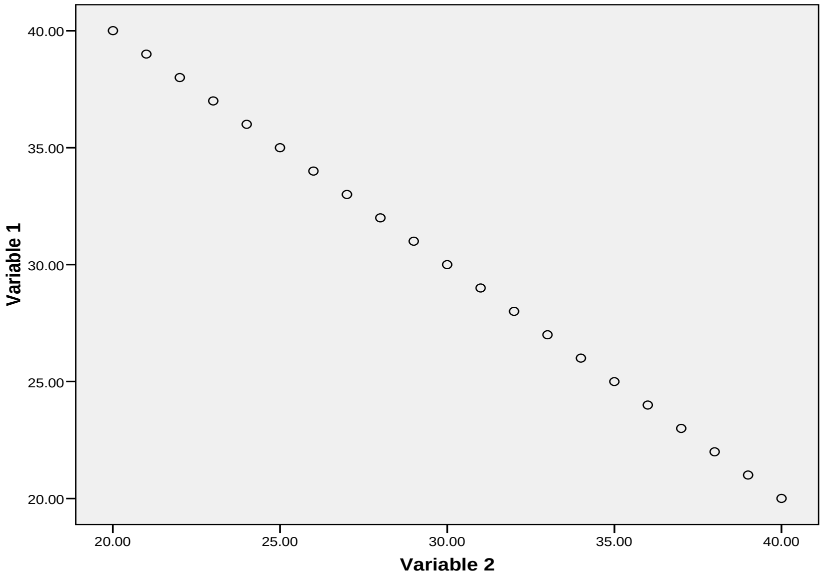
- Correlation tests the relationship between a continuous independent variable and a continuous dependent variable.

- Correlation tests produce an **r value** and a **p value**.

- The **r value** is always between -1 and +1, and indicates the degree to which the two variables are related.

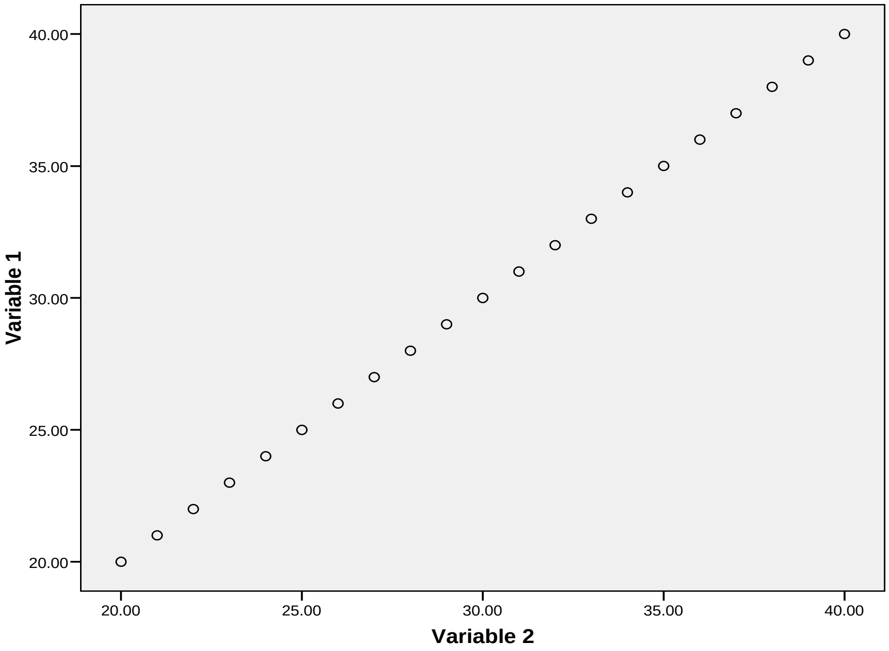
- A **negative r value** indicates that as the value of one variable increases, the value of the other variable decreases (referred to as a negative correlation)

Example of a graph of a negative correlation



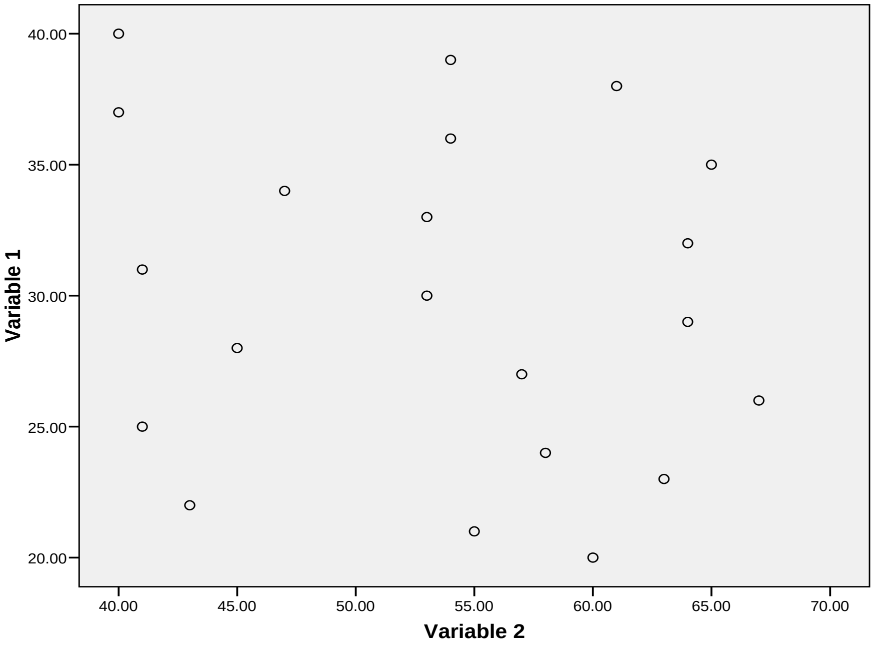
- A **positive r value** indicates that as one variable increases, the other variable also increases (referred to as a positive correlation)

Example of a graph of a positive correlation



- An **r value of zero** indicates no relationship between variables

Example of a graph indicating no correlation between variables



Research Scenario: Correlation

In the general child population, a number of contextual factors have been linked to emotional problems in children and youth, including:

1) low income,

2) negative parenting behavior (i.e. hostile or coercive parenting),

3) family conflict (including family violence and verbal abuse), and

4) low self-efficacy of the child's primary caregiver (i.e. the extent to which the primary caregiver feels mastery over her/his life)

Although the relationship between 1) income, 2) parenting, 3) family conflict, and 4) primary caregiver self-efficacy and children's emotional problems has been established in the general child population, much less is known about how these four factors might contribute to emotional problems among children in immigrant families.

You are interested in testing the relationship between income, parenting behavior, family conflict and primary caregiver self-efficacy among a randomly selected sample of 379 children of immigrant parents in Los Angeles County. Within each family one primary caregiver and one child is interviewed with a structured interview format. This sample includes children ages 9 to 15. **First, we will look at the separate bivariate relationships between each independent variable and the dependent variable. Then we will include <u>all</u> variables in a simultaneous multivariate model (see 12, "Multivariate Statistics").**

<u>Bivariate Analyses</u>

1) Identify the <u>**independent variables**</u> and <u>**their levels of measurement**</u>

   There are four continuous independent variables in this research scenario:

   1) **Independent Variable: <u>Poverty</u>**, measured with income-to-needs ratio, which is a ratio that takes the total family income and divides it by a poverty threshold determined by the federal government. An income-to-needs ratio of 1 means that the family's income is exactly proportional to

the family's financial needs. An income-to-needs ratio of less then 1 indicates more need than income (i.e. poverty) and an income-to- needs ratio of more than 1 indicates more income than need (i.e. not in poverty). This is a continuous variable.

2) **Independent Variable: Parenting behavior**, measured with the Home Observation and Measurement of the Environment (HOME) inventory, which includes a series of self-report questions that the parent answers in relation to their own parenting behaviors. Questions are focused on frequency and type of emotional support and cognitive stimulation provided to the child. This is a continuous variable.

3) **Independent Variable: Family conflict**, measured with Family Conflict scale, which includes a series of self-report questions that the child answers in relation to the ways in which the family communicates with one another and solves problems. This is a continuous variable.

4) **Independent Variable: Primary Caregiver (PCG) Self-Efficacy**, measured with the Pearlin Self-Efficacy Scale, which includes a series of self-report questions that the parent answers in relation to their own level of self-efficacy (e.g. feeling that one has control over one's life). This is a continuous scale.

2) Identify the **dependent variable** and **level of measurement**

The **dependent variable is children's emotional problems**, which is measured with the Behavior Problem Index, internalizing sub-scale, which includes a series of self-report questions that the parent answers in relation to the emotional well-being of their child (e.g. depression, anxiety, withdrawal etc). This is a continuous scale.

3) State the **null hypotheses** (there will be four, since there are four independent variables)

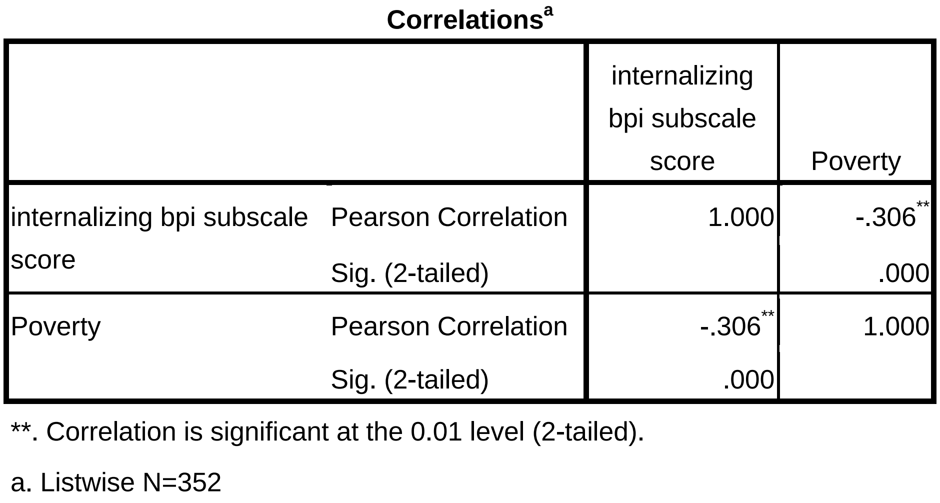
4) State the **alternative hypotheses** (there will be four, since there are four independent variables)

5) Why is **correlation** the appropriate statistical test in this scenario?

6) **Results (**SPSS output) (alpha is set at .05)

Independent Variable: **Poverty**

Dependent Variable: Children’s emotional problems (BPI internalizing subscale score)

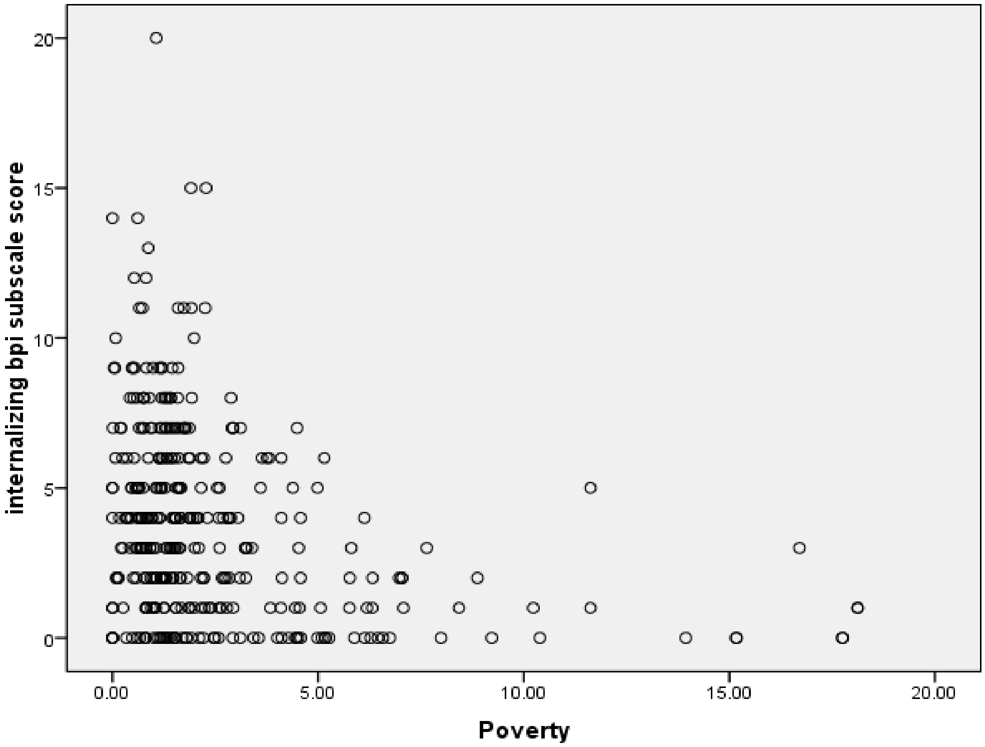


Correlations[a]

		internalizing bpi subscale score	Poverty
internalizing bpi subscale score	Pearson Correlation	1.000	-.306**
	Sig. (2-tailed)		.000
Poverty	Pearson Correlation	-.306**	1.000
	Sig. (2-tailed)	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=352



Independent Variable: **Parenting behavior**

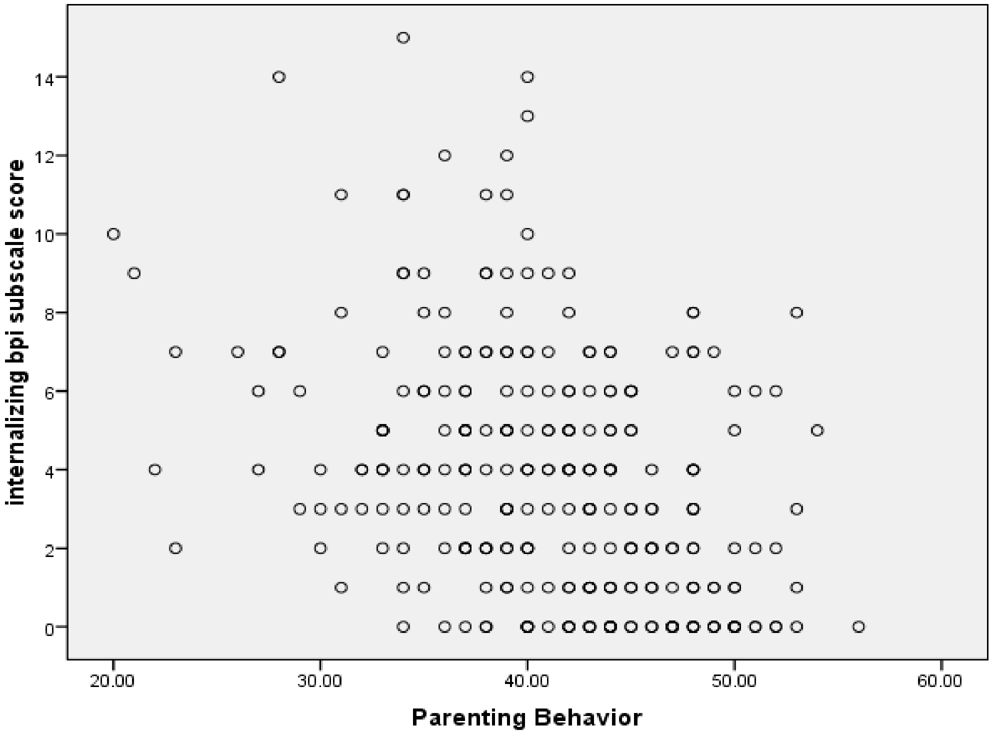
Dependent Variable: Children's emotional problems (BPI internalizing subscale score)

Correlations[a]

		internalizing bpi subscale score	Parenting Behavior
internalizing bpi subscale score	Pearson Correlation	1.000	-.382**
	Sig. (2-tailed)		.000
Parenting Behavior	Pearson Correlation	-.382**	1.000
	Sig. (2-tailed)	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=280



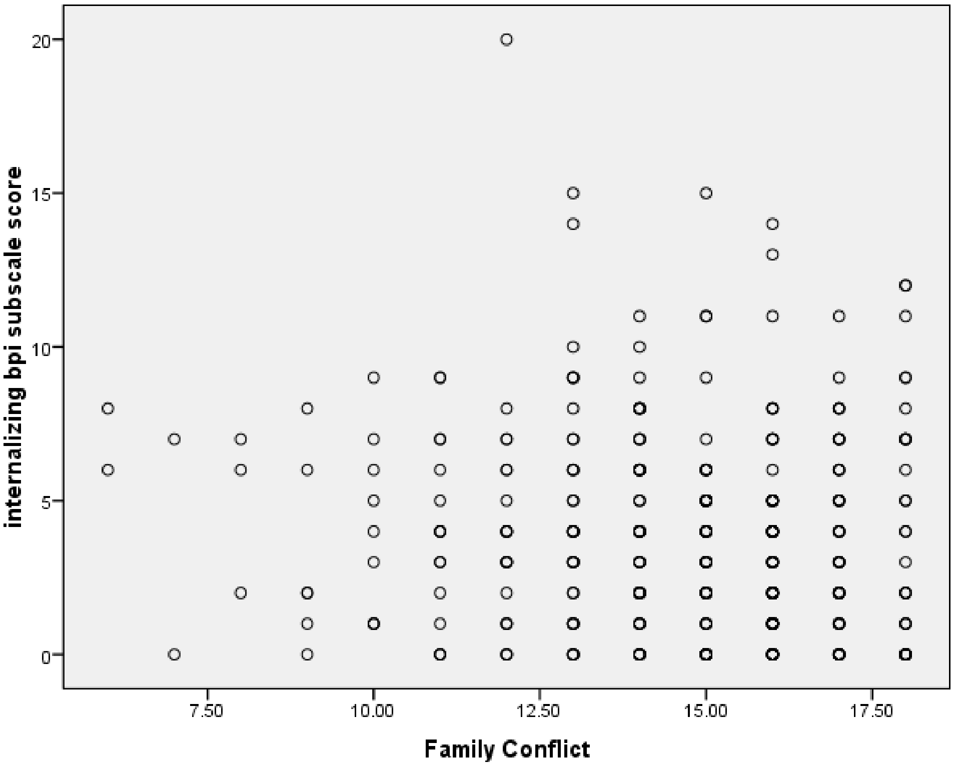
Independent Variable: **Family Conflict**

Dependent Variable: Children’s emotional problems (BPI internalizing subscale score)

Correlations[a]

		internalizing bpi subscale score	Family Conflict
internalizing bpi subscale score	Pearson Correlation	1.000	-.091
	Sig. (2-tailed)		.090
Family Conflict	Pearson Correlation	-.091	1.000
	Sig. (2-tailed)	.090	

a. Listwise N=346



Independent Variable: **Primary Caregiver Self-Efficacy**

Dependent Variable: Children's emotional problems (BPI internalizing subscale score)

Correlations[a]

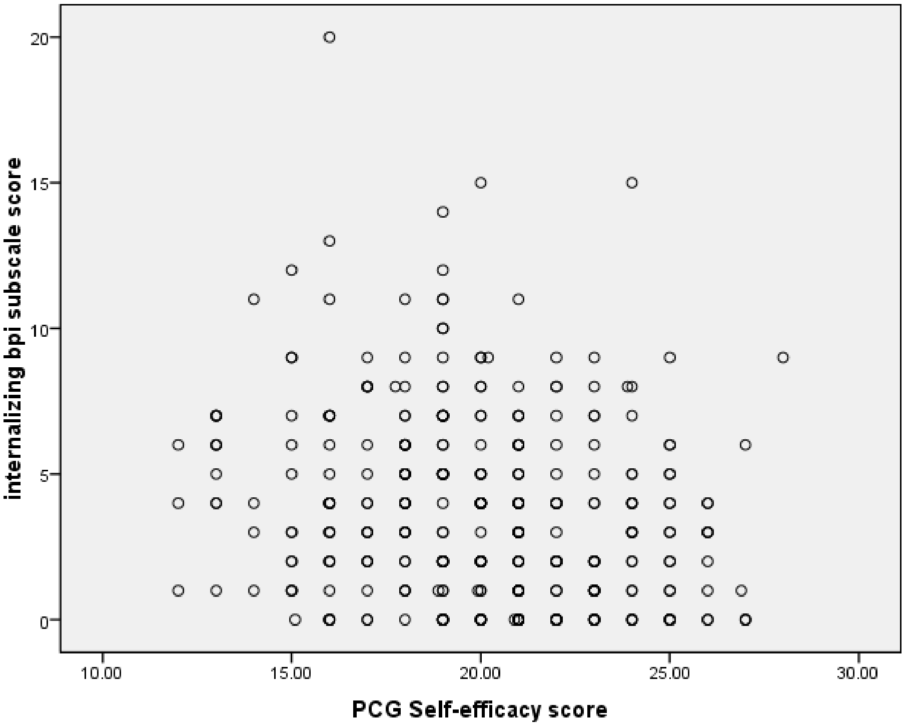
		internalizing bpi subscale score	PCG Self-efficacy score
internalizing bpi subscale score	Pearson Correlation	1.000	-.222**
	Sig. (2-tailed)		.000
PCG Self-efficacy score	Pearson Correlation	-.222**	1.000
	Sig. (2-tailed)	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations[a]

		internalizing bpi subscale score	PCG Self-efficacy score
internalizing bpi subscale score	Pearson Correlation	1.000	-.222**
	Sig. (2-tailed)		.000
PCG Self-efficacy score	Pearson Correlation	-.222**	1.000
	Sig. (2-tailed)	.000	

a. Listwise N=350



7) Describe **results** and **decision to accept or reject the null hypotheses (use APA)**

1) For the independent variable of **poverty:** *r*(352) = -. 306, *p* < .001. Since the p value of *p* < .001 is less than the alpha of .05, we can reject the null hypothesis and conclude that there is a negative relationship between the income/needs ratio and

children's emotional problems—as the income/needs ratio decreases, children's emotional problems increase (this is a negative correlation, we can tell by the negative sign on the r value).

2) For the independent variable of **parenting behavior**: *r*(280) = -.382, *p* < .001. Since the p value of *p* < .001 is less than the alpha of .05, we can reject the null hypothesis and conclude that there is a negative relationship between parenting behavior—as positive parenting behavior decreases, children's emotional problems increase (this is a negative correlation, because of the negative sign on the r value).

3) For the independent variable of **family conflict**: *r*(346) = -.091, *p* = 090. Since the p value of p = .090 is greater than the alpha of .05, we cannot reject the null hypothesis and so we conclude there is not a relationship between family conflict and children's emotional problems

4) For the independent variable of **primary caregiver self-efficacy**: *r*(350) = -.222, *p* <. 001. Since the p vale of *p* < .001 is less than the alpha of .05, we can reject the null hypothesis and conclude that there is a negative relationship between primary caregiver self-efficacy and children's emotional problems—as primary caregiver self-efficacy decreases, children's emotional problems increase (this is a negative correlation, because of the negative sign on the r value.

8) Provide a **discussion of these results**

Results of the Pearson correlation indicated a statistically significant relationship between children's emotional problems and income, parenting behavior and primary caregiver self-efficacy. Specifically the lower a family's income to needs ratio, the higher the level of emotional problems for the child; the lower a parent's level of parenting behavior (i.e. low levels of emotional support and cognitive stimulation), the higher the level of emotional problems for the child; and the lower the level of primary caregiver self-efficacy, the higher the level of emotional problems for the child. There was no significant correlation between family conflict and children's emotional problems.

What can we say about:

- Meaning and implications

- Limitations/areas for future research

II. **Multivariate Statistics**

a. What are multivariate statistics?

- Multivariate statistics allow you to determine the impact of an independent variable on a dependent variable while factoring out the influence of potentially confounding (i.e. extraneous) variables.

b. Types of multivariate statistics:

1. **Multiple linear regression**

- The **dependent variable must be continuous**

- The independent variables (or control variables) may be continuous or categorical

2. **Binomial logistic regression**

- The **dependent variable must be categorical with only 2 categories**

- The independent variables (or control variables) may be continuous or categorical

c. **Dummy coding** in multiple linear regression and binomial logistic regression:

- If an independent/control variable is categorical, then **dummy coding** (AKA creating indicator variables) is necessary for proper analysis with multiple regression. This involves creating a separate variable for each category within the categorical variable and using a "baseline" category with which to compare all other categories.

- For instance, race/ethnicity is a very common demographic variable that is included in many multivariate statistical models. We normally think of race/ethnicity as one categorical (nominal) variable with multiple categories within it (i.e. White, African American, Latino, Asian/Pacific Islander, Other). However, to include race/ethnicity in a multivariate model, we need to use a procedure called dummy coding (AKA creating indicator variables) to create several *dichotomous* variables. To do this, the one variable of race/ethnicity is re-coded (in SPSS) into 4 indicator variables:

1. **White: Value labels: 0 = Not White 1 = White**
2. **African American: Value Labels: 0 = Not African American 1 = African American**
3. **Latino: Value labels: 0 = Not Latino, 1 = Latino**
4. **Asian/Pacific Islander: Value labels: 0 = Not API, 1 = API**

Why four and not five? Because we don’t need to create dummy variables for all five original attributes. The analysis treats the missing dummy variable as a baseline with which to compare all others. (If you did code all five and tried to run the multivariate analysis, your analysis would be in error.)

- One indicator variable is chosen as the “baseline” to which all other racial/ethnic categories are then compared. For instance, if White is chosen as the baseline, then the statistical output provided by SPSS will indicate a comparison between African Americans and Whites, Latinos and Whites, and APIs and Whites with respect to the dependent variable. How do you chose which variable to exclude as a dummy variable, hence using it as a baseline variable? The decision is based on a combination of theory and standard research practice—often an “Other” category is typically as the baseline. However, sometimes you are interested in comparing all others to White (or another ethnicity), so that variable would be used as the baseline.
- You can also create a dummy variables from “Gender”, where Female = 1 and Male = 2. You would recode “Gender” as a dichotomous dummy variable called “Female.gender” where 0=not Female and 1=Female. “Male” would be the baseline, and the analysis would then compare females to males. We use “0” and “1”

since the interpretation of the results focuses on having either a presence or absence of the variable.

Here’s what it looks like for ethnicity, using “Other” as a baseline. Compare these two alternate ways of coding “ethnicity”:

Data set with one ethnicity variable (having five attributes or categories):

Subject.ID	*Ethnicity* 1=White 2=Latino 3=Afr Amer 4=Asian/PI 5=Other
1	3
2	3
3	1
4	4
5	1
6	2
7	2
8	5
9	2
10	2

Recoded data set with *four* ethnicity indicator (dummy) variables (and one, the “Other” is excluded):

Subject.ID	*White* 0=non-White 1=White	*Latino* 0=non-Latino 1=Latino	*Afr.amer* 0=non-African American 1=African American	*Asian.pi* 0=non-Asian/PI 1=Asian/PI
1	0	0	1	0
2	0	0	1	0
3	1	0	0	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	0	1	0	0
8	0	0	0	0
9	0	1	0	0
10	0	1	0	0

***Is “Other” really *missing*? Hint—look at subject ID 8.

d. **<u>Sample size requirements</u>** for multivariate statistics

- General rule of thumb is there needs to be at **least 10 people in the sample for every independent or control variable** included in the model.

III. Example of a multiple linear regression

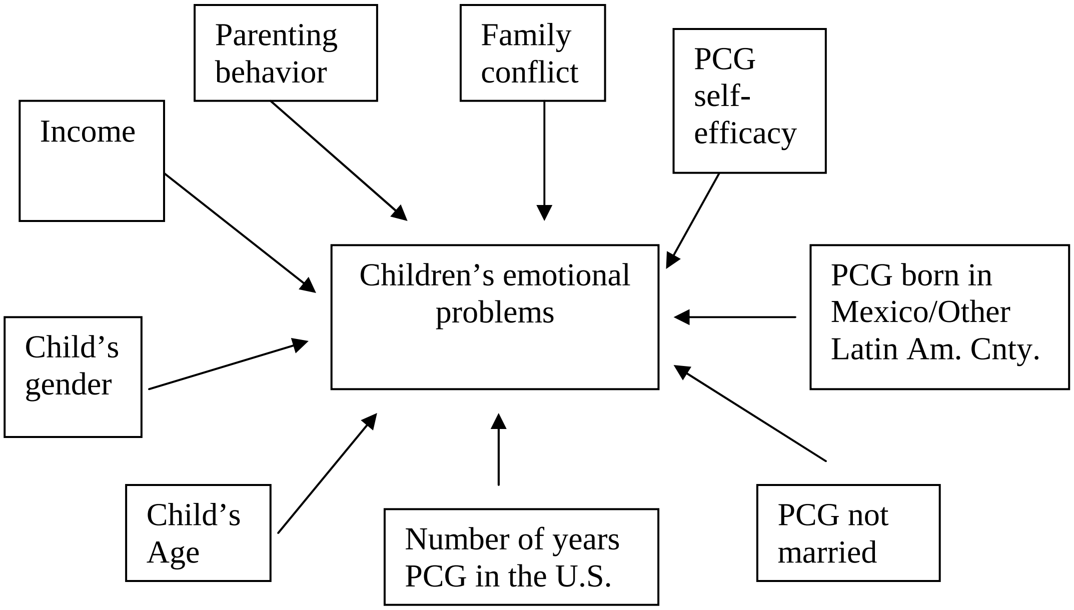
Research Scenario: Multiple Linear Regression

Findings from the correlation analyses in the previous research scenario indicated that income, parenting behavior and primary caregiver self-efficacy are all significantly related to emotional problems among children of immigrant parents. With the exception of the non-significant correlation between family conflict and children’s emotional problems, these findings are similar to findings based on non-immigrant children.

You are interested in finding out if these significant relationships remain after controlling for potentially confounding variables in a multivariate statistical model. You decide to use multivariate statistics and add the following control variables to the model: child age, child gender, number of years primary caregiver has been in the U.S., the marital status of the primary caregiver and the country of origin for the primary caregiver.

***What do we mean by “Model”?

<u>Variables in the model:</u>


Parenting behavior
Family conflict
PCG self-efficacy
Income
Children's emotional problems
PCG born in Mexico/Other Latin Am. Cnty.
Child's gender
Child's Age
Number of years PCG in the U.S.
PCG not married

1. Identify the **<u>independent and control variables</u> and <u>their levels of measurement</u>**

   1) **Independent Variable: <u>Poverty</u>**, measured with income-to-needs ratio, continuous

   2) **Independent Variable: <u>Parenting behavior</u>**, measured with HOME inventory, continuous

   3) **Independent Variable: <u>Family conflict</u>**, measured with Family Conflict scale, continuous

   4) **Independent Variable: <u>Primary Caregiver (PCG) Self-Efficacy</u>**, measured with Pearlin Self-Efficacy Scale, continuous

   5) **Control Variable: <u>Child's gender</u>**, categorical, male as baseline

   6) **Control Variable: <u>Child's age</u>**, continuous

   7) **Control Variable: <u>Number of years PCG has been in the U.S.,</u>** continuous

8) **Control Variable: <u>PCG born in Mexico/Other Latin American country</u>**, categorical, PCG born in any other foreign country is the baseline

9) **Control Variable: <u>PCG unmarried</u>**, categorical, PCG married is the baseline

2. Identify the **<u>dependent variable</u>** and **<u>level of measurement</u>**

Children's emotional problems, measured with Behavior Problem Index, internalizing sub-scale, a continuous scale.

3. State the **<u>null hypotheses</u>**

1. **<u>Null hypothesis for the overall model</u>**: There is no relationship between the combined influence of the independent and control variables and the dependent variable of children's emotional problems.

2. **<u>Null hypothesis for each independent and control variable:</u>**

   a. There is no relationship between **<u>poverty</u>** and children's emotional problems, after controlling for the influence of the other variables in the model.

   b. There is no relationship between **<u>parenting behavior</u>** and children's emotional problems after controlling for the influence of the other variables in the model.

   c. Etc…for each independent and control variable

4. State the **alternative hypotheses**

1. **Alternative hypothesis for the overall model:** There is a relationship between the combined influence of the independent and control variables and the dependent variable of children's emotional problems.

2. **Alternative hypothesis for each independent and control variable:**

- There is a relationship between **poverty** and children's emotional problems, after controlling for the influence of the other variables in the model.

- There is a relationship between **parenting behavior** and children's emotional problems after controlling for the influence of the other variables in the model.

- Etc…for each independent and control variable

5. Why is **multiple linear regression** the appropriate statistical test?

6. **Results** (SPSS output) alpha is .05

When you run a multiple regression in SPSS, you get three tables:

1) **Model Summary** (to get adjusted R Square),

2) **ANOVA** (to get F statistic and p value for overall model),

3) **Coefficients** (to get standardized coefficient beta and the p value) .

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

- The **Model Summary** tells you how well your overall model (with all of the independent variables and control variables combined), predicts or explains the dependent variable

- **Adjusted R Square** is the percent of variance in the dependent variable that is explained by the overall model. This means that when all of the independent and control variables are combined, the adjusted R Square tells us how much these variables explain or are contributing to the dependent variable.

- The **adjusted R Square** is used rather than the R Square because it is adjusted for the fact that we are using a sample and not the total population.

- The larger the **adjusted R square**, the better your overall model.

- In this case, the adjusted R Square is .240, which tells us that approximately 24% of the variance in children's emotional problems is explained by all of the independent and control variables included in the model. This means that 76% of the variance in immigrant children's emotional problems is explained by other variables not included in the model.

- A general rule of thumb is that any Adjusted R-Square over 30% is noteworthy.

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

- The ANOVA table within the multiple regression tests the significance of the overall regression model.

- This means that the F statistic (10.192) and the p value (p<.001), indicate whether all of the independent and control variables combined in the model are significantly related to the dependent variable.

- If the p value in the ANOVA table of a multiple linear regression is non-significant (i.e. the p value is greater than the alpha of .05), then it is unlikely that any of the independent or control variables are significantly related to the dependent variable.

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

- The coefficients table tells you how much each independent variable and control variable is contributing to, or explaining the dependent variable, after controlling for all of the other variables in the model.

- To evaluate the statistical significance of each independent and control variable, we look **at two values in the coefficients table**:

  - **Standardized Coefficient Beta (b-coefficient):** These values are between – 1 and + 1 and reflect the overall strength and direction of the relationship between the independent/control variable and the dependent variable.

- ♦ A **negative value** on the standardized beta coefficient indicates a negative linear relationship (as the independent variable decreases, the dependent variables increases, after controlling for the influence of the other variables in the model).

- ♦ A **positive value** on the standardized beta coefficient indicates a positive linear relationship (as the independent variable increases, the dependent variable increases, after controlling for the influence of the other variables in the model).

\

- o **P value (sig.)** : if lower than alpha (.05), then reject null and conclude there is a relationship between the independent/control variable and the dependent variable after adjusting, or controlling for the influence of the other variables in the model.

7. Describe **results** and **decision to accept or reject the null hypotheses (use APA)**

- o The overall multiple regression was significant (adjusted R-square = .24, F (9, 253) = 10.192, *p* < .001).
  - ♦ Since the p value of *p* < ,001 is less than the alpha of .05 we can reject the null hypothesis and conclude that there is a significant relationship between the combined influence of the independent and control variables on children's emotional problems.

- o Significant variables in the model included:

1) **Poverty** (B = -.146, *t*(263) = -2.25, *p* = .025). As the income to needs ratio decreases (we can tell the direction of the relationship by the negative sign on the standardized beta coefficient), emotional problems increase after controlling for other variables in the model.

We can reject the null hypothesis because the p value of .025 is less than the alpha of .05.

2) **Parenting behavior** (B = -.230, *t*(263) = -3.92, *p* <. 001). As the quality of parenting behavior decreases (negative sign on the standardized beta coefficient), children's emotional problems increase, after controlling for other variables in the model. We can reject the null hypothesis because the p value of *p* < .001 is less than the alpha of .05.

3) **PCG self-efficacy** (B = -.149, *t*(263) = -2.63, *p* = .009). As PCG self-efficacy decreases (negative sign on the standardized beta coefficient), children's emotional problems increase, after controlling for other variables in the model. We can reject the null hypothesis because the p value of .009 is less than the alpha of .05.

4) **PCG born in Mexico or other Latin American Country** (B = .196, *t*(263) = 3.07, *p* = .002). PCGs who were born in Mexico or other Latin American Countries were more likely to have children with emotional problems than PCGs who were born in another foreign country (we know they are more likely to have emotional problems because of the positive sign of the standardized beta coefficient). We can reject the null hypothesis because the p value of .002 is less than the alpha of .05.

8. Provide a **discussion of these results**

Results of the multiple linear regression model indicated four variables were significantly related to emotional problems among children of immigrant parents, while controlling for the influence of other variables. Specifically, children with higher levels of emotional problems tend to: live in families with low income to needs ratios; have parents who demonstrate parenting behaviors characterized by low levels of emotional support and cognitive stimulation; and have a primary caregiver with low self-efficacy, even after controlling for the influence of child age, child gender, primary caregiver marital status, length of time primary caregiver, family conflict, country of origin, and the other independent variables.

Interestingly, country of origin was also found to be significantly related to children's emotional problems; children whose primary caregiver was Latino (e.g. from Mexico or other Latin/Central American Country) had higher levels of emotional problems than children whose primary caregiver was from a different foreign country.

What can we say about:

- Meaning and implications of these results

- Limitations and areas for future research