

SAN JOSÉ STATE UNIVERSITY

College of Social Work
S. W. 242
Spring 2009
Edward Cohen

Week 11

April 3, 2009

- **Review of statistical tests so far**
- **Labs 2b & 3**
- **Bivariate and Multivariate Linear Regression**
- **Review of Analysis Plan for Final Paper**

Concepts you should know:

- Bivariate linear regression
- Multivariate statistics
- Multivariate (or multiple) linear regression
- Multivariate model
- Control variables
- R^2 and adjusted R^2
- Analysis of variance in linear regression
- F ratio in linear regression
- Regression coefficient (B coefficient)
- Standardized coefficient (Beta coefficient)
- t -test in linear regression

I. Review—what statistical procedure would answer the research question?

A. Does a group case management intervention result in fewer overall costs than an individual case management intervention?

B. For decisions about those eligible vs. not eligible for SSI (Supplemental Social Security Income), does ethnicity of applicant matter?

C. 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?

D. Does the change in a living skills scale score vary by type of treatment: individual therapy, individual therapy with case management, or case management without therapy?

E. 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](#)

II. Bivariate and Multivariate (or multiple) linear regression

A. What is “regression”?

Regression is a form of correlation analysis; it can be either bivariate or multivariate.

- Bivariate regression predicts the value of a dependent (or outcome) variable from an observed independent (or predictor) variable. The dependent variable must be continuous. The independent variable is most often continuous. (Bivariate regression is not used much—instead use correlation for two continuous variables. Use *t*-tests or ANOVA if the IV is categorical.)
- Multivariate (or multiple) regression predicts the value of a dependent (or outcome) variable from *two or more* observed independent (or predictor) variables. The independent variables (or control variables) may be continuous or categorical.

B. You can say ““The IV is predictive of the DV, *controlling for other IVs*”

C. 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”
5. So you can say about *any* independent variable, “controlling for the effects of all other variables...”

D. Bivariate example

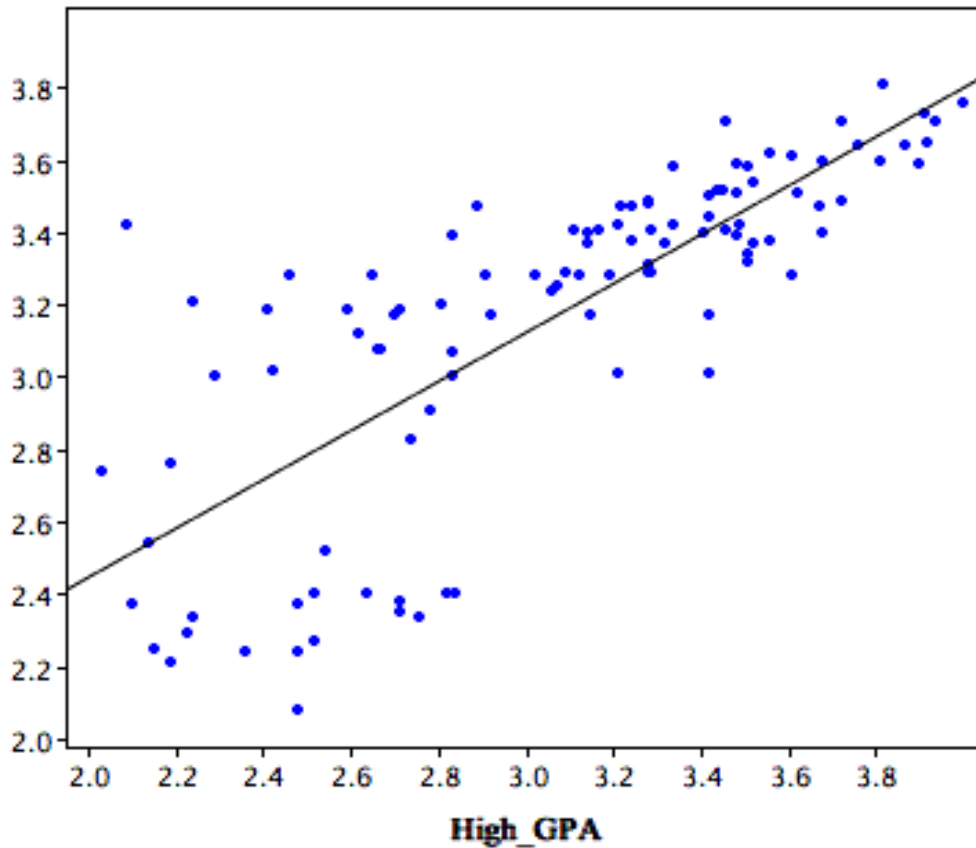
1. Null and Alternative Hypotheses

- 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.
- H_0 (Null): There is no relationship between strength-based comments and behavioral incidents in children

2. Linear regression – the best fitting line (or, “least squares regression”)

The following shows a plot of high school students’ GPA scores (independent variable) against subsequent university GPA scores. The dots are the observed points of intersection between the two variables, for each student.

Univ_GPA



The line results from “least squares regression” – using a calculus formula to find the line that *on average has the least amount of distance from the observed values to the predicted ones (those on the line)*.

That results in an equation that describes the line—its beginning point and slope:

$$Y = c + bX$$

Y = dependent variable

X = independent variable

b = slope of the line or the regression coefficient of the X variable (literally, “b times X”); *the change in Y for a one-unit change in X*.

c = constant (or the point of intercept at the Y axis) (sometimes you’ll see “a” used as the symbol)

For the GPA example:

$$\text{University GPA} = 1.097 + (0.675)(\text{High School GPA})$$

This equation can be used to predict a new high school student’s university GPA score.

E. Example of a multiple linear regression

1. 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...
2. H_A : The number of strength-based comments has a larger impact on behavioral incidents in children than seriousness of diagnosis, gender and length of time in care
3. H_0 (Null): There is no relationship between strength-based comments and behavioral incidents in children

The multivariate formula notation is:

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

Y = dependent variable

X = independent variables

b = regression coefficient for each X

c = constant

k = number of independent variables

F. Typical uses of multiple regression--while regression can be used to predict outcomes, the procedure is most often used to:

1. Determine whether there is a relationship between a primary IV and DV, controlling for other variables
 2. Determine the strength and direction of the relationship between a primary IV and DV, while controlling for other variables
 3. Determine the effects of other independent variables (such as control variables) to the DV, and which variables have greater effect
- Note: the rules of causality still apply. It's the research design that supports causality, not the statistic.

G. Research Scenario: Multiple Linear Regression

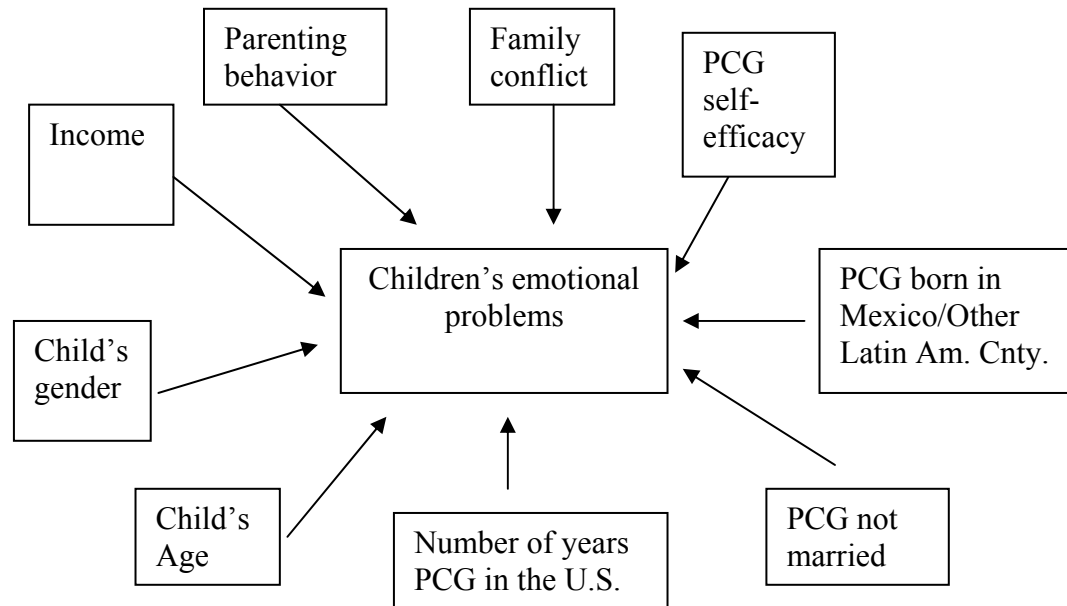
Findings from the correlation analyses in this scenario (discussed last class) indicated that income, parenting behavior and primary caregiver self-efficacy are all significantly related to emotional problems among children of immigrant parents. With the exception of the non-significant correlation between 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. (Confounding variables are those that could also affect the DV—they “confound” the relationship between the primary IV and DV.) 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”?

Variables in the multivariate model:



Eight Steps to Hypothesis Testing:

1. Identify the independent and control variables and their levels of measurement

Variable Name	As measured by...	Level of measurement
Income/need (indicator of poverty)	Income-to-needs ratio (total family income divided by federal poverty threshold). A ratio of 1 means family income is exactly proportional to family's financial needs (by federal standards). A higher ratio indicates higher financial resources, less need	Continuous
Parenting behavior	Home Observation and Measurement of Environment (HOME) Inventory. The	Continuous

Variable Name	As measured by...	Level of measurement
	higher the score, the more positive the parenting behavior	
Family conflict	Family Conflict Scale—child response—questions about how family communicates and solves problems. The lower the score, the lower the perceived amount of conflict	Continuous
Primary caregiver (PCG) self efficacy	Pearlin Self-Efficacy Scale—parent response to questions about self efficacy (feeling that one has control over one's life). Higher scores indicate higher self-efficacy	Continuous
Marital status	Marital status of PCG (married or unmarried)	Categorical (nominal)
Child's gender	Male or female	Categorical (nominal)
Child's age	Age in years	Continuous
Years in US	Number of years primary caregiver in US	Continuous
Birthplace of PCG	Birthplace of primary caregiver (Mexico vs. other Latin American country)	Categorical (nominal)

2. Identify the dependent variable and level of measurement

Behavioral Problem Index (Internalizing subscale) -- the higher the score, the more severe the emotional problems; continuous level of measurement

3. State the null hypotheses

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

Null hypothesis for each independent and control variable:

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

There is no 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

4. State the alternative hypotheses

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.

Alternative hypothesis for each independent and control variable:

There is a negative relationship between income/need and children's emotional problems, after controlling for the influence of the other variables in the model.

There is a relationship negative 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?

The dependent variable is continuous, and there two or more independent variables.

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

When you run a multiple linear regression in SPSS (Analyze→Regression→Linear), 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

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. (The adjusted R^2 is the one reported. The unadjusted R^2 is sensitive to the number of variables in the model; just adding new variables will increase it. So, R^2 has to be adjusted accordingly.)

The larger the adjusted R square, the better your overall model. (In many studies one group of independent variables will be entered first, than another group of variables added in a separate regression to see the change in the adjusted R^2 .)

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. We won't be too concerned about the size of the adjusted R^2 .

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*.

The F ratio in linear regression is the Regression Mean Square divided by the Residual (or error) mean square. "Sums of squares" are calculated for all observed values, and all "error" values (those distances between the plotted points and the estimated regression line.) "Mean square" is the sums of squares divided by the corresponding degrees of freedom.

The point of the F ratio is to determine the overall importance of the regression line values, compared to the error values. For the statistical test, the null is "the effect of the regression model (the independent variables as a group) is no better than what can be achieved by chance." Or, you can say "the regression model is not a significant predictor of the dependent variable."

Here, the F statistic (10.192) and the p value ($p < .001$), indicate that the model (independent variables as a group) is 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 the model is not significantly related to the dependent variable. Further testing of other models (e.g. by deleting some of the variables from the original model) would then be required.

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. If you were to re-run the regression after deleting one variable, the other coefficients would likely change.

There are two types of regression coefficients—Unstandardized, and Standardized.

Interpreting unstandardized coefficients: *The regression coefficient (B) indicates 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.*

A negative value on the B coefficient indicates a negative linear relationship (as one variable increases, the other variable decreases, after controlling for the influence of the other variables in the model). A positive coefficient indicates a positive relationship between the IV and DV.

Applied to our example:

- *Income/need: B = -.144:*
“As income increases by one unit, the children’s BPI score *decreases* by .144 of a unit, controlling for other variables in the model.”
- *Country of origin: B = 1.475:*
“Compared to those born in any other county, children of those parents born in Latin American countries show a 1.475 increase in BPI scores (i.e. more likely to have emotional problems), controlling for other variables in the model.”

Standardized Coefficient Beta (β -coefficient): These values are between -1 and $+1$ and are the regression coefficients you would get if you standardized all the variables, then ran the regression.

The standardized beta (β) is in standard deviation units (like z scores), so that all β s can be compared to each other (e.g. the sizes of the coefficients can be compared to determine which ones account for the most variance in the dependent variable)

The β is interpreted as “a one standard deviation change in the independent variable results in a (β value) change in the DV.”

The standardized β is reported in the Results section (see below), though usually not interpreted, except to compare the relative weights of the variable effects on the DV.

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)

How we report the R-square, F test and coefficient results:

The overall regression model significantly explains the variance of children’s behavioral problems (adjusted R-square = .24, $F(9, 253) = 10.192, p < .001$). Since $p < .001$ 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.

Coefficient results support the hypothesized negative relationship between the independent variables income/need index ($\beta = -.146, p = .025$), parenting behavior ($\beta = -.230, p < .001$), primary caregiver self-efficacy ($\beta = -.149, p = .009$), and the dependent variable BPI internalizing score. Parenting behavior seems to have the largest influence on reducing children’s behavioral problems. The children of primary caregivers born in Mexico or other Latin American country were more likely than those born in other foreign countries to have higher BPI scores (worse behavior) ($\beta = .196, p = .002$), controlling for other variables. Hypothesized effects of family conflict and the number of years the primary caregiver was in the U.S. were not

supported. Also, gender and age of child, and primary caregiver's marital status were not related to children's behavioral problems though served as control variables in the model.

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 of study and further research...