

Weeks 12  
April 17, 2009

- Review for test
- Dummy variables in multivariate regression
- Lab 4: Multiple Regression
- Small group work on Analysis Plan
- Writing Abstracts

I. Review for test

A. Choosing Statistical Tests

1. Know when to use statistical procedures--in the abstract, e.g. knowing level of measurement and number of IVs, DVs; and for specific research scenarios
2. Test will cover these:
  - a. Dependent (or Paired) groups  $t$  test
  - b. Independent groups  $t$  test
  - c. Multivariate regression (but not logistic regression)

B. Hypothesis Testing--as in the first exam, you will be asked to go through the 8 steps of hypothesis testing for a given scenario

1. What is (are) IV(s)? Type of measure
2. What is DV? Type of measure
3. Identify null hypothesis
4. Identify alternate hypotheses
5. Choose statistical test and alpha
6. Interpret SPSS output
7. Justify decision to reject or not reject null hypothesis, and other Results
  - Know how to interpret both significant *and* non-significant findings
  - For  $t$  tests: interpret differences in means to support research hypothesis
  - Find significant regression coefficients; support research hypotheses?

- In regression, identify IV with largest impact using the Standardized Beta coefficients
- Understand how to interpret coefficients for dummy variables

## 8. Interpret results

- Did findings answer research question(s)?
- Study limitations
  - Research design
  - Sampling strategy and representativeness
- Implications of findings for social work practice and/or policy
- Suggestions for further research

## II. Dummy Variables in Multivariate Linear and Logistic Regression

***Dummy variables are transformed nominal or ordinal variables whose attributes are coded into dichotomous variables.*** A dummy variable is dichotomous, e.g. the variable named “Latino” has only two attributes: 1=Latino; 0=Not Latino.

If an independent/control variable is categorical (either nominal or ordinal), then dummy coding is necessary for proper analysis with multiple regression. This involves creating a separate variable for each category (attribute) of the categorical variable and using a “baseline” category with which to compare all other categories.

Why do we need dummy variables? Consider “ethnicity”—if coded 1=White, 2=AA, 3=Latino, etc., then the regression formula sees this as a continuous variable (1, 2, 3, 4, etc.), which is not accurate. A “one unit change in ethnicity” makes no sense, since ethnicity is really a categorical (nominal) variable.

If the original variable has  $k$  attributes, you create  $(k - 1)$  dummy variables. Why  $k - 1$ ? Because we don’t need to create dummy variables for all of the original attributes. Doing this would create redundant information, because if you know that out of four attributes, all cases for three new dummy variables are coded either “1” or “0”, then you would know (and the regression formula would know) which cases are coded “1” or “0” for the fourth dummy variable. The multivariate analysis treats the missing dummy variable as a baseline with which to compare all others. (If you did code all  $k$  attributes and tried to run the multivariate analysis, your analysis would be in error.)

For instance, race/ethnicity is a very common demographic variable that is included in many multivariate statistical models as a control variable, if not the primary independent variable. 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). Here’s how it would look in a data file:

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

<i>Subject.ID</i>	<i>Ethnicity</i>
	1=White 2=Latino 3=African American 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

However, to include race/ethnicity in a multivariate model, we need to create dummy variables for  $k - 1$  attributes ( $5 - 1 = 4$ ).

<b>Variable Name</b>	<b>Coding</b>
White	0=Not White 1=White
African American	0=Not African American 1=African American
Latino	0=Not Latino 1=Latino
Asian/PI	0=Not Asian/PI 1=Asian/PI
<b>Other</b>	<b>Variable not created</b>

One indicator variable is chosen as the “baseline” to which all other racial/ethnic categories are then compared. We use “0” and “1” since the interpretation of the results focuses on having either a presence or absence of the variable. Here, we re-coded the dummy variables using the excluded variable (“Other”) as the baseline. In the analysis, coefficients for these dummy variables would be in comparison to “Other”. In another example, if White is chosen as the baseline, then the regression coefficients provided by SPSS would be interpreted as a comparison

between African Americans and Whites, Latinos and Whites, and APIs and Whites with respect to the dependent variable.

How do you choose 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 used 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. This goes back to your original research questions, which should be guided by the latest evidence (literature review), a theoretical basis, and culturally appropriate research methods.

The recoded ethnicity data set looks like this:

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

<i>Subject.ID</i>	<i>White</i> 0=non-White 1=White	<i>Latino</i> 0=non-Latino 1=Latino	<i>Afr.amer</i> 0=non-African American 1=African American	<i>Asian.pi</i> 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

\*\*\*The “Other” category is not really *missing*. Look at subject ID 8.

If “Gender” is one of your IVs, you would also create a dummy variable from the original variable Female = 1 and Male = 2. You would recode “Gender” as a dichotomous dummy variable called “Female” where 1=Female and 0=Not female. “Male” would be the baseline, and the analysis would then compare females to males. Alternatively, you could instead have created a variable called “Male” where 1=Male and 0=Not male. It doesn’t matter as long as you set it up right in SPSS so that you can make sense of the output (and as long as you don’t end up with two variables—“Female” and “Male”!)

What does this mean for interpreting dummy variable regression coefficients? Just like all other independent variables in a multivariate model, each dummy variable will have its own coefficient in the multiple regression output.

Example: Let’s say you were interested in the effects of treatment for depression (the lower the depression score the lower the depression, and the dependent variable is the change in depression

from pre- to post-), controlling for ethnicity (Latino, Asian, African American and White) and comparing a treatment and control group. You create  $k - 1$  dummy variables for ethnicity—one each for Latino, Asian, and African American. The White category is your baseline. You would have regression (B) coefficients for three dummy variables: Latino, Asian, and African American. Your “treatment group” variable would also be a dummy variable (1=experimental, 0=control). Let’s say the coefficient for Latino is  $-.80$  and it’s statistically significant. You would interpret the coefficient as “for Latinos in the sample, depression scores were reduced by  $.80$  compared to the depression score of Whites (sampled Latinos’ depression score was lower than that of Whites).” Saying “for Latinos in the sample...” is the same thing as saying “a one unit change in the Latino variable...”

Let’s say the “treatment group” coefficient (B) is  $-.75$  and it’s statistically significant. The interpretation is “The treatment group, compared to the control group, had a  $.75$  reduction in depression scores controlling for ethnicity.”

\*\*\*Awkward language alert: Don’t confuse “control group” and “controlling for”.

### III. Final Paper--Writing the Analysis Plan

#### **Example of Proposed Analysis Section**

*See the example Analysis Plan on p. 42-43 of the Rubin, Babbie & Lee Supplement. Your Analysis Plan (for the Final Paper) should have a table like the one on p. 43 to describe all univariate, bivariate and multivariate statistical procedures as they apply for your study. They might not all apply, although all papers (even those with qualitative-only designs) should include a plan to analyze univariate statistics of the sample. As an introduction to the table, your papers should also have a very brief narrative summary, as in:*

#### Analysis Plan

Narrative description to accompany table:

The analysis plan for this study is summarized in Table \_\_\_\_\_. Univariate analyses will be conducted on all independent and dependent variables in order to describe the sample. Bivariate analyses will be conducted to explore the relationship among each independent variable (age, quality of family life, social support, gender, and ethnicity) and the dependent variable (life satisfaction). Multivariate analysis will consist of multiple linear regression of the independent variables on the dependent variable to address extent to which age, quality of family life, and social support are related to and predict life satisfaction, controlling for gender and ethnicity.

For the qualitative component, the narrative text (*or other types of data as relevant to your study*) will be coded for themes related to how participants view life satisfaction and its relationship of any of the independent variables in their own lives, as well as other factors not included in the quantitative analyses.

*You can cut and paste this text for your own purposes. Note that this paragraph summarizes but is not redundant with the table.*

### IV. Writing Abstracts

(see separate handout)