Solving Standard Multiple Regression Problems

We have used regression analysis to examine relationships among interval level dependent variables and interval and dichotomous level independent variables. However, many important independent variables, such as race, religion, marital status, etc., are all nominal level variables. We can extend our use of multiple regression to include these variables by dummy coding nominal level variables. We can also apply dummy-coding to ordinal level variables and avoid any contention about using ordinal level variables in statistics requiring interval level data.

Dummy-coding is done by converting the categories of the nominal level variable into a set of dichotomous variables. The number of dichotomous variables needed to dummy code a variable is one less than the number of categories in the original variable. Thus, if a variable had three categories, we would retain its information using two dummy-coded variables. If a variable had four categories, it would require us to replace it with three dummy-coded variables, and so forth.

The basic idea behind dummy coding is to select one category of the variable as the “reference” category, and compare all other categories to it. A difference in means would be reported as the difference between one of the retained categories and the reference category. For example, if we were studying participation in extracurricular activities by class standing, we would have categories for freshmen, sophomores, juniors, and seniors. If we selected freshmen as the reference category, we would interpret the b coefficients as the difference between freshmen and sophomores, between freshmen and juniors, and between freshmen and seniors. This dummy-coding strategy is referred to as “indicator” coding to distinguish it from “deviation” coding (also called “effects” coding).

Deviation coding lets us compare each category of the independent variable to the average for all categories. Instead of interpreting the b coefficients as the difference between freshmen and sophomores, we would interpret the deviation-coded variable as the difference between sophomores and all students. Deviation coding, like indicator coding, requires us to designate one category as the reference category that is excluded from the regression.

An example of indicator coding is shown in the following table:

<table>
<thead>
<tr>
<th>Categories of Original Variable:</th>
<th>New Variable SophStudents</th>
<th>New Variable JrStudents</th>
<th>New Variable SrStudents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sophomores</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Juniors</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Seniors</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

If a student was in the freshman reference category, they would have zeros for all three of the new variables. If a student was a sophomore, they would have a one for the new variable SophStudents, and a zero for the other new variables. If a student was a junior, they would have a one for the new variable JrStudents, and a zero for the other new variables. If a student was a senior, they would have a one for the new variable SrStudents, and a zero for the other new variables.
Deviation coding differs by assigning a negative one (-1) to the reference category rather than a zero, as shown in the following table:

<table>
<thead>
<tr>
<th>Categories of Original Variable:</th>
<th>New Variable SophStudents</th>
<th>New Variable JrStudents</th>
<th>New Variable SrStudents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>Sophomores</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Juniors</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Seniors</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Interpretation of the relationships based on the two-coding schemes are different because each is making a different comparison. Indicator coding compares each category to the reference category, while deviation coding compares each category to the mean of all categories. Since the comparisons are different, the slope(b) coefficients for the categories will be different under the two schemes.

In some SPSS procedures, dummy-coding is done automatically if you declare the variable to be a factor (general linear models) or a categorical covariate (logistic regression). These procedures offer a limited number of choices over which category is treated as the reference group, usually the last or first category. To choose another category, we would have to recode the variable so that our chosen group had the highest or lowest code number.

We can also use SPSS Recode commands to dummy code any variable and use the results in any procedure that we want.

The script which we will use for this week’s problems does deviation coding for each non-metric variable, including variables that are dichotomous and could be used in the regression analysis without dummy coding. The rationale for dummy-coding dichotomous variables is to make their interpretation consistent with the interpretation for nominal variables, i.e. comparing the category to the mean for both categories.
The Problem in Blackboard

The Statement about Level of Measurement

1. The first statement in the problem asks about level of measurement. Standard multiple regression requires the dependent variable and the metric independent variables be interval level, and the non-metric independent variables be dummy-coded if they are not dichotomous.

2. The only way we would violate the level of measurement would be to use a nominal variable as the dependent variable, or to attempt to dummy-code an interval level variable that was not grouped.
Marking the Statement about Level of Measurement

- The metric dependent variable "occupational prestige score" [prestg80] was interval level, satisfying the requirement for dependent variables.
- "Income" [rincom98] is ordinal level, but the problem calls for treating it as metric, applying the common convention of treating ordinal variables as interval level.
- "Highest academic degree" [degree] is ordinal level and could be treated as metric, but the problem calls for dummy-coding it, so we will satisfy the requirement by treating it as non-metric.
- The non-metric independent variable "race" [race] was nominal level, but will satisfy the requirement for independent variables when dummy coded.

Satisfying the Assumptions of Multiple Regression

- The next four statements identify the strategies that we will follow to attempt to satisfy the assumptions of multiple regression. If we fail to satisfy the requirement for independence of variables, we halt interpretation of the problem.

Once we satisfy the assumptions, we skip the remaining strategies. If we do not satisfy the assumptions with any of the strategies, we will continue to interpret the analysis, but we should mention the failure to satisfy assumptions in our findings.
Using the Script to Evaluate Assumptions

We will use a new script to evaluate regression assumptions for analyses that include metric and non-metric variables.

Assuming that you have downloaded the script from the course web site to My Documents, select the Run Script command from the Utilities menu.

Opening the Script

In the Run Script dialog box, navigate to My Documents and highlight the file named:

SatisfyingRegressionAssumptionsWithMetricAndNonmetricVariables.SBS

With the script file highlighted, click on the Run button to use the script.
Changes in the Script

The elements of the script have been rearranged to support the addition of non-metric variables. Overall, its actions are very similar to the script we used previously.

The script now included a list box for selecting non-metric variables (nominal, dichotomous, and ordinal variables treated as non-metric).

The script will create dummy-coded variables, but the user must specify the reference category by selecting from the drop down list.

Selecting the Dependent Variable

First, highlight the dependent variable `prestg80` in the list of variables in the data editor.

Second, click on the `>>` button to move the selected variable to the Dependent variable text box.

The problem calls for a standard multiple regression of "occupational prestige score" [prestg80] on "income" [rincom98], "highest academic degree" [degree] and "race" [race].
Selecting Non-metric Independent Variables - 1

The problem calls for a standard multiple regression of "occupational prestige score" [prestg80] on "income" [rincom98], "highest academic degree" [degree] and "race" [race]. The instructions tell us to treat "highest academic degree" [degree] and "race" [race] as non-metric variables.

First, highlight the non-metric independent variable degree in the list of variables in the data editor.

Second, click on the >> button to move the selected variable to the Non-metric independent variables text box.

The instructions tell us to dummy code "highest academic degree" [degree] using "0=lt high school" as the reference category.

Click on the drop down list for Reference category and select 0=LT HIGH SCHOOL.

Selecting the Reference Category for the Non-metric Independent Variable
Selecting Non-metric Independent Variables - 2

The problem calls for a standard multiple regression of "occupational prestige score" [prestg80] on "income" [rincom98], "highest academic degree" [degree] and "race" [race]. The instructions tell us to treat "highest academic degree" [degree] and "race" [race] as non-metric variables.

First, highlight the second non-metric independent variable race in the list of variables in the data editor.

Second, click on the >> button to move the selected variable to the Non-metric independent variables text box.

The instructions tell us to dummy code "race" [race] using "3=other" as the reference category.

Click on the drop down list for Reference category and select 3=OTHER.
Selecting Metric Independent Variables

First, highlight the metric independent variable rincom98 in the list of variables in the data editor.

Second, click on the >> button to move the selected variable to the Metric independent variables list box.

The problem calls for a standard multiple regression of "occupational prestige score" [prestg80] on "income" [rincom98], "highest academic degree" [degree] and "race" [race]. The instructions tell us to treat "income" [rincom98] as a metric variable.

NOTE: the reference category only appears when its non-metric independent variable is highlighted. When another list box has the highlight, the reference category is not displayed.

Having included all of the variables mentioned in the problem, we click on the OK button to request the output.
The Variables Included in the Analysis

We can look at the variables listed in the table of Descriptive Statistics to make certain we have included the correct variables.

The list includes four dummy-coded variables for degree, excluding the reference category of Less Than High School.

The list includes two dummy-coded variables for race, excluding the reference category of Other.

Evaluating the Assumption of Independence of Variables

The tolerance values for all of the independent variables are larger than 0.10: "income" [rincom98] (0.747), "survey respondents who had graduated from high school" [degree_1] (0.916), "survey respondents who had a junior college degree" [degree_2] (0.813), "survey respondents who had a bachelor's degree" [degree_3] (0.866), "survey respondents who had a graduate degree" [degree_4] (0.683), "survey respondents who were white" [race_1] (0.818) and "survey respondents who were black" [race_2] (0.873). Multicollinearity is not a problem in this regression analysis.
Evaluating the Assumption of Linearity

Evaluating the Assumption of Homoscedasticity - 1

The homogeneity of error variance is tested with the Breusch-Pagan test. For this analysis, the Breusch-Pagan statistic was $5.379$. The probability of the statistic was $p = 0.039$, which was greater than the alpha level of significance of 0.01. The null hypothesis that "the variance of the residuals is the same for all values of the independent variable" is not rejected. The research hypothesis that "the variance of the residuals is different for some values of the independent variable" is not supported. The assumption of homogeneity of error variance is satisfied.
Evaluating the Assumption of Homoscedasticity - 2

The scatterplot of residuals (vertical axis) against predicted values (horizontal axis) demonstrates the consistent spread of residuals across the predicted values which reinforces the conclusion that the assumption of homoscedasticity is satisfied.

Evaluating the Assumption of Normality - 1

Regression analysis assumes that the errors or residuals are normally distributed. The Shapiro-Wilk test of studentized residuals yielded a statistical value of 0.992, which had a probability of $p = .533$, which was greater than the alpha level for diagnostic tests ($p = .010$). The null hypothesis that "the distribution of the residuals is normally distributed" is not rejected. The research hypothesis that "the distribution of the residuals is not normally distributed" is not supported. The assumption of normality of errors is satisfied.
Evaluating the Assumption of Normality - 2

The q-q normality plot shows the distribution of residuals plotted against the expected frequencies if the distribution is normal. The alignment of the points on the line reinforces the conclusion that the assumption of normality is satisfied.

Regression analysis assumes that the errors (residuals) are independent and there is no serial correlation. No serial correlation implies that the size of the residual for one case has no impact on the size of the residual for the next case. The Durbin-Watson statistic tests for the presence of serial correlation among the residuals. The value of the Durbin-Watson statistic ranges from 0 to 4. As a general rule of thumb, the residuals are not correlated if the Durbin-Watson statistic is approximately 2, and an acceptable range is 1.50 - 2.50. The Durbin-Watson statistic for this problem is 2.00 which falls within the acceptable range. The analysis satisfies the assumption of independence of errors.
Marking the Statement for Satisfying Assumptions

The model satisfied the assumptions of multiple regression without transforming any variables or excluding any survey respondents.

Had we not satisfied the assumptions, we would have used the script to remove extreme outliers and re-evaluated the assumptions.

If that were not successful, we would test transformations for the metric variables, and if that still is not successful, we would remove extreme outliers from the model using transformations.

The Statement for the Sample Size Requirement

The next statement to be evaluated is the requirement for a minimum sample size.
The analysis included 7 independent variables: 1 for the covariate "income" [rincom98] plus 4 dummy-coded variables for the factor "highest academic degree" [degree] and 2 dummy-coded variables for the factor "race" [race]. The number of cases available for the analysis was 167, satisfying the requirement for 112 cases based on the rule of thumb that the required number of cases should be the larger of the number of independent variables x 8 + 50 or the number of independent variables + 105.

Since we satisfied the sample size requirement, we mark the check box.
The Statement for Overall Relationship

Having satisfied level of measurement, assumptions, and sample size requirements, we now interpret the statistical relationships for the model.

First, we will check the correctness of the significance and strength of the overall relationship.

Since the script runs a standard multiple regression when it generates the output for evaluating assumptions, we can use its output to answer the remaining questions.

The relationship between occupational prestige score and the combination of income, highest academic degree, and race was statistically significant ($F(7, 159) = 13.50, p < .001$). The null hypothesis that "all of the partial slopes (b coefficients) = 0" is rejected, supporting the research hypothesis that "at least one of the partial slopes (b coefficients) is not equal to 0".

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>13479.35</td>
<td>7</td>
<td>1924.93</td>
<td>13.40</td>
<td>000*</td>
</tr>
<tr>
<td>Residual</td>
<td>21000.31</td>
<td>159</td>
<td>132.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>33479.66</td>
<td>166</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: [Constant], RESPONDENTS INCOME, RACE OF RESPONDENT-BLACK, RS HIGHEST DEGREE=HIGH SCHOOL, RS HIGHEST DEGREE=BACHELOR, RS HIGHEST DEGREE=JUNIOR COLLEGE, RACE OF RESPONDENT-WHITE, RS HIGHEST DEGREE=GRADUATE

Applying Cohen's criteria for effect size (less than .10 = trivial; .10 up to 0.30 = weak; .30 up to .50 = moderately strong; .50 or greater = strong), the relationship was correctly characterized as strong (Multiple R = .611).

Since the overall relationship was statistically significant and the Multiple R of .611 supported characterization as strong, the check box is marked.
The Statements for Individual Relationships

The next block of statements are interpretations of the relationships between individual predictors and the dependent variable.

A statement is correct if it is statistically significant and the direction of the relationship is accurate. None, some, or all of the statements might be true.

Survey respondents who had higher incomes had more prestigious occupations
Survey respondents who had graduated from high school had less prestigious occupations compared to the average for all survey respondents
Survey respondents who had a junior college degree had less prestigious occupations compared to the average for all survey respondents
Survey respondents who had a bachelor's degree had more prestigious occupations compared to the average for all survey respondents
Survey respondents who had a graduate degree had more prestigious occupations compared to the average for all survey respondents
Survey respondents who were white had less prestigious occupations compared to the average for all survey respondents
Survey respondents who were black had more prestigious occupations compared to the average for all survey respondents

Evaluating the Relationship Between Income and Occupational Prestige

The statement that "survey respondents who had higher incomes had more prestigious occupations" is correct. The individual relationship between the independent variable "income" [rincom98] and the dependent variable "occupational prestige score" [prestg80] was statistically significant, $\beta = .297$, $t(159) = 4.09$, $p < .001$. We reject the null hypothesis that the partial slope (b coefficient) for the variable income = 0 and conclude that the partial slope (b coefficient) for the variable income is not equal to 0. The positive sign of the b coefficient (0.780) means that higher values of income were associated with higher values of occupational prestige score.
Marking the Statement of Relationship Between Income and Occupational Prestige

The statement that "survey respondents who had higher incomes had more prestigious occupations" is correct, so we mark the check box.

The next statement focuses on the relationship between graduating from high school and occupational prestige.

Evaluating the Relationship High School Graduation and Occupational Prestige

The individual relationship between the independent variable "survey respondents who had graduated from high school" [degree_1] and the dependent variable "occupational prestige score" [prestg80] was statistically significant, $\beta = -.225$, $t(159) = -3.43$, $p < .001$.

We reject the null hypothesis that the partial slope (b coefficient) for the variable survey respondents who had graduated from high school = 0 and conclude that the partial slope (b coefficient) for the variable survey respondents who had graduated from high school is not equal to 0. The negative sign of the b coefficient (-5.240) means that survey respondents who had graduated from high school had less prestigious occupations compared to the average for all survey respondents.
Marking the Relationship of High School Graduation to Occupational Prestige

The statement that "survey respondents who had graduated from high school had less prestigious occupations compared to the average for all survey respondents" is correct, so we mark the check box.

Evaluating the Relationship of Junior College Graduation to Occupational Prestige

The individual relationship between the independent variable "survey respondents who had a junior college degree" [degree_2] and the dependent variable "occupational prestige score" [prestg80] was not statistically significant, $\beta = -.041$, $t(159) = -.59$, $p = .555$. We are not able to reject the null hypothesis that the partial slope (b coefficient) for the variable survey respondents who had a junior college degree = 0.
The statement that "survey respondents who had a junior college degree had less prestigious occupations compared to the average for all survey respondents" is not correct, so we do not mark the check box.

The next statement focuses on the relationship between graduating from college and occupational prestige.

The individual relationship between the independent variable "survey respondents who had a bachelor's degree [degree_3] and the dependent variable "occupational prestige score" [prestg80] was statistically significant, $\beta = .192, t(159) = 2.85, p = .005$. We reject the null hypothesis that the partial slope (b coefficient) for the variable survey respondents who had a bachelor's degree = 0 and conclude that the partial slope (b coefficient) for the variable survey respondents who had a bachelor's degree is not equal to 0.

The positive sign of the b coefficient (5.860) means that survey respondents who had a bachelor's degree had more prestigious occupations compared to the average for all survey respondents.
Marking the Relationship of College Graduation to Occupational Prestige

The statement that "survey respondents who had a bachelor's degree had more prestigious occupations compared to the average for all survey respondents" is correct, so we mark the check box.

The next statement focuses on the relationship between graduate degrees and occupational prestige.

Evaluating the Relationship of Graduate Degrees and Occupational Prestige

The individual relationship between the independent variable "survey respondents who had a graduate degree" [degree_4] and the dependent variable "occupational prestige score" [prestg80] was statistically significant, $\beta = .265$, $t(159) = 3.49$, $p < .001$. We reject the null hypothesis that the partial slope (b coefficient) for the variable survey respondents who had a graduate degree = 0 and conclude that the partial slope (b coefficient) for the variable survey respondents who had a graduate degree is not equal to 0.

The positive sign of the b coefficient (9.230) means that survey respondents who had a graduate degree had more prestigious occupations compared to the average for all survey respondents.
Marking the Relationship of Graduate Degrees and Occupational Prestige

The statement that "survey respondents who had a graduate degree had more prestigious occupations compared to the average for all survey respondents" is correct, so we mark the check box.

The next statement focuses on the relationship between being white and occupational prestige.

Evaluating the Relationship of Being White and Occupational Prestige

The individual relationship between the independent variable "survey respondents who were white" [race_1] and the dependent variable "occupational prestige score" [prestg80] was statistically significant, β = -.166, t(159) = -2.39, p = .018. We reject the null hypothesis that the partial slope (b coefficient) for the variable survey respondents who were white = 0 and conclude that the partial slope (b coefficient) for the variable survey respondents who were white is not equal to 0.

The negative sign of the b coefficient (-5.430) means that survey respondents who were white had less prestigious occupations compared to the average for all survey respondents.
The statement that "survey respondents who were white had less prestigious occupations compared to the average for all survey respondents" is correct, so we mark the check box.

The next two statements focus on the relationship between being black and occupational prestige.

The statement that "survey respondents who were black had less prestigious occupations compared to the average for all survey respondents" is not correct. The individual relationship between the independent variable "survey respondents who were black" [race_2] and the dependent variable "occupational prestige score" [prestg80] was not statistically significant, $\beta = -0.062$, $t(159) = -0.92$, $p = 0.360$. 
Marking the Relationship of Being Black and Occupational Prestige

The statement that "survey respondents who were black had less prestigious occupations compared to the average for all survey respondents" is not correct. The statement that "survey respondents who were black had more prestigious occupations compared to the average for all survey respondents" is not correct. Neither statement is marked.
Logic Outline for Standard Multiple Regression Problems

1. Level of measurement ok?
   - Yes
   - No
     - Do not mark check box
     - Mark: Inappropriate application of the statistic
     - Stop

2. Ordinal level variable treated as metric?
   - Yes
     - Consider limitation in discussion of findings
   - No
     - Run Script to Evaluate Regression Assumptions for Model 1: Original Variables, All Cases

   - Use α for diagnostic tests

3. Normality ok?
   - Yes
     - Mark check box for Model 1
   - No
     - Don’t mark check box for Model 1

   - Linearity ok
   - Homoscedasticity ok?
   - Independence of variables ok?
   - Independence of errors ok?

4. If no extreme outliers are identified in Model 1, skip this step

5. Run Script to Evaluate Regression Assumptions for Model 2: Original Variables, Excluding Extreme Outliers

6. Normality ok?
   - Yes
     - Mark check box for Model 2
   - No
     - Don’t mark check box for Model 2

   - Linearity ok
   - Homoscedasticity ok?
   - Independence of variables ok?
   - Independence of errors ok?
Logic Outline for Standard Multiple Regression Problems - 2

Run Script to Evaluate Regression Assumptions for Model 3: Transformed Variables, All Cases

Normality ok? Linearity ok Homoscedasticity ok? Independence of variables ok? Independence of errors ok?

Yes → Mark check box for Model 3

No → Don’t mark check box for Model 3

Run Script to Evaluate Regression Assumptions for Model 4: Transformed Variables, Excluding Extreme Outliers

If no extreme outliers are identified in Model 3, skip this step

Normality ok? Linearity ok Homoscedasticity ok? Independence of variables ok? Independence of errors ok?

Yes → Mark check box for Model 4

No → Don’t mark check box for Model 4

Interpret Model 1
Mention violations of assumptions as limitations to the analysis

Sample size ok?

Yes

No → Do not mark sample size Consider limitation to findings
Logic Outline for Standard Multiple Regression Problems - 3

Mark check box for correct sample size

Overall relationship (F-test Sig ≤ α)?
- Yes: Use α for statistical tests
- No: Do not mark check box for overall relationship

Correct adjective used to describe effect size?
- Yes: Mark check box for overall relationship
- No: Do not mark check box for overall relationship

Individual relationship (t-test Sig ≤ α)?
- Yes: Use α for statistical tests
- No: Do not mark check box for individual relationship

Correct interpretation of direction of relationship?
- Yes: Mark check box for individual relationship
- No: Repeat evaluation of individual relationships for other variables

Repeat evaluation of individual relationships for other variables