

|  |
| --- |
| Computational Assignment #3: OLS Regression Modeling with Categorical Variables  *MSDS 410* |

This third computational assignment builds on your prior modeling and computing experiences. You may begin to work on this assignment anytime you wish.

**Data:** The data for this assignment is the Nutrition Study data: NutritionStudy.CSV It is a 16 variable dataset with n=315 records. The data was obtained from medical record information and observational self-report of adults. The dataset consists of categorical, continuous, and composite scores of different types. A data dictionary is not available for this dataset, but the qualities measured can easily be inferred from the variable and categorical names for most of the variables. As such, higher scores for the composite variables translate into having more of that quality. The QUETELET variable is essentially a body mass index. It can be googled for more detailed information. It is the ratio of BodyWeight (in lbs) divided by (Height (in inch))^2. Then the ratio is adjusted with an adjustment factor so that the numbers become meaningful. Specifically, QUETELET above 25 is considered overweight, while a QUETELET above 30 is considered obese. There is no other information available about this data.

**Objective:** Use multiple regression to predict CHOLESTEROL using models with categorical variables. Please note: This assignment is not prescriptive of what you “should do” as an analysis. It is intended to give you experience conducting and reporting on different kinds of multiple regression models.

**Tasks:** To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

For these analyses, let the response variable be: Y = CHOLESTEROL. The remaining variables will be considered explanatory variables, X’s.

1. For all of the categorical variables in the dataset, recode the text based categories into numerical values that indicate group. For example, for the VITAMIN variable, you could code it so that: 1=regular, 2=occasional, 3=never. Save the categorical variables to the dataset.
2. For the VITAMIN categorical variable, fit a simple linear model that uses the categorical variable to predict the response variable Y=CHOLESTEROL. Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Recode the VITAMIN categorical variable so that you have a different set of indicator values. For example, you could code it so that: 1=never, 2=occasional, 3=regular. Re-fit an OLS simple linear model using the new categorization. Report the model, interpret the coefficients, discuss test results, etc. What is going on here?
3. Create a set of dummy coded (0/1) variables for the VITAMIN categorical variable. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL (Y). Remember, you need to leave one of the dummy coded variables out of the equation. That category becomes the “basis of interpretation.” Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 2). What has changed?
4. For the VITAMIN categorical variable, use the NEVER categorical as the control or comparative group, and develop a set of indicator variables using effect coding. Save these to the dataset. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 3). What has changed? Which do you prefer? Why?
5. Discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:
6. if ALCOHOL = 0
7. if 0 < ALCOHOL < 10
8. if ALCOHOL >= 10

Use these categories to create a set of indicator variables for ALCOHOL that use effect coding. Save these to your dataset.

1. At this point, you should have effect coded indicator variables for VITAMIN and 2 effect coded indicator variables for ALCOHOL. Create 4 product variables by multiplying each of the effect coded indicator variables for VITAMIN by the effect coded indicator variables for ALCOHOL. This is all pairwise products of the effect coded variables. Now, we are going to test for interaction. Fit an OLS multiple regression model using the 4 VITAMIN and ALCOHOL effect coded indicator variables plus the 4 product variables to predict CHOLESTEROL. Call this the full model. For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL. Conduct a nested model F-test using the Full and Reduced Models described here. Be sure to state the null and alternative hypothesis, make a decision regarding the test, and interpret the result. Obtain a means plot to illustrate any interaction, or lack thereof, to help explain the result.
2. There are 2 other categorical variables in this dataset, namely GENDER and SMOKE. Do these variables interact amongst themselves or with VITAMIN or ALCOHOL when it comes to modeling CHOLESTEROL? Obtain means plots to see if there is interaction. Conduct nested model F-tests to rule out randomness as the explanation for observed patterns. Report your findings.
3. Please write a reflection on your experiences from this assignment.