

|  |
| --- |
| Computational Assignment #4: OLS Regression Modeling with Continuous  and Categorical Variables  *MSDS 410* |

This fourth computational assignment builds on your prior modeling and computing experiences with assignment #3. 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 continuous and 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. Consider the continuous variable, FIBER. Is this variable correlated with Cholesterol? Obtain a scatterplot and appropriate statistics to address this question.
2. Fit a simple linear regression model that uses FIBER to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss the goodness of fit.
3. For the ALCOHOL categorical variable, create a set of dummy coded (0/1) indicator variables. Fit a multiple linear model that uses the FIBER continuous variable and the ALCOHOL dummy coded variables to predict the response variable Y=CHOLESTEROL. Remember to leave one of the dummy coded variables out of the model so that you have a basis of interpretation for the constant term. Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. This is called an Analysis of Covariance Model (ANCOVA)
4. Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, make a scatterplot of the actual values of CHOLESTEROL(Y) by FIBER (X), but color code by the different groups of the ALCOHOL variable. If you compare the two scatterplots, does the ANCOVA model appear to fit the observed data very well? Or, is a more complex model needed?
5. Create new interaction variables by multiplying the dummy coded variables for ALCOHOL by the continuous FIBER(X) variable. Save these product variables to your dataset. Now, to build the model, start with variables in your ANCOVA model from task 4) and add the interaction variables you just created into the multiple regression model. Don’t forget, there is one category that is the basis of interpretation. DO NOT include any interaction term that is associated with that category. This is called an Unequal Slopes Model. Fit this model, and save the predicted values. Plot the predicted values for CHOLESTEROL (Y) by FIBER(X). Discuss what you see in this graph. In addition, report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics.
6. You should be aware that the models of Task 4) and Task 5) are nested. Which model is the full and which one is the reduced model? Write out the null and alternative hypotheses for the nested F-test in this situation to determine if the slopes are unequal. Use the ANOVA tables from those two models you fit previously to compute the F-statistic for a nested F-test using Full and Reduced models. Conduct and interpret the nested hypothesis test. Are there unequal slopes? Discuss the findings.

1. Now that you’ve been exposed to these modeling techniques, it is time for you to use them in practice. Let’s examine more of the NutritionStudy data. Use the above practiced techniques to determine if SMOKE, VITAMINS, or GENDER interacts with the FIBER variable and influences the amount of CHOLESTEROL. Formulate hypotheses, construct essential variables (as necessary), conduct the analysis and report on the results. Which categorical variables are most predictive of CHOLESTEROL, in conjunction with FIBER.
2. Please write a reflection on your experiences.
3. Extra Credit: Feel free to explore models that have other continuous variables, as well as interactions of categorical variables. The more you do, the more extra credit you can accumulate.