Modeling Assignment 5: Modeling with Categorical Explanatory Variables – ANOVA, ANCOVA, and Unequal Slopes Models

Assignment Overview

In this assignment we will use multiple regression to predict CHOLESTEROL. The catch is that we will be using categorical variables as the explanatory variables in different combinations with continuous variables. The absence, presence and combinations of such variables gives rise to different kinds of situations that you may have seen or heard of in the past. For example, in MSDS 401, you were exposed to Analysis of Variance (ANOVA) for making mean comparisons of 2 or more groups. It turns out that a regression model using a categorical variable, appropriately dummy coded, is equivalent to an ANOVA model. This assignment is not prescriptive of what you "should do" as an analyst. Rather, it is intended to give you experience conducting and reporting on different kinds of multiple regression models that utilize categorical explanatory variables.

The data for this assignment is from a study of nutrition. The Nutrition Study Data 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 variables translate into having more of that quality. There is one variable, called QUETELET, that 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 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.

Preparatory Work

In this fifth Modeling Assignment, we are working with a brand-new dataset. When you are in such a situation, in addition to Sample Population determination and EDA to understand the data, you may also need to do some transformations on the existing variables. If you intend to use categorical variables as explanatory variables in your models, you will have to potentially recode variables or construct dummy coded variables prior to any modeling.

When you import the Nutrition Study Data into R or EXCEL, you will notice that there are 4 categorical variables. These are: SMOKE, GENDER, VITAMINUSE, and PRIORSMOKE. Some of these variables use numbers to indicate the levels of the categorical variables, others use text. For regression modeling purposes, you will most likely need to transform these variables, or construct new dummy coded variables. How you do this is as follows:

a) For any dichotomous categorical variable (i.e. a categorical variable with 2 levels), you want to recode such a variable so that the values (or numbers) that indicate the level are set to 0 and 1. The GENDER and SMOKE variables are like this. Often, an analyst will just create a new variable, like d_GENDER, that is the coded version of GENDER.
b) For categorical variables with 3 or more levels, you will need to construct a set of dummy coded (0/1) variables to indicate the levels. The VITAMINUSE and PRIORSMOKE variables are like this. Please see the Module 5 Classroom for directions on how to construct dummy coded variables. Each level must have its own dummy coded variable. As such, there should be 3 dummy coded variables for VITAMINUSE. Similarly, there will be 3 dummy coded variables for PRIORSMOKE.
c) Some analysts like to take continuous variables and discretize or convert them into categorical. For example, the ALCOHOL variable may be easier to work with or interpret results if it were converted into a variable called ALCOHOL CONSUMPTION with levels like: None, Some, A lot. In doing this, you could discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:

    1 if ALCOHOL = 0

    2 if 0 < ALCOHOL < 10

    3 if ALCOHOL >= 10

    Once you have the levels for the new ALCOHOL CONSUMPTION categorical variable, you would then dummy code these levels.

In preparation for modeling, you need to create dummy coded variables for the categorical variables in the Nutrition Study data set. Construct the ALCOHOL CONSUMPTION categorical variable and create dummy coded variables for it.

Assignment Tasks

For the tasks in this assignment, the response variable will be: CHOLESTEROL (Y). The remaining variables will be considered explanatory variables (X's).

1. Obtain descriptive statistics (n, mean, s, and any others you want) for Y by the PRIORSMOKE variable. Use the PRIORSMOKE variable as a factor in an ANOVA to test for mean differences in Cholesterol between PRIORSMOKE groups. Report and interpret these results.
2. Fit a linear regression model that uses the dummy coded variables for PRIORSMOKE to predict Cholesterol (Y). Call this Model 1. Remember: you need to leave one of the dummy coded variables out of the equation. That category becomes the "basis of interpretation." Report the prediction equation and interpret each coefficient in the context of this problem. Report the coefficient and ANOVA tables from this regression model. Discuss how the results from the regression model compare and contrast to the results from the ANOVA model in Task 1.
3. Model 1 illustrates the ANOVA model as a Linear Regression Model. Let's go a step further. Start with Model 1 and add in the continuous variable FAT. In other words, you are using

FAT and PRIORSMOKE to predict Cholesterol, but you are using dummy coded variables for the PRIORSMOKE categorical variable. More specifically, fit a multiple linear model that uses the FAT continuous variable and the PRIORSMOKE dummy coded variables to predict the response variable CHOLESTEROL (Y). 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 prediction model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics, if it is relevant. This is called an Analysis of Covariance Model (ANCOVA). Call this Model 2.

4. Use the ANCOVA Model 2 from Task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FAT (X), but color code the records for the different groups of PRIORSMOKE. What do you notice about the patterns in the predicted values of Y? Make a second scatterplot of the actual values of CHOLESTEROL(Y) by FAT (X), but color code the data points by the different groups of the PRIORSMOKE 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 product variables by multiplying each of the dummy coded variables for PRIORSMOKE by the continuous FAT(X) variable. Name and save these product variables to your dataset. Now, to build the Unequal Slopes Model, start with the ANCOVA model, Model 2, from Task 3). Add in the interaction variables you just created. You now should have a multiple regression model with the predictor variables of: FAT, two dummy coded PRIORSMOKE variables, and two product variables. This is called an Unequal Slopes Model – call it Model 3. Fit Model 3 and report the prediction equation, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, leverage, influence, and Outlier statistics, if warranted.
6. Use Model 3 to obtain predicted values. Plot the predicted values for CHOLESTEROL (Y) by FAT(X). Discuss what you see in this graph.
7. You should be aware that Model 2 and Model 3 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 to determine if the slopes are unequal. Use the ANOVA tables from Models 2 and 3 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 in this situation? Discuss the findings.
8. 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 modeling approach to determine if the categorical variables SMOKE, ALCOHOL CONSUMPTION or GENDER, along with the continuous variables FAT variable are predictive 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?
9. Please write a conclusion / reflection on your experiences in this assignment.

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present

your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign5_YourLastName.pdf.