Robin Singh



|  |
| --- |
| 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.

Table

Description automatically generated

1. 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?

Text

Description automatically generated

Y = 232.634 + 5.001\*B1

Y is the amount of cholesterol, B1 is the Vitamin usage as determined previously (1=never, 2=occasional, 3=regular).

Hypothesis test:

Null: B1= 0

Alt: B1!= 0

t-value=0.577

t-value to analyze against is approximately 2.5.

0.577<2.5, which means the null hypothesis cannot be disproved, and has a significant chance of being true.

For goodness of fit, the p-value is 0.5642, meaning there is a high chance of the null hypothesis being true.

We also have a very low R^2 value of 0.001063, and an adjusted R^2 of -0.002128, indicated a poorly fit model.

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generated

There is no change, besides in direction of line which is now negative, upon flipping the variables in the model.

Text

Description automatically generated

1. 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?

Text

Description automatically generated

Y=246.599 + -9.908\*B1 + -1.156\*B2

Y is the amount of cholesterol, -9.908 is the effect of regularly using Vitamins (B1), and -1.156 is the effect of occasionally using vitamins (B2), all compared to never using vitamins, which was the variable that was omitted form the model.

Hypothesis test:

Null: B1=B2=0

Alt: B1 != B2 != 0

The t-values for B1 and B2 are -0.571 and -0.060, which are both less than the threshold of around 2.5, meaning that the null hypothesis cannot be disregarded because it may hold.

The p-value is 0.8262 which is high, indicating the null hypothesis holds.

The R^2 value is 0.001223 and adjusted R^2 is -0.005179, indicating a poor fitting model.

Chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generatedChart, box and whisker chart

Description automatically generated

There are not many differences in the model. It still indicates that vitamins alon are a poor predictor of cholesterol.

1. 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?

Text, letter

Description automatically generated

Y=242.911 -6.220\*B1 + 2.532\*B2

Y is the amount of cholesterol, -6.22 is the effect of regularly using Vitamins (B1), and 2.532 is the effect of occasionally using vitamins (B2), all compared to never using vitamins, which was the variable that was control variable.

Hypothesis test:

Null: B1=B2=0

Alt: B1 != B2 != 0

The t-values for B1 and B2 are -0.607 and -0.223, which are both less than the threshold of around 2.6, meaning that the null hypothesis cannot be disregarded because it may hold.

The p-value is 0.8262 which is high, indicating the null hypothesis holds. This is the same as the dummy coding.

The R^2 value is 0.001223 and adjusted R^2 is -0.005179, indicating a poor fitting model, same as before.

Chart

Description automatically generated Chart

Description automatically generated

Chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart

Description automatically generated

I do like this way better, as it takes into account the control variable in some way, including it in the model.

1. Discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:
2. if ALCOHOL = 0
3. if 0 < ALCOHOL < 10
4. if ALCOHOL >= 10

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

Table

Description automatically generated

The ‘No Alcohol’ is the control variable.

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.

Graphical user interface, text

Description automatically generated

For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL.

Text

Description automatically generated

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.

Null: B5=B6=B7=B8=B9=0

Alt: One does not equal 0

F-test= ((5394372-5342216)/(8-4))/( 5394372/(315-8))=0.7421

F-stat to analyze against=2.37

0.7421<2.37

We cannot reject the null hypothesis because the f-stat for our model is lower than the f-stat to analyze against, meaning that the four additional variables in the full model make a significant difference in our model.

Chart

Description automatically generated with low confidenceChart, box and whisker chart

Description automatically generated

1. 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.

Text

Description automatically generated

F-test:

Null:B9=B10=0

Alt: One of them is not zero

Graphical user interface, text, application

Description automatically generated

We can reject the null hypothesis that this has no relations combined with the Vitamin and Alcohol variables on the Cholesterol variable because the f-stat value is significantly greater at 9.8232, meaning there is evidence supporting the alternate hypothesis.

Individually, Smoke has a higher impact on the Cholesterol variable with an F-stat of 21.342, while gender significantly less with 2.2597.

Graphical user interface, text, application

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

1. Please write a reflection on your experiences from this assignment.

This assignment was initially easy but got a little confusing as things went on. Some of the questions were a little tough to understand in terms of what exactly they wanted, and many were more loaded than others. Still, this taught a lot about how to incorporate categorical features into linear models, and how much of an impact they can truly have on the overall prediction. I did not know how various techniques of implementing them may have different impacts on the overall model.