Data Analysis 504
Logistic Regression
Assignment 4

You use logistic regression when you have a binary dependent variable and either interval/ratio scale or binary independent variables. While chi-square statistics allow you to use a nominal scale dependent variable, logistic regression has much greater flexibility in the use of control variables than chi-square. Also, in logistic regression, you can determine the exact relationship between the independent variable and the dependent variable (much like OLS regression analysis). To use the logistical regression model, get into your data.

TO ACCESS THE DATA FROM SPSS

- Click the **Folder/File** icon visible in the bottom left hand corner of your computer screen. It is located next to the "Type here to search" box.
- Double Click on drive **Storage (S:)**, then on **Public Shares**, then on **Class Data**, then on **GSSW**.
- Scroll down to the bottom of the page where you will find many copies of the PSID dataset Double click **PSID1999to2019_1**. You can also click on **PSID1999to2019**_2, or **PSID1999to2019**_3, or any of the other versions of PSID1992to2019 in SPSS. We have these different versions so that no two people are working on the same data. Make sure you do not choose **PSID1992to2015**.sas or **PSID1992to2015**.sas7bdat (these are for another type of software).

Logistic Regression

To run a multiple logistic regression analysis, go to **Analyze**, then to **Regression**, then to **Binary Logistic**.
Right click on the variables to sort them alphabetically and to display the variable name.

Selecting your variables & running the regression

In this assignment, you will examine which factors help explain why some people get cancer.

You'll use the variable **Cancer15** as your dependent variable, which has a value of 1 when the person has cancer and a value of 0 when the person does not have cancer. Click on **Cancer15**, and click on the arrow to move this into the box labeled *dependent variable*.

For your *independent and control variables*, or covariates, click on **age13** (age of the person, interval/ratio), **White13** (race of white, dummy variable), **Gupoor13** (did this person grow up poor, dummy variable), **Younghealthex** (was the health status of this person excellent from ages 0 to 16, dummy variable), **Smoke13 (**whether the person smoked in 2015, dummy variable) and **famsiz13** (the number of people in the family, interval/ratio). Click on **OK**.

Your research question is: Is the occurrence of cancer related to such factors as the age, whether the person is white (relative to all other races), whether the person grew up in a poor household (relative to all other households), whether the person had excellent health growing up (relative to those who had very good, good, fair, or poor health growing up), whether the person smokes (relative to those who do not smoke), and their family size. Each of the coefficient estimates is determined while controlling for the effects of the other variables within the model.

Interpreting Your Results

To determine the significance levels for the coefficient estimates for the variables, look at the bottom table (labeled variables in the equation) and examine the column of your output labeled Sig. These are your two-tailed significance levels. If this column indicates a level under .05, you can reject your null hypothesis at the 5% level of significance for a two-tailed test. A value of .6160 indicates that this coefficient estimate is significant at the 61.07% level -- higher than the 5% level. For all variables that are above the .05 level, fail to reject the null hypothesis (you can adjust the significance levels in the Sig column of your output for one tailed tests). A Sig value of .0000 indicates that the coefficient is significant at the 0% level (or is in the far tail portion of the distribution), and is thus highly significant.

Questions:

1. **Which of the independent variables (or the coefficient/b estimates) have a significant relationship with the dependent variable? Which do not have a significant relationship with the dependent variable?**

Significant: Age; White

Not significant: growing up poor; childhood health being excellent; smoking; family size

2. **Can you support your hypothesis that age is a significant predictor for likelihood of cancer diagnosis?**

Yes, age is significant at all levels with a p-value of .000

3. **Use the Exp(B) or the odds ratio column to interpret the likelihood of having a value of 1 for your significant dependent variable.** [For dummy variables (independent variables), you are examining the likelihood of being in the 1 category versus being in the 0 category. Thus, if you are examining the effects of being depress (independent variable) on the likelihood of suicide (dependent variable) and get an odds ratio of 1.5, this indicates that those who are depressed are 1.5 times as likely or 50% more likely to take their life by suicide than those who are not depressed. If you are examining the effects of the number of times you spank the child affects the child's likelihood of having poor health and get an odds ratio of 1.03, this indicates that for every time the child is spanked, the likelihood of having poor health increases by 3 percent.]

   Two significant findings are Age (I/R variable) & Race(White is a nominal variable)

   Formula for interpreting our findings: (OR-1) *100

   Age Exp(B)/OR value = 1.055
   (1.055-1)*100 = .055*100 = 5.5%
   For each additional year older someone is their likelihood of having a cancer diagnosis increases by 5.5%.

   White Exp(B)/OR value = 1.893
   (1.893-1)*100 = .893*100 = 89.3%
   Those who are white are 89.3% more likely to have cancer compared to those who are a race other than white.

4. **How much of the variation of the dependent variable is explained by the set of independent variables?**

   Take the R-Square value and multiple by 100 to turn it into a percent.

   According to the Cox & Snell R-Square value, our model explains 5.5% of the variation of our DV(getting cancer).

According to the Nagelkerke R-Square value, our model explains 13.4% of the variation of our DV(getting cancer).

5. **Is your model significant? How do you know this?**

Yes, looking at our model significance level, which is .000, our model is significant at all levels.

You are now done your lass computer assignment !

****

Note: If the dependent variable in your paper is a nominal scale variable, you will be using logistic regression analysis for this final paper. If you would like, go to the data set you are using for your paper, and use the dependent variable that you'll be using for your paper. Use some of the control and independent variables you will be using for your paper. What do you find?