**University of Calgary**

**Cumming School of Medicine**

**MDCH 640 Fall 2018**

**Fundamentals of Epidemiology**

**Student Name: Mili Roy**

**ID Number : 10128986**

**Submitted to: Dr. Kirsten M. Fiest**

**Dr. Reed Beall**

**Assignment#8**

**Logistic Regression**

**Date: 29th November 2018**

The purpose of this study is to measure associations between exposure and disease and investigating the effect of extraneous variable. The data used for this assignment was taken from the Canadian community health survey (CCHS) conducted by statistics Canada. Although it is cross sectional data, it is assumed the data are incidence data and cases were considered new rather than prevalent cases.

The CCHS is a cross-sectional survey designed to estimate the prevalence of diseases in the Canadian population. Thus, we will be estimating prevalence odds ratios using logistic regression.

**QUESTION 1**

Using Logistic Regression, estimate the crude odds ratio of asthma (CCC\_031) using being a smoker (SMK\_202) as the exposure. (SMK\_202 has 3 categories; you should combine two of these). Provide an interpretation for the odds ratio and the 95% confidence interval.

OR with CI; interpretation; justification for combining; discussion of results

**Table 1: Frequency table for the variable smoker**

|  |  |  |
| --- | --- | --- |
| Smoker | Frequency | percent |
| yes | 29110 | 22.27 |
| No | 101596 | 77.73 |
| Total | 130706 | 100 |

**Table 2: Crude odds ratio of asthma using smoker as exposure using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 1.123 | <.001 | 1.073 1.175 |

Table 1 reports the frequency and percentage of the smoking status. It is pertinent to mention that; this variable is categorized into two groups from three groups. People who were daily and occasional smoker they were considered as smoker and not at all category was considered as non-smoker. Frequency is also checked after reducing one category. Out of 1,30,706, only 22.27% respondents were found with smoking status implies the highest percentage (77.73) of the study population were free of smoking. From table 2, since the crude odds ratio from logistic regression 1.123>1, we can say, there is a positive association between smoker and asthma patients. According to the table, individuals with being smoker are 1.123 times more likely to develop asthma than those without having smoking status. Also, 95% confidence interval shows the statistical significant relationship between smoking and asthma since the interval does not contain the null value 1. In addition, we are 95% confident that based on data, the true odds ratio for smoking will lie somewhere between 1.073 and 1.175. The confidence is narrow and affirms the accuracy the accuracy of the estimate. Also, smoking category is combined in two categories all the disease and exposure variable should be dichotomized as 0 or 1 in logistic regression.

**QUESTION 2**

Using logistic regression, determine if gender modifies or confounds the relationship between asthma and smoking.

OR with CI for female and male; interpretations; discussion of results; modifier? confounder? why or why not?

**Table 1: Sex specific odds ration using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| Asthma (disease) | Odds ratio | P value | 95% confidence interval |
| smoking (exposure) | 0.991 | 0.802 | 0.921 1.065 |
| sex | 1.292 | <.001 | 1.233 1.352 |
| interaction | 1.279 | <.001 | 1.165 1.404 |

**Table 2: Odds ratio and 95% CI for male using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| Male | 0.990 | <.001 | 0.921 1.065 |

**Table 3: Odds ratio and 95% CI for female using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| Female | 1.267 | <.001 | 1.194 1.344 |

As we know, effect measure modification means stratum-specific estimates of outcome-exposure association differ by levels of a third variable, we need to investigate whether the relationship between asthma and smoking are influenced by effect modifier using logistic regression.

From table 1, we reported the odds ratio for asthma and smoking by modifying the third variable sex. Since the interaction between sex and smoker is found significant from Wald test (p value<.001), there is evidence of effect measure modification. It means the estimates are heterogenous by third variable. Odds of female smoker have 1.267 times more likely to develop asthma than odds of female who do not smoke. From table 1, from the 95% CI, we found that sex is significant as the confidence interval does not include 1 (1.233-1.352) but the exposure variable smoking is not significant (0.921-1.065). Similarly, from table 3, the crude odds ratio of logistic regression of asthma and smoking relationship, female was found significant (1.194-1.344) whereas male is not significant as the confidence interval includes 1 (0.921-1.065). Also, we are 95% confident that the true odds ratio for female population will lie somewhere between 1.194 and 1.344, respectively. Therefore, asthma and smoking relationship is mixed by the effect modifier gender and it is modified by female.

**QUESTION 3**

Which odds ratio(s) would you report and why? (i.e. Stratum-specific, Adjusted, or Crude)

OR with CI stated; reasoning

**Table 1: Odds ratio and 95% CI for male using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| Male | 0.990 | <.001 | 0.921 1.065 |

**Table 2: Odds ratio and 95% CI for female using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| Female | 1.267 | <.001 | 1.194 1.344 |

Stratum specific or sex specific odds ratio is reported since the relationship between smoking and asthma are modified by effect measure because of having interaction effect between sex and smoking. Therefore, crude OR is not applicable in this situation. Crude OR is unadjusted whereas considering the interaction effect is adjusted.

**QUESTION 4**

**How do your results throughout Question 1 compare to your results using stratified analysis in Assignment 7?**

**Table 1: Crude odds ratio of asthma using smoker as exposure using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| asthma (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 1.123 | <.001 | 1.073 1.175 |

**Table 2**: **Sex specific odds ratio and CI for asthma and smoking**

|  |  |  |
| --- | --- | --- |
| Sex | OR | 95% Confidence Interval |
| Male | 0.991 | 0.920 1.065 |
| Female | 1.267 | 1.194 1.344 |
| Crude | 1.123 | 1.072 1.176 |
| M-H combined | 1.145 | 1.094 1.198 |

Test of homogeneity (M-H) chi2(1) = 26.77 P value<.001

In this study in question 1 we assessed the association between outcome and one exposure variable. Table 1 reveals, smoking is significant variable (p value<.001) and there is a positive association between asthma and smoking (OR=1.123>1). The odds of smoker were 1.123 times more likely to develop asthma than odds of non-smoker. However, stratified analysis is powerful statistical technique that allows us to test for effect modification and confounding. In table 2, we gain more insight from analysis as stratified analysis helps to measure the association between two categorical variables by adjusting for third categorical variable; gender. From table 2, we found, strata specific odd ratios are substantially different (male=0.991, female=1.267) implies the impact of effect modifier. Moreover, test of homogeneity explains the presence of effect modifier. Significant p value (<.001) tells that sex is effect modifier and interestingly, it is modified by female strata as the confidence interval does not include 1 ( 1.194-1.344). Finally, we can summarize stratified analysis is better than crude analysis (unadjusted logistic regression)

as it reports the strata specific estimates and helps to make decision regarding the effect of third variable.

**QUESTION 5**

Using logistic regression, estimate the odds ratio of emphysema (CCC\_91E) using being a smoker (SMK\_202) as the exposure. Provide an interpretation for the odds ratio and the 95% confidence interval.

OR with CI; interpretation; justification for combining; discussion of results

**Table 1: Frequency table for the variable smoker**

|  |  |  |
| --- | --- | --- |
| Smoker | Frequency | percent |
| yes | 29110 | 22.27 |
| No | 101596 | 77.73 |
| Total | 130706 | 100 |

**Table 2: Odds ratio of emphysema using smoker as exposure using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 2.009 | <.001 | 1.797 2.246 |

Table 1 reports the frequency and percentage of the smoking status. It is pertinent to mention that, this variable is categorized into two groups from three groups. People who were daily and occasional smoker they were considered as smoker and not at all category was considered as non-smoker. Frequency is also checked after reducing one category. Out of 1,30,706, only 22.27% respondents were found with smoking status implies the highest percentage (77.73) of the study population were free of smoking. From table 2, since the crude odds ratio 2.009>1, we can say, there is a positive association between smoker and Emphysema patients. According to the table, odds of individuals with being smoker are 2.009 times more likely to develop Emphysema than odds of those without having smoking status. Also, 95% confidence interval shows the statistical significant relationship between smoking and Emphysema since the interval does not contain the null value 1. In addition, we are 95% confident that based on data, the true odds ratio value will lie somewhere between 1.794 and 2.246. The confidence is wide, and we are not certain about the accuracy of the estimate.

**QUESTION 6**

Using logistic regression, determine if less than very good perceived health (GEN\_01) modifies the relationship between emphysema and smoking.

OR with CI for strata; interpretations; discussion of results; modifier? why or why not?

**Table 1: Frequency table for the variable Heath**

|  |  |  |
| --- | --- | --- |
| Health | Frequency | Percent |
| Healthy | 72848 | 55.67 |
| less than very good perceived health | 58003 | 44.33 |
| Total | 130851 | 100 |

**Table 2: Odds ratio of emphysema using smoker as health using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 1.739 | <.001 | 1.542 1.961 |
| Health(covariate) | 0.137 | <.001 | 0.113 0.166 |
| Interaction | 1.066 | 0.716 | 0.757 1.499 |

**Table 3: Crude odds ratio of emphysema using smoker as exposure using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| Less than good health | 1.739191 | <.001 | 1.542 1.961 |

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| Healthy | 1.853159 | <.001 | 1.345 2.551 |

Table 1 reports the frequency and percentage of the self perceived health status. It is pertinent to mention that, this variable is categorized into two groups from five categories (excellent, very good, good, fair, poor). People who belonged to excellent and very good health categories they were categorized as healthy and good, fair, poor categories were considered as unhealthy categories and coded as “less than very good perceived health”. Frequency is also checked after reducing the categories. Out of 1,30,851, 55.67% respondents perceived good health and whereas 44.43% of the respondents belong to less than very good perceived health. Note that, it is necessary to combine the categories as 2x2 cross table is required to calculate the odds ratio.

From table 2, we reported the odds ratio using logistic regression for emphysema and smoking by modifying the third variable less than very good perceived health. Since the interaction between health and smoking variable is insignificant (p value=0.716>.05), there is an indication that the third variable did not modify the relationship. Also, 95% confidence interval includes the null value 1 implies the interaction effect is insignificant. Consequently, we can check whether we have the impact of confounding. Table 2 and 3 report strata specific odds ratio. From the table 2 and 3, note that, people who perceive less than very good health with smoking status are 1.739 times more likely to develop emphysema than those without having smoking status. Similarly, odds of healthy people with smoking status were 1.853 times more likely to develop emphysema than odds of people who were not smoker. Also, we are 95% confident that the true odds ratio for healthy population will lie somewhere between 1.345 and 2.551, and the true odds ratio for people who perceive less than very good health will lie somewhere between 1.542 and 1.961 which is narrower and more precise than the CI of healthy people.

**QUESTION 7**

Using logistic regression, determine if less than very good perceived health (GEN\_01) confounds the relationship between emphysema and smoking.

Confounder? Why or why?

**Table 1: Odds ratio of emphysema and smoking using logistic regression with health confounder**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 1.752 | <.001 | 1.566 1.961 |
| Health(covariate) | 0.140 | <.001 | 0.113 0.166 |

**Table 2: Crude odds ratio of emphysema and smoking using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 2.009 | <.001 | 1.797 2.246 |

From table 1 and 2, we ran the logistic regression with health (confounder) and without health. It is evident that the adjusted OR (1.753>1) and crude OR (2.009>1) are different than each other in terms of magnitude. Also, the difference is more than 10% () in this case. Thus, less than very good perceived health confounds the relationship.

**QUESTION 8**

Which odds ratio would you report and why? (i.e. Stratum-specific, Adjusted, or Crude)

OR with CI stated; reasoning

**Table 1: Odds ratio of emphysema and smoking using logistic regression with health confounder**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 1.752 | <.001 | 1.566 1.961 |
| Health(covariate) | 0.140 | <.001 | 0.113 0.166 |

**Table 2: Crude odds ratio of emphysema and smoking using logistic regression**

|  |  |  |  |
| --- | --- | --- | --- |
| emphysema (disease) | odds ratio | p value | 95% confidence interval |
| smoking (exposure) | 2.009 | <.001 | 1.797 2.246 |

As from the comparison between table 1 and 2, as we found the health specific odds ratio is different than crude odds ratio, we have the impact of confounding. Thus, it is advisable to report adjusted odds ratio. Table 1 presents the odds ratio of the variable smoker (1.752) which is positively associated with emphysema disease. In addition, the odds of smoker were 1.752 more likely to develop emphysema than non-smoker in the presence of confounder. Also, we are 95% confident that the true odds ratio of smoker will lie somewhere between 1.566 and 1.961 which is very wide and less precise.

**QUESTION 9**

Briefly compare and contrast the stratified analysis and modeling approach for assessing the presence of effect measure modification and/or confounding.

Compare & contrast.

Stratified analysis and modelling are closely related procedures. Stratified analysis and modeling approach both are useful to detect effect measure and/or confounding. Stratification is a foundational skill for epidemiologists, but in current research it mostly functions as a preliminary procedure for more advanced analysis using statistical models. When these models are encountered in the literature, they can be very complex and difficult to understand. For example, they may include many variables and it may not be clear which, if any, are actual confounders. The procedure for selecting variables to include in the model are often unclear as well. The motivation for presenting complex models is to report estimates that are simultaneously adjusted for many potential confounders, but this approach comes at the cost of diminished transparency. Stratified extraneous variables at a time, so it takes a great deal of space to present stratified analysis for a variety of extraneous variables. In modeling we have beta coefficients and we predict the average of disease variable based on exposure variable. In linear regression models, beta can be interpreted as differences where in logistic regression, they are the log odds ratios. Modeling assist with some of the difficulties that arise in stratified analysis. Regression models deal more effectively with the sparse data problems that afflict stratified analysis. They are more capable of controlling multiple variables and joint confounding and deal better with continuous variables, such as age. Whereas, in stratification, continuous variables must be broken down, often arbitrarily, into categories.