Manav Dalmiya
MATH242
Professor An

Asthma and Its Risk Factors:
A Statistical Exploration of Asthma Status on Its Risk Factors.

1. Abstract:

This report will describe a logistic regression model on asthma status in the United States based on predictive demographic, biological, environmental, and behavioral factors. This research was motivated by an interest in understanding how much these predictive variables affect asthma status. The final model indicates that exposure to bad air quality is the most predictive factor influencing asthma status.

2. Introduction:

Asthma is a chronic disease in which an individual's bronchial passages become inflamed, causing symptoms such as wheezing, shortness of breath, chest tightness, and coughing. Asthma has many different variants including exercise-induced, allergy-induced, occupational, and childhood. The aim of this research project was to attain an understanding of how much specific predictive factors of asthma status affect the likelihood of having asthma. Modeling the likelihood of asthma is very important as the CDC estimates that over 25 million people in the United States have asthma by 2011. This is nearly a tenth of the overall US population and this number is only rising. Moreover, the CDC estimates that asthma-related costs due to medical expenses and missed school days and workdays were over $56 billion in 2007. The insights gathered from analyzing these predictive factors can inform a number of domains like behavior, policy, and medicine about how to reduce future increases in asthma status in the United States.

3. Methods:

The data used in this research was collected from the CDC's 2007-08 National Health and Nutrition Examination Survey. Surveys on Demographics, Air Quality, Income, Medical Conditions, Occupation, Physical Activity, Respiratory Health, Smoking – Cigarette Use, Smoking – Household Smokers, and Smoking – Recent Tobacco Use were merged and used for analysis. The models used in this research were determined by using a logistic regression model on the following variables in our dataset:

1. Female: Sex of observed person.

2. Family History of Asthma: Whether a person in the observed person's family has asthma.

3. Bad Air Quality: Whether the observed person has changed their activity in a day based on poor air quality.

4. Overweight: Whether the observed person has being told he or she is overweight by a doctor

5. Daily Activity: How many hours the observed person is active at work or recreationally in the day. Proxy variable

6. Years Smoked: How many years the observed person has smoked.

7. Industrial Dust and Fumes: Whether the observed person is exposed to industrial dust and fumes at work or home.

Some variables that medical literature suggests are predictive of asthma status like frequent respiratory illness were not included in the data set and regression due to their omission in the NHANES surveys. Upon analysis, other variables like Industrial Dust and Fumes status were found to return "yes" values for each observation and were excluded from the final model. To effectively evaluate the likelihood of asthma status given the certain indicators, several behavior and environmental variables in the data had to be recoded as proxy variables such as the Daily Activity, Years Smoked, and Bad Air Quality. In this process, unfortunately, Second Hand Smoking status was dropped because its recode values were not returning the correct values. In addition to these proxy variables, the logistic regression also included controls for demographic and biological factors like Family History of Asthma, Overweight, and Gender. After excluding incomplete cases, the final data set contained 6307 observations. Three logistic regression models were used to test the relationship between asthma status: a control model using only demographic factors, a full model using all the variables, and a final model including on statistically significant variables.

4. Results:

A univariate analysis of asthma status and its predictor factors reveals that only 842 of the total 6307 observations reported having asthma or 15.83% of the total sample has asthma. This is interesting as the sample has roughly 10% less positive responses for asthma than the American population does according to the CDC. The dataset includes nearly same number of women as men with 3193 and 3114 observations respectively. 755 people were affected by bad air quality whereas 5552 people who were not affected by bad air quality. The dataset also reveals that 2008 of the 6307 observations were diagnosed as overweight by his or her doctor.

The three logistic models used to analyze the data as are follows:

Control Model: $\text{logit}(\text{Asthma}=1) = \beta_0 + \beta_1\text{Female} + \beta_2\text{BadAir} + \beta_3\text{Overweight} + \beta_4\text{Family} + u$.

Full Model: $\text{logit}(\text{Asthma} = 1) = \beta_0 + \beta_1\text{Female} + \beta_2\text{BadAir} + \beta_3\text{Overweight} + \beta_4\text{Family} + \beta_5\text{Active} + \beta_6\text{YearsSmoked} + \beta_7\text{Dust} + u$.

Final Model: $\text{logit}(\text{Asthma}=1) = \beta_0 + \beta_1\text{BadAir} + \beta_2\text{Overweight} + \beta_3\text{Family} + u$

The coefficients derived from the regression returned the following values.

	Model Results		
Variables	**Control Model**	**Full Model**	**Final Model**
Intercept	0.100408***	0.09214***	0.0729357 ***
Female	0.02176*	0.01856	-
Bad Air Quality	-	0.08669***	0.0972818***
Overweight	0.078134***	0.07545***	0.0595751***
Family History	0.019886***	0.01955***	0.1573142***
Hrs Daily Activity	-	0.0002035	-
Dust	-	NA	-
Years Smoked	-	-0.00004921	-
C-Index	0.6093509	0.6408967	0.6607311

The regression on the control model indicates that each control variable is statistically significant at the 1% critical level with the exception of gender status, which is only significant at the 10% critical level. With a reported c-index of 0.6093509, this model is weakly predictive of asthma status. As the control model on biological and demographic risk factors returned statistically significant variables, forward selection was implemented to include all variables into the model.

The full model regression yielded a number of interesting results. The coefficient on female was no longer statistically significant at the 1% level. Likewise, the added behavioral variables on hours of daily activity and years smoked were found to lack statistical significance. One source of confusion with this model stemmed from the variable on industrial dust and fumes returning an “na” value. Upon further inspection, all values of the dust variable equaled 1, indicating that all observed people had some kind of industrial dust or fume exposure in their lifetime and that it would need to be omitted from the model. Another source of intrigue was that despite containing a number of insignificant variables, the c-index of 0.640897 was higher that the control model's c-index of 0.6093509.

Given the results from the full model regression, a final model was created regressing asthma status on only statistically significant variables. Each included variable was again found to be statistically significant. As expected, this model was the most predictive of asthma status with a c-index of 0.6607311.

5. Discussion:

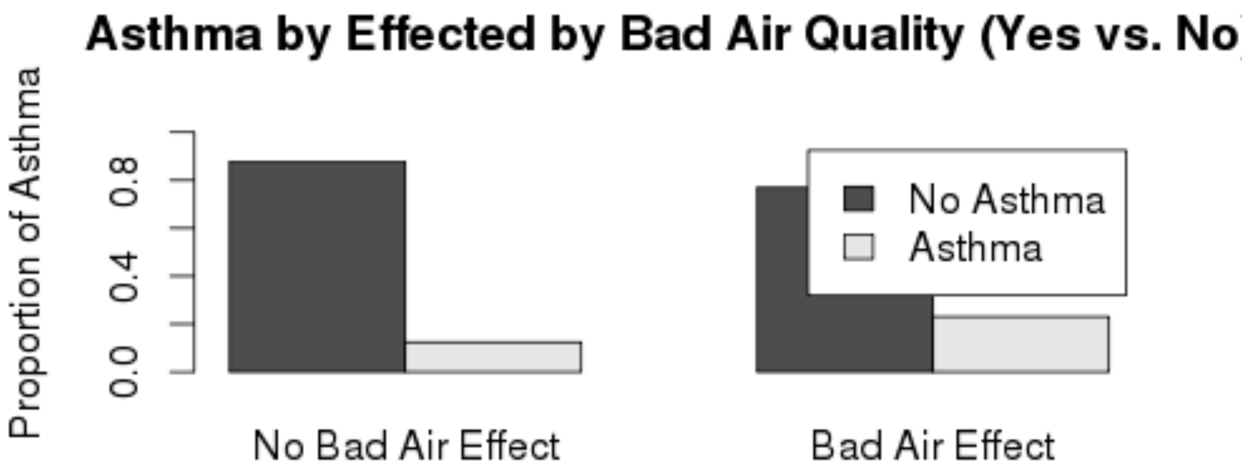
The final regression model reveals that the strongest predictors of asthma status are changing one's activities due to bad air quality, being overweight, and a family history of asthma. An analysis of the coefficient values yielded in the final model shows that β_0 equals 0.0729357, which implies that the observed sample has a 7.29357% likelihood of having asthma regardless of their status on the other included variables. The bad air quality variable returned a β_1 value of 0.0972818, implying that those who did change their activities due to bad air quality conditions are 9.72818% more likely to have asthma. The overweight status variable returned a β_2 value of 0.0595751, indicating those who have been diagnosed as overweight are 5.95751% more likely to have asthma. Finally, the coefficient on family history of asthma returned a β_3 value of 0.1573142, which means that members of the observed population whose family member have asthma are 15.73142% more likely to have asthma as well. As the beta values indicate, family history of asthma is the most significant variable for predicting asthma status.

By placing the final model outcomes in context, uncertainty arises about claiming a causal relationship between these predictive variables and asthma status. While it is possible that living in an area with poor air quality and choosing to change one's daily activities is predictive of asthma status, it is also reasonable to assume that people with asthma are more likely to change his or her activities due to poor air quality and those who do not. The potential for this suggests that the model may suffer from bias. Further analysis suggests that this may be the case for overweight status. In regards to the overweight status variable, it is possible that people who are overweight are more likely to suffer from asthma, it is also possible that those with asthma will become overweight as he or she might be limited from exercising. Given the constraints of the data and the model, specifically in regard to when an individual was diagnosed with asthma and being overweight, we cannot statistically account for this possibility.

6. Conclusion:

The complexities incurred while dealing with this data set serve a strong example of the difficulty associated with bio-statistical modeling. Without having a survey designed for an exact model, questions of bias in the model's finding arise. Future efforts to address the topic of asthma status must heed the bias concern by accounting for the impact of time of diagnosis. Though the final model returns a c-index suggesting weak to medium predictive capability, the model overall certainly has room for improvement.

8. Figures:


Asthma by Effected by Bad Air Quality (Yes vs. No)
Proportion of Asthma
0.0
0.4
0.8
No Asthma
Asthma
No Bad Air Effect
Bad Air Effect

Bibliography:

Center for Disease Control, “Asthma in the US”, (2011) http://www.cdc.gov/vitalsigns/asthma.

Center for Disease Control, “Fast Stats: Asthma”, (2012) http://www.cdc.gov/nchs/fastats/asthma.htm.