Econometrics Final Project

Respiratory Diseases: Nature vs. Nurture

By Hussain Raza, Maclain Pfeiffer, Yoga Ramachandran

December 6th, 2020

Introduction:

Throughout human history, respiratory diseases have been one of the leading causes of death and disability in the world. On a global scale, nearly 334 million people suffer from asthma, and of that group, 14% of which are children[1]. Lung cancer, with an average 1.6 million people diagnosed each year, is the leading cause of malignant respiratory disease and is the cause of 15% of all cancer related deaths. Past research has focused on the relationship between respiratory diseases and potential causes like smoking and air pollution, emphasizing the effects of particulates on the human body for determining causation. Some research has also focused on the presence of respiratory illness in the workplace, investigating how specific jobs and working conditions affect the risk of illness in an individual, or factors such as physical activity and BMI, factors which would not seem to affect respiratory illness on the surface, but have an effect nonetheless.

While many factors have been listed as catalysts towards respiratory health, many official health organizations tend to only focus on singular factors as causation of an illness, such as attributing many respiratory illnesses as being caused by smoking and air pollution, without listing other factors contributing to illness. In a way, it diminishes the work that many researchers have completed and ignores their contributions. To rectify this error, our goal in this study is to accumulate and determine the defining risk factors in everyday life which most commonly contribute to respiratory diseases, using a nationwide sample across the United States as an informative measure on behaviors and habits impacting the risks and dangers related to respiratory disease development. A common public policy to curb harmful, but not illegal, behavior is to impose sin taxes on activities deemed dangerous to public health. However, through our research we hope to examine whether respiratory diseases are largely caused by behavioral factors, and therefore support policies such as sin taxes, or whether environmental factors have more of a bearing. The data will indicate if it is better to focus on the individual tendencies or the environment we build in order to minimize the risk of respiratory disease.

With our research query leading us to determine whether the risk of respiratory illness in individuals is affected more by environmental or behavioral factors, we looked to investigate a large sample of real world data for comparison purposes, settling on IPUMS NHIS survey data collected over several years. After using a combination of logit/probit models we found that, contrary to our initial hypothesis, environmental variables such as vapors and particulates at work and living with residential smokers, affect the risk of developing a respiratory disease more than behavioral factors. Due to factors outside of one's control, such as genetics and air pollution, it is highly encouraged to avoid environmental risks such as regions of poor air quality to minimize one's risk of contracting a respiratory disease.

Literature Study:

In order to develop a general awareness of possible risk factors for investigation, we began our study by accumulating scholarly articles related to respiratory diseases and how each

[1] Forum of International Respiratory Societies. The Global Impact of Respiratory Disease – Second Edition. Sheffield, European Respiratory Society, 2017.

was affected by human behaviors and environmental features. In most instances, we found articles related to smoking, environmental hazards, and health behaviors. According to Anil Mukherjee and Zhongjian Zhang[2], asthma is described as a chronic inflammatory disease resulting from a combination of genetic and environmental factors. The paper highlights the genetic predisposition of individuals to develop allergic reactions to substances like pollen, antibiotics and perfumes, that are not known to cause immune response. It is underlined that environmental factors like tobacco smoke, air pollutants, allergens in the air, and diet can also trigger allergic asthma in genetically susceptible individuals. From these observations, we can glean the assumption that given a family history of respiratory conditions due to genetics or past decision-making, genetics can impact one's risk of developing a respiratory condition.

According to John Mullahy and Paul Portney[3], individual respiratory health is a function of cigarette smoking (cigarettes smoked per time period), environmental variables (air pollution, climatological factors), individual influences and other unmeasurable individual factors. The authors cite that if smoking is negatively correlated with one's vulnerability to respiratory illness and if the respiratory health is correlated with one of the unobserved factors, then the partial effect of smoking on respiratory health would be biased. The paper also pointed to the choice not to smoke as a variable positively affecting respiratory health. The Forum of International Respiratory Societies[4] also found that some respiratory diseases such as lung cancer are avoidable given one has control over their own tobacco usage, as well as attempting to reduce risk factors like air pollution. Some authors have even suggested that tobacco smoke including passive (second hand) smoke exposures is the leading cause of respiratory disease along with air pollution and workplace exposure to unsafe air, as around 2 billion people are exposed to indoor as well as outdoor air pollution.[5] And due to the growing concern of climate change created by human activity, air pollution becomes more prevalent and widespread, which may lead to higher numbers of patients with respiratory illnesses.[6] While the statements and research provided are not new when discussing the topic of respiratory illnesses, these variables listed can be used within our analysis to further this study's credibility and help to bolster the evidence of our findings.

In addition, BMI and overall health factors, such as abdominal fat, can contribute as risk factors towards developing respiratory illnesses. Even though BMI as such does not indicate the specific distribution of fat in the body, BMI is generally recognised as an indicator of health risk. An increase in BMI can lead to reduced lung capacity as well as impaired exercise capacity, both

---

[2] Mukherjee, Anil B., and Zhongjian Zhang. "Allergic Asthma: Influence of Genetic and Environmental Factors." *Journal of Biological Chemistry*, 23 Sept. 2011, www.jbc.org/content/286/38/32883.short.

[3] Mullahy, John, and Paul R. Portney. "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health." *Journal of Health Economics*, North-Holland, 5 Mar. 2002, www.sciencedirect.com/science/article/pii/016762969090017W.

[4] Forum of International Respiratory Societies. The Global Impact of Respiratory Disease – Second Edition. Sheffield, European Respiratory Society, 2017.

[5] Ferkol, Thomas, and Dean Schraufnagel. "The Global Burden of Respiratory Disease." *Annals of the American Thoracic Society*, vol. 11, no. 3, 2014, pp. 404–406., doi:10.1513/annalsats.201311-405ps.

[6] D'Amato, Gennaro, et al. "Climate Change, Air Pollution and Extreme Events Leading to Increasing Prevalence of Allergic Respiratory Diseases." *Multidisciplinary Respiratory Medicine*, vol. 8, no. 1, 2013, p. 12., doi:10.1186/2049-6958-8-12.

contributing to increased risk of respiratory illness. And although a relationship between obesity and asthma exists, as stated by Magali Poulain and his group, it remains to be ascertained if the association is causal or by chance.[7] Some studies have also pointed out that health behaviors and habits done by individuals can have adverse effects on one's risk of respiratory health, and given a higher amount of health knowledge, individuals can reduce their own risks of obtaining a respiratory disease, infectious or not[8]. This inference on risk leads us to believe that with lower BMIs and better health behaviors such as more exercise or physical activity can reduce the risk of respiratory disease, as higher amounts of exercise and physical activity causes more utilization of respiratory pathways as oxygen is utilized more frequently by the human body.

Parallels drawn from another research[9], emphasize the importance of occupational history while evaluating the respiratory outcomes of patients with asthma. According to Kjell Torén and Paul Blanc, multiple population analyses indicate that 16.3% of all adult cases of asthma were caused by occupational exposures, with a 17.3% attributable risk over the general population of being afflicted with asthma. In addition, respiratory risk factors can be attributed to and traced to specific working industries. As noted by Roel Vermeulen and his associates[10], employment within construction or production industries such as the metals industry or the plastics industry have a positive correlation with individual cases of bronchitis and asthma, in addition to the individual time spent working in specific industries. J.C. McDonald's studies further corroborate these findings, stating that the heightened risks in those industries relate to the particulates and working conditions common throughout, such as metals, grains, wood dusts, solders, and welding fumes[11]. It is safe to assume that if these particulates are prevalent or can be replicated in other industries and working environments, then the risk for respiratory disease can be increased, thus causing occupation to become another possible and viable risk factor. However, we should also take the presence of second-hand smoke into account from other coworkers, as the harmful vapor content at work can contain both second-hand smoke inhalation and work-related particulates.

In this study, data is collected from IPUMS, a U.S public data record containing economic, health, and personal survey information collected from across the country. Many of the papers we found do not mention or utilize IPUMS data, which provides us with a unique perspective and opportunity to test previous research. In performing our study, we are looking to prove the validity of previous papers on a broader scale, using a large random sample of the country's inhabitants to discover how well specific variables correlate with respiratory disease

---

[7] Poulain, Magali, et al. "The Effect of Obesity on Chronic Respiratory Diseases: Pathophysiology and Therapeutic Strategies." *CMAJ*, CMAJ, 25 Apr. 2006, www.cmaj.ca/content/174/9/1293.full.

[8] Sun, Xinying, et al. "Determinants of Health Literacy and Health Behavior Regarding Infectious Respiratory Diseases: a Pathway Model." *BMC Public Health*, vol. 13, no. 1, 2013, doi:10.1186/1471-2458-13-261.

[9] Torén, Kjell, and Paul D Blanc. "Asthma Caused by Occupational Exposures Is Common – A Systematic Analysis of Estimates of the Population-Attributable Fraction." *BMC Pulmonary Medicine*, vol. 9, no. 1, 2009, doi:10.1186/1471-2466-9-7.

[10] Vermeulen, Roel, et al. "Respiratory Symptoms and Occupation: a Cross-Sectional Study of the General Population." *Environmental Health*, vol. 1, no. 1, 2002, doi:10.1186/1476-069x-1-5.

[11] McDonald, J C, et al. "Incidence by Occupation and Industry of Acute Work Related Respiratory Diseases in the UK, 1992-2001." *Occupational and Environmental Medicine*, vol. 62, no. 12, 2005, pp. 836–842., doi:10.1136/oem.2004.019489.

risks. By pursuing this objective, our results may be very similar to those of previous papers, and may not provide more unique insights other than reaffirming those already mentioned contributions or disproving their overall significance.

Data Description:

We identified NHIS (National Health Interview Survey) data from IPUMS for 2010 and 2018 specifically for this analysis[12]. The NHIS is a large national representative survey of the civilian population of the United States of America that has been conducted since 1963. Briefly, NHIS is a household survey containing longitudinal data collected continuously throughout the year. For these chosen survey years, about 43,000 households and about 46,000 households were sampled for the years 2010 and 2018, respectively, with about 86,000 relevant observations in each sample year. The data available for respiratory diseases includes chronic bronchitis, asthma, lung cancer, larynx cancer, and mouth/tongue/lip cancer. However, our collected data is significant, due to the fact that our sample is dealing with an immense population size, which would chart in the millions for collected data points. Due to the complexity and time constraints needed to create an accurate model with millions of data points, a much smaller sample size was taken in order to make an inference and informed decision on the entire population. IPUMS solves the sampling problem by surveying for the explicit purpose of determining correlations between variables, and collecting a much smaller sample in comparison to the population being investigated.

```
> stargazer(data.df, type = 'text')

================================================================================
Statistic              N      Mean   St. Dev.   Min   Pctl(25) Pctl(75)   Max
--------------------------------------------------------------------------------
Disease              86,737 0.024    0.153      0       0        0        1
HEALTH               86,737 2.133    1.054      1       1        3        5
BMI                  86,737 8.386    14.896   0.000   0.000    21.490   99.990
SmokingAge           86,737 87.276   24.643     6       96       96       96
HomeSmokeExposure    86,737 1.033    0.180      1       1        1        2
WeekdaySittingHours  86,737 2.046    9.480      0       0        2        99
DailyHrsSleep        86,737 2.118    4.258      0       0        6        99
WorkVaporExposure    86,737 0.230    0.525      0       0        0        9
Exercise             86,737 0.716    2.210      0       0        0        28
AGE                  86,737 35.251   22.417     0       16       52       85
MotherLungCan        86,737 0.071    0.464      0       0        0        9
FatherLungCan        86,737 0.074    0.453      0       0        0        9
--------------------------------------------------------------------------------
```

[12] Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King and Kari C.W. Williams. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [https://www.nhis.ipums.org/]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D070.V6.4

Respiratory disease criteria

IPUMS NHIS has data on hand for a host of different types of diseases, of which we selected those that were related to the lungs and general respiratory system. Binary data was collected for the relevant diseases, including asthma, bronchitis, and cancers of the lung, larynx, mouth, tongue, and lip. In addition, discrete numerical recordings were used to determine the age at which the survey participants had contracted the relevant disease, if at all. A respondent was categorized as having a respiratory disease based on their aggregate answers to the disease questions. Positive responses were qualified by any indication of disease, which tallied to 2073 cases out of 86,737 respondents. Only respondents who answered with a positive or negative response to the respiratory disease questions were included in the calculations. Individuals who chose not to respond or who did not have an answer were filtered out from the data.

Health, BMI, Sleep

Health is a self-reported variable and thus, is subject to response bias tendencies. Answers were recorded on a five-point scale ranging from “excellent” (1) to “poor” (5). The distribution is heavily skewed right with over 50% of the data consisting of “excellent” to “very good”. Unknown responses were filtered out.

Body Mass Index (BMI) is a feature based on metrics regarding respondent weight and height. The index is meant to measure the weight to height ratio in order to determine an obesity coefficient. IPUMS uses the following formula:

BMI = [Weight(kg)] / [[Height(m)]squared]

The above formula has remained in consistent use over time, with exceptions for 1991, 1993, and 1995. Data for this feature was left unfiltered due to a normal distribution of values across respondents.

Sleep is another self-reported value that is largely consistent, but has had moments of shifting survey standards. Reliability is assured as outliers of unexpectedly high numbers are accounted for by those of advanced age and low health.

Physical Activity

Physical activity is a feature derived from (i) the time period of moderate physical activity that lasted at least ten minutes and (ii) the number of units of those sessions per week. The distribution of this variable is skewed in favour of those who do not exercise, which is an accurate reflection on the population tendency to not exercise.

The flip side of exercise is the amount of time spent sitting in place, and our data shows that the average adult spends about two hours sitting outside of work. Given the number of non-work hours in the day are limited, this is a high proportion of inactive time.

Smoke and Smokers

The first smoke factor is the behavioral choice to smoke cigarettes, which we have recorded as the number of cigarettes per day, the age respondent began smoking, and the frequency with which once currently smokes. Combining these aspects of individual behavior, we determined a new feature to track whether or not someone was a smoker.

The second smoke factor is whether the respondent lives in an environment affected by other smokers. In 2005 the question was altered to specify whether any of other residents smoked "in a usual week". The data here overlaps with the distribution of smokers against non-smokers with marginally fewer respondents indicating that they live with a smoker than those who said they smoke.

The third smoke factor is an environmental factor meant to measure the effect of dangerous fume exposure at work. Collected in 2010, this variable covers any employment an adult respondent has had over the twelve months leading up to the survey.

Genetics

In 2000 family history collection with respect to parental lung cancer was started and resulted in data for both the biological mother and father. Our data tracks whether or not either biological parent had lung cancer. Minimal data transformations were required by our team as the survey had begun programmatically filling in 'not ascertained' responses.

Economic Strategy:

The hypothesis we will be testing is that behavioral factors such as smoking have the greatest impact on the likelihood of contracting a respiratory disease. It is motivated by the general advocacy of public policy against behaviors such as smoking and in favour of physical exercise. However, there are constant environmental factors outside of our control that also affect the chance of getting a respiratory disease. Our research will address a host of different factors in order to determine if smoking maintains the highest influence.

The data we have on hand has a binary variable of interest, which means that we will have to use a logistic regression model to estimate the likelihood of contracting a respiratory disease. Linear regression, while it can be estimated, will not bear any fruits for our analysis because all our data points will plot themselves on one of two horizontal lines at the values of 0 and 1. The linear model will then fit itself along some diagonal across the data or neatly along $y = 0$ or $y = 1$.

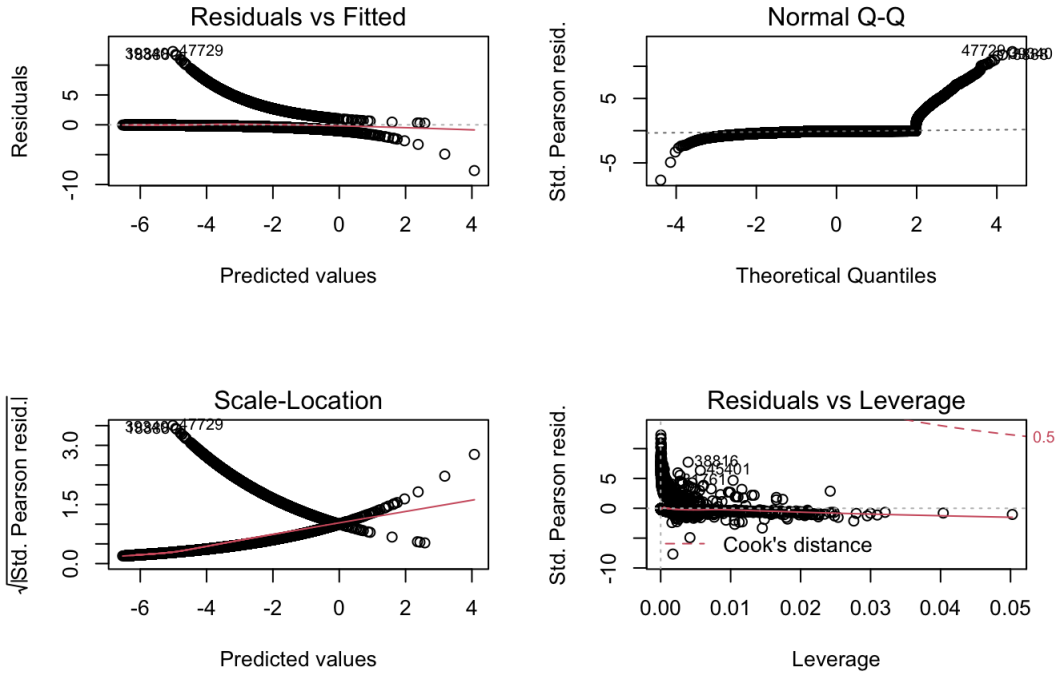
Logistic regression (logit and probit) are best accustomed to dealing with binary dependent variables and thus best suited to helping determine correlation between related environmental and behavioral factors. More specifically, logistic regression is useful to predict the presence or absence of a characteristic based on a set of explanatory variables by regressing a binary dependent variable on a set of independent predictor variables. In our case, the model will predict the probability of contracting a respiratory disease.

In collecting our data, we collected as many respiratory disease variables as possible to determine correlation. The NHIS data from 2010 and 2018 contains information on several

respiratory diseases, each disease containing at least 1 variable to denote the presence of the disease in an individual. We took specific defining variables of a disease's presence, such as survey answers where individuals were told by a health professional if they had chronic bronchitis in the last 12 months, and accumulated all into a single binary variable which denoted if someone possessed a respiratory disease from the survey population. We then collected risk factor variables, using the aforementioned research articles as a reference for variable suitability in the model. After collecting our risk factor variables, such as variables for BMI and predisposition to smoking, we filtered and removed those variables which gave similar or identical information to our main choice variables to analyze.

In order to ensure our model was as accurate as possible, we also needed to remedy any heteroskedasticity and multicollinearity problems in our model, as either problem can increase the declared significance of a variable and thus give false results. To address heteroskedasticity in our logit model, which is defined as the unequal spread of data points or residuals, we employed a regression diagnostic plot to check for heteroskedasticity, as showcased in Figure 1. The regression diagnostic plot in Figure 1 provides an almost straight horizontal line without much deviation. This proves the absence of heteroskedasticity in the regression analysis.

Figure 1: Regression Diagnostic Plot for Residuals vs. fitted values



To further understand and address heteroskedasticity, we also employed the use of a Breusch-Pagan test, which determines the presence of heteroskedasticity using p-values below 0.05 to declare an absence of heteroskedasticity in the model.

```
bptest(logit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  logit1
## BP = 7834, df = 15, p-value <2e-16
```

We will approach multicollinearity using a tiered process. First we will identify variables with high levels of correlation, at least 0.80. Then we will use the variance inflation factors function (vif) in the car package in R to determine the ratio of our multiple regression model against a single variable regression model. If the coefficients of the vif analysis are less than five, then we will assume there is no multicollinearity.

Results:

Linear Regression

A straight line cannot accurately predict the change in probability according to a unit increase in an explanatory variable because it is in fact predicting the result. As shown in Figure X, the linear regression mostly predicts implausible real life scenarios between the scenarios of having and not having a disease.

The linear model was attempted with the final variables after addressing multicollinearity in the logistic regression. Consequently, every variable is significant at very minimum the 95% confidence level. However, the ultimate indication of failure is represented by the very low R-squared value of 0.076, meaning that only 7.67% of the population behavior can be explained by the model.

Figure 2: Linear Regression Summary

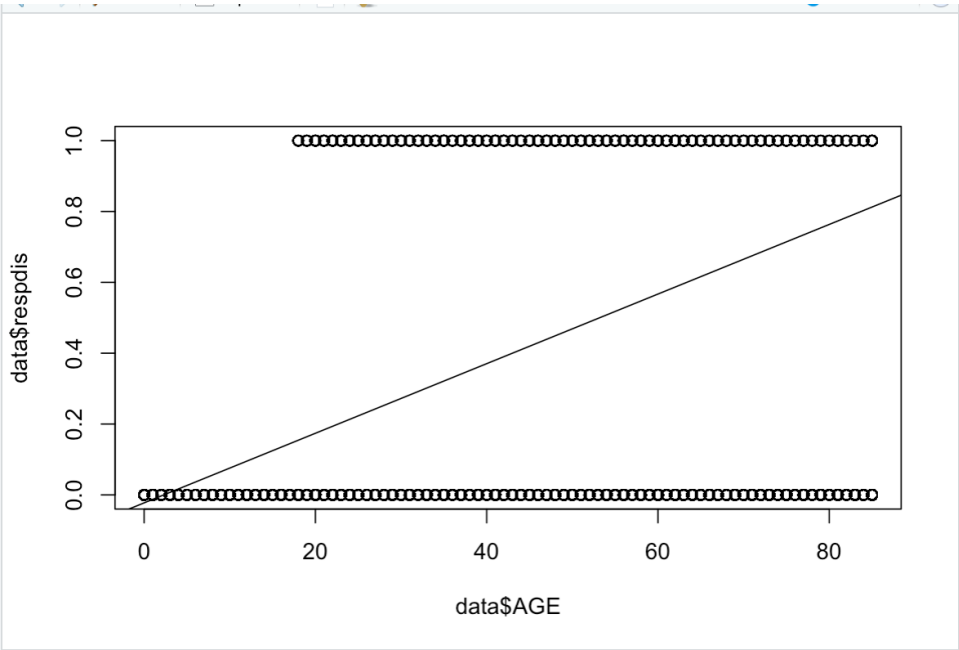
```
Call:
lm(formula = respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES +
    SITWKDAYHR + HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN +
    BFLGCAN, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34147 -0.02988 -0.00258  0.00753  0.98463

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.251e-02  4.685e-03  -4.805 1.55e-06 ***
HEALTH       9.825e-03  5.125e-04  19.169  < 2e-16 ***
BMI          1.237e-03  5.606e-05  22.056  < 2e-16 ***
SMOKAGEREG  -3.752e-04  2.536e-05 -14.796  < 2e-16 ***
SMOKHOMRES   4.153e-02  3.067e-03  13.543  < 2e-16 ***
SITWKDAYHR   4.157e-04  5.640e-05   7.370 1.73e-13 ***
HRSLEEP     -4.400e-04  1.716e-04  -2.564 0.010337 *
WEXVAPOR     2.518e-02  1.249e-03  20.160  < 2e-16 ***
MOD10FWK     9.261e-04  2.578e-04   3.592 0.000328 ***
AGE         -7.171e-05  2.520e-05  -2.846 0.004432 **
BMLGCAN      5.278e-03  1.110e-03   4.753 2.01e-06 ***
BFLGCAN      9.203e-03  1.141e-03   8.063 7.52e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1468 on 86725 degrees of freedom
Multiple R-squared:  0.07647,   Adjusted R-squared:  0.07636
F-statistic: 652.8 on 11 and 86725 DF,  p-value: < 2.2e-16
```

Figure 3: Linear Regression Model



Logistic Regression

A 20-factor predictor logistic model was fitted to the data to test the research hypothesis regarding the relationship between the likelihood of a respiratory disease and various independent factors like smoking, health, age of the individual, occupation, physical activity etc. According to our model, the logarithmic odds of an individual developing respiratory disease are positively correlated to overall health status, BMI, hours of sleep, hours sitting outside work, exposure to vapors, gas and dust at work, the presence of residential smokers, and parental history of having a respiratory disease. This means that for variables such as BMI, higher values indicate a higher risk to develop a respiratory disease. In the case of our overall health status variable, as it changes from excellent (a value of 1) to poor (a value of 5), the likelihood of individuals developing a respiratory disease surge. In contrast, respiratory disease risks were negatively related to age at which respondents started smoking regularly and age at which parents were diagnosed with a respiratory disease. Given all other factors being held constant, the earlier an individual started smoking, the more likely they would develop a respiratory disease.

Figure 4: Logistic Regression model after adjusting for multicollinearity

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.166  -0.181  -0.104  -0.079   3.129
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.243807   0.143958  -43.37  < 2e-16 ***
## HEALTH       0.427677   0.022943   18.64  < 2e-16 ***
## BMI          0.028770   0.001144   25.16  < 2e-16 ***
## SMOKAGEREG  -0.008399   0.000684  -12.29  < 2e-16 ***
## SMOKHOMRES   0.461797   0.068328    6.76  1.4e-11 ***
## SITWKDAYHR   0.006754   0.001279    5.28  1.3e-07 ***
## HRSLEEP      0.021091   0.002992    7.05  1.8e-12 ***
## WEXVAPOR     0.781782   0.032586   23.99  < 2e-16 ***
## MOD10FWK     0.054622   0.006043    9.04  < 2e-16 ***
## AGE          0.009367   0.001398    6.70  2.1e-11 ***
## BMLGCAN      0.071678   0.023940    2.99   0.0028 **
## BFLGCAN      0.122185   0.023330    5.24  1.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19577  on 86736  degrees of freedom
## Residual deviance: 14939  on 86725  degrees of freedom
## AIC: 14963
##
## Number of Fisher Scoring iterations: 8
```

Figure 5: Probit Regression Summary

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.864  -0.176  -0.082  -0.056   3.178
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.178522   0.069292  -45.87  < 2e-16 ***
## HEALTH       0.189879   0.010890   17.44  < 2e-16 ***
## BMI          0.015559   0.000622   25.02  < 2e-16 ***
## SMOKAGEREG  -0.003743   0.000331  -11.30  < 2e-16 ***
## SMOKHOMRES   0.237818   0.035466    6.71  2.0e-11 ***
## SITWKDAYHR   0.003419   0.000662    5.17  2.4e-07 ***
## HRSLEEP      0.012261   0.001692    7.25  4.3e-13 ***
## WEXVAPOR     0.369689   0.016213   22.80  < 2e-16 ***
## MOD10FWK     0.025264   0.003137    8.05  8.1e-16 ***
## AGE          0.003548   0.000642    5.53  3.2e-08 ***
## BMLGCAN      0.036198   0.012660    2.86   0.0042 **
## BFLGCAN      0.064665   0.012479    5.18  2.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19577  on 86736  degrees of freedom
## Residual deviance: 14575  on 86725  degrees of freedom
## AIC: 14599
##
## Number of Fisher Scoring iterations: 8
```

Overall evaluation of the logistic model

Unfortunately, a normal adjusted R^2 doesn't function well with logistic models, as it cannot explain the proportion of variance well as compared to linear models. Instead, we utilized a pseudo R^2 to explain the variance in our model and determine if the model would be suitable for our research. The pseudo R^2 of 0.24 retrieved from our model helps prove that the model is a good fit for population data, as the model will only fit data points existing at our binary values of 0 and 1, due to the nature of a logit/probit model.

```
PseudoRsq<-pR2(logit2)
```

```
## fitting null model for pseudo-r2
```

```
PseudoRsq
```

```
##      llh  llhNull       G2 McFadden     r2ML     r2CU
## -7.5e+03 -9.8e+03  4.6e+03  2.4e-01  5.2e-02  2.6e-01
```

To test our individual predictors, we used a Wald Chi-Square statistic test to test the significance of every regression coefficient. To illustrate, Table 3 includes two predictors, BMI and the age when smoking regularly began (SMOKAGEREG). Both these predictors show statistical significance in the outcome of respiratory disease, as shown by their low p-values. The remaining predictors also follow similar significance values when used in the Wald Chi-squared test.

Figure 6: Wald Test Chi-Square statistics summary

```
regTermTest(logit2,'BMI')
```

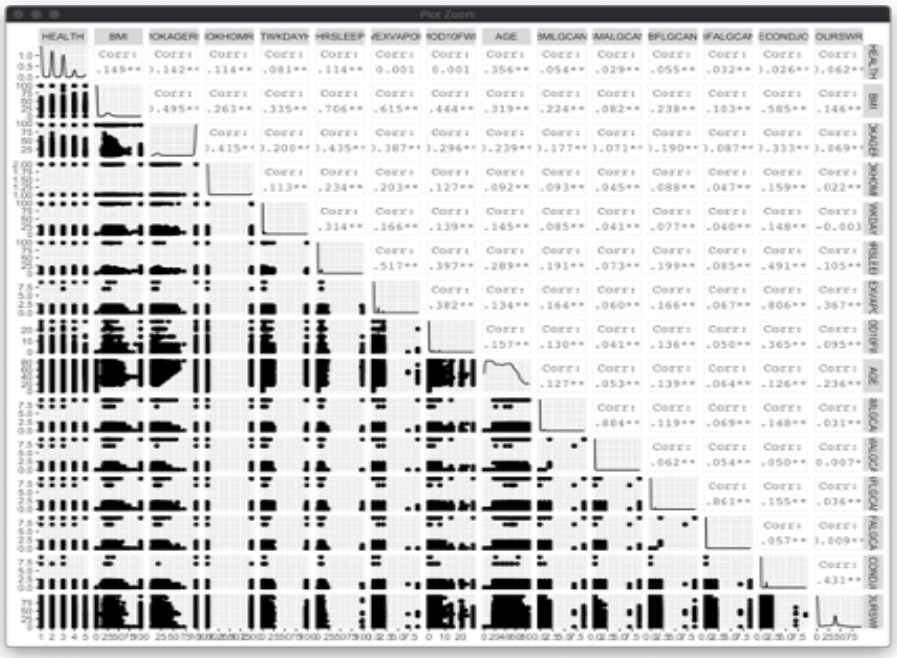
```
## Wald test for BMI
##  in glm(formula = respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES +
##      SITWKDAYHR + HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN +
##      BFLGCAN, family = binomial(link = "logit"), data = data)
## F =  633  on  1  and  86725  df: p= <2e-16
```

```
regTermTest(logit2,'SMOKAGEREG')
```

```
## Wald test for SMOKAGEREG
##  in glm(formula = respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES +
##      SITWKDAYHR + HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN +
##      BFLGCAN, family = binomial(link = "logit"), data = data)
## F =  151  on  1  and  86725  df: p= <2e-16
```

Upon narrowing the scope of our explanatory variables to just the significant ones, we began to look for multicollinearity. To deal with multicollinearity, we utilized a correlation matrix on our explanatory variables to find and remove concerning variables which could prove problematic to our model. As shown by Figure 2, correlation values of 80% were considered as the threshold for removing any problematic variables. In compliance with this guideline, variables indicating age at which parents were diagnosed with a respiratory disease, second job, hours of work were removed from the model. After accounting for multicollinearity in our model, the model was optimized.

Figure 7: Correlation matrix for explanatory variables



Along with our correlation matrix, we also employed a variance inflation factor(VIF) test to measure how much variance in our variables was being affected by multicollinearity in the model. As declared by the VIF test, any variable which has a VIF value closer to 1 indicates there are no significant multicollinearity problems with the variable[13] This test was performed and found no significant multicollinearity issues within our variables at large.

```
car::vif(logit2)
```

```
##     HEALTH        BMI SMOKAGEREG SMOKHOMRES SITWKDAYHR    HRSLEEP   WEXVAPOR
##        1.2        1.2        1.3        1.2        1.1        1.1        1.2
##   MOD10FWK        AGE    BMLGCAN    BFLGCAN
##        1.1        1.2        1.0        1.0
```

We also utilized a log-likelihood ratio test to help us demonstrate whether the model is better fit to the data compared to the first model with multicollinearity present. A p-value of less than 0.05 for the model provides evidence in favor of the former model.

```
lrtest(logit1, logit2)
```

```
## Likelihood ratio test
##
## Model 1: respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES + SITWKDAYHR +
##     HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN + BMALGCAN +
##     BFLGCAN + BFALGCAN + SECONDJOB + HOURSWRK
## Model 2: respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES + SITWKDAYHR +
##     HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN + BFLGCAN
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1  16  -7387
## 2  12  -7469 -4   164     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(logit1,logit2, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES + SITWKDAYHR +
##     HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN + BMALGCAN +
##     BFLGCAN + BFALGCAN + SECONDJOB + HOURSWRK
## Model 2: respdis ~ HEALTH + BMI + SMOKAGEREG + SMOKHOMRES + SITWKDAYHR +
##     HRSLEEP + WEXVAPOR + MOD10FWK + AGE + BMLGCAN + BFLGCAN
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     86721      14775
## 2     86725      14939 -4     -164   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[13] Yoo, Wonsuk, et al. "A Study of Effects of MultiCollinearity in the Multivariable Analysis." *International Journal of Applied Science and Technology*, U.S. National Library of Medicine, Oct. 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4318006/.

Assessment of predicted probabilities

In assessing the effects of each variable in our model, we investigated the predicted probabilities of each variable while holding all other values constant. Table 5 below shows the predicted probability of developing respiratory disease decreasing from 0.25 to 0.12 as a result of one's health status increasing from poor to good (5 to 3).

Figure 8: Predicted Probability for the Logistic Regression Model

```
predict(logit2, data.frame(HEALTH=5, BMI=35,
                           SMOKAGEREG=20,
                           SMOKHOMRES=1,SMOKWKEXPYR=1,
                           QUITYRS=0, SITWKDAYHR=2,
                           HRSLEEP=7,
                           WEXVAPOR=1,MOD10FWK=0,AGE=60,
                           BMLGCAN=1,BFLGCAN=1), type = 'response')
```

```
##    1
## 0.25
```

```
predict(logit2, data.frame(HEALTH=3, BMI=35,
                           SMOKAGEREG=20,
                           SMOKHOMRES=1,SMOKWKEXPYR=1,
                           QUITYRS=0, SITWKDAYHR=2,
                           HRSLEEP=7,
                           WEXVAPOR=1,MOD10FWK=0,AGE=60,
                           BMLGCAN=1,
                           BFLGCAN=1), type = 'response')
```

```
##    1
## 0.12
```

Similarly, keeping all the factors constant, increasing BMI from 35 to 40 increases the predicted probability of developing respiratory disease by 12%, from 0.25 to 0.28.

```
predict(logit2, data.frame(HEALTH=5, BMI=40,
                           SMOKAGEREG=20,
                           SMOKHOMRES=1,SMOKWKEXPYR=1,
                           QUITYRS=0, SITWKDAYHR=2,
                           HRSLEEP=7,
                           WEXVAPOR=1,MOD10FWK=0,AGE=60,
                           BMLGCAN=1,
                           BFLGCAN=1), type = 'response')
```

```
##    1
## 0.28
```

Odds Ratio

The odds ratio defines the ratio of probability of getting the success of an outcome over the odds of failure. For a unit increase in BMI, the odds of getting a respiratory disease changes by a factor of 1.02 compared to the odds of not getting a respiratory disease. For a unit increase in smoking age, the odds of getting a respiratory disease changes by a factor of 0.99.

```
exp(logit2$coefficients)
```

```
## (Intercept)      HEALTH         BMI  SMOKAGEREG  SMOKHOMRES  SITWKDAYHR
##      0.0019      1.5337      1.0292      0.9916      1.5869      1.0068
##     HRSLEEP    WEXVAPOR    MOD10FWK         AGE     BMLGCAN     BFLGCAN
##      1.0213      2.1854      1.0561      1.0094      1.0743      1.1300
```

Limitations:

Unfortunately, limitations are prevalent while using any model and must be acknowledged. In the case of a logistic model itself, the model can suffer from overfitting, meaning it fits more to the data and holds less predictive power over our variables. As a result, our model may not be well suited for predicting the presence of respiratory disease. However, the more pressing concern towards our results stem from the data sample being used. For example, the initial data from IPUMS contained thousands of NIU (not in universe) variables, or unknown value codes denoting values which were not recorded in the study. If these were handled improperly, the model could contain an imbalance of values, which can cause correlation measures to misrepresent and fail to properly classify a variable's relation towards respiratory disease. Some variables which were found within our data may also suffer from omitted variable bias, as variables which could affect the correlation or impact of a variable on respiratory disease risk may not be included within the data or our model. One such example of this is air pollution, which does affect multiple variables and their outcome on respiratory disease, yet is not included or noted within the NHIS survey data. Because there are variables which haven't been included within our data which may have correlation with other variables, we cannot correct for this bias if it exists in our model. In addition, because the data has been collected from real world participants of varying backgrounds with no control groups present, we can assume that some of our variation described with our model will lack explanation.

Next steps for this research would be to address the unexpected results for certain factors such as the average hours of sleep and moderate physical activity, both of which should decrease the chance of getting sick, but are causing a positive effect in the model. The omitted variable bias mentioned above could address the possibility of unregistered poor air quality negating the benefits of exercise. Another possible limitation could be simultaneous causality bias between the number of hours slept and disease because there is a positive relationship between age, disease, and sleep. Results that were surprisingly not significant and deserve further investigation are factors such as whether someone is a smoker, number of cigarettes per day, and occupation type. Smoking is said to be a leading cause of lung cancer in the United States, however our results indicated that secondhand smoke is more of a risk than being an active smoker. This could be a consequence of the high number of NIU responses.

The results of our research indicate that work exposure is by and large the leading cause of respiratory disease, increasing one's chances to over twice the likelihood by simply being exposed. Next in significance is the general health of the individual, followed by exposure to smoke at home, both of which increase the risk by about half. While health is partly reliant on good dietary and exercise habits, there are many factors outside an individual's control. Genetics, bacteria, and unfortunate accidents can affect anyone at any time and are consequences of our

environment, just as the exposure to smoke at work or at home. Given the strong case for environmental factors, it is our recommendation that individuals should limit their exposure to regions of poor air quality to minimize their risk of contracting a respiratory disease.

References:

D'Amato, Gennaro, et al. "Climate Change, Air Pollution and Extreme Events Leading to Increasing Prevalence of Allergic Respiratory Diseases." *Multidisciplinary Respiratory Medicine*, vol. 8, no. 1, 2013, p. 12., doi:10.1186/2049-6958-8-12.

Ferkol, Thomas, and Dean Schraufnagel. "The Global Burden of Respiratory Disease." *Annals of the American Thoracic Society*, vol. 11, no. 3, 2014, pp. 404–406., doi:10.1513/annalsats.201311-405ps.

Forum of International Respiratory Societies. The Global Impact of Respiratory Disease – Second Edition. Sheffield, European Respiratory Society, 2017.

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King and Kari C.W. Williams. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [https://www.nhis.ipums.org/]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D070.V6.4

McDonald, J C, et al. "Incidence by Occupation and Industry of Acute Work Related Respiratory Diseases in the UK, 1992-2001." *Occupational and Environmental Medicine*, vol. 62, no. 12, 2005, pp. 836–842., doi:10.1136/oem.2004.019489.

Mukherjee, Anil B., and Zhongjian Zhang. "Allergic Asthma: Influence of Genetic and Environmental Factors." *Journal of Biological Chemistry*, 23 Sept. 2011, www.jbc.org/content/286/38/32883.short.

Mullahy, John, and Paul R. Portney. "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health." *Journal of Health Economics*, North-Holland, 5 Mar. 2002, www.sciencedirect.com/science/article/pii/016762969090017W.

Poulain, Magali, et al. "The Effect of Obesity on Chronic Respiratory Diseases: Pathophysiology and Therapeutic Strategies." *CMAJ*, CMAJ, 25 Apr. 2006, www.cmaj.ca/content/174/9/1293.full.

Sun, Xinying, et al. "Determinants of Health Literacy and Health Behavior Regarding Infectious Respiratory Diseases: a Pathway Model." *BMC Public Health*, vol. 13, no. 1, 2013, doi:10.1186/1471-2458-13-261.

Torén, Kjell, and Paul D Blanc. "Asthma Caused by Occupational Exposures Is Common – A Systematic Analysis of Estimates of the Population-Attributable Fraction." *BMC Pulmonary Medicine*, vol. 9, no. 1, 2009, doi:10.1186/1471-2466-9-7.

Vermeulen, Roel, et al. "Respiratory Symptoms and Occupation: a Cross-Sectional Study of the General Population." *Environmental Health*, vol. 1, no. 1, 2002, doi:10.1186/1476-069x-1-5.

Yoo, Wonsuk, et al. "A Study of Effects of MultiCollinearity in the Multivariable Analysis." *International Journal of Applied Science and Technology*, U.S. National Library of Medicine, Oct. 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4318006/.