ORIE 3120: Final Report

Health Indicators in Big Cities Data Analysis

**Table of Contents**

Abstract …………………………………………………………………………………………... 2

[Introduction](https://docs.google.com/document/d/12e51OhLBlWMGYHWZUmdmjywI7FNDnIpF-sYTzIvWS3U/edit#heading=h.s6ppa8btox1p) ……………………………………………………………………………………..... 3

Research Questions ……………………………………………………………………………..... 3

Question1 ……………………………………………………………………………………..…. 4

Question [2](https://docs.google.com/document/d/12e51OhLBlWMGYHWZUmdmjywI7FNDnIpF-sYTzIvWS3U/edit#heading=h.h57avqy2gh8b) ………………………………………………………………………………………...5

Linear Regression ………………………...…………………………………………………6

Logistic Regression ……………………...…………………………………………..………6

Testing Assumptions ….…………………………………………………………..…………6

Constraints ………………………………………………………………………..…………6

Discussion ………...…………………………………………………………………………6

Question [3](https://docs.google.com/document/d/12e51OhLBlWMGYHWZUmdmjywI7FNDnIpF-sYTzIvWS3U/edit#heading=h.5rdieuagzfu7) ………………………………………………………………………………………...8

[Conclusion](#_k8tgjqgg3vm) ………………………………………………………………………….……………. 8

References ………………………………………………………………………….……………10

Appendix A ………………………………………………………………………….…………. 11

Appendix B ………………………………………………………………………….…………..15

**Figures and Tables**

[Figure 1](https://docs.google.com/document/d/18NDqlGVkdxs4z9m_iEIkarq_HfDhezLThFAVlJUi4c4/edit#bookmark=id.u7rr27hqi6wm) Causes of Death Across All Cities ………………….…………………….…………..4

[Figure 2](https://docs.google.com/document/d/18NDqlGVkdxs4z9m_iEIkarq_HfDhezLThFAVlJUi4c4/edit#bookmark=id.ryhfggv9zlm)  Lung Cancer Mortality vs. City ……………………………………………………….5

[Figure 3](https://docs.google.com/document/d/18NDqlGVkdxs4z9m_iEIkarq_HfDhezLThFAVlJUi4c4/edit#bookmark=id.2wm8znc68a08)  Heart Disease Mortality vs. City ……………………………………………………...5

[Figure 4](https://docs.google.com/document/d/18NDqlGVkdxs4z9m_iEIkarq_HfDhezLThFAVlJUi4c4/edit#bookmark=id.ae3rsqgdjmj8) Percent of Smokers vs. Lung Cancer ………………...……………..……………….. 6

## **Abstract**

Mortality rates in the United States are caused by a multitude of underlying factors, such as obesity, poor nutrition, and poverty. To decrease these rates, one must target the root of the issue. For example, although heart disease might be linked to obesity, one must understand what causes obesity in the first place, such as food deserts and a failure to self-regulate. By analyzing the health circumstances that influence causes of death, this report will determine the best strategy to minimize mortality rates across cities in the U.S. This will be accomplished through various data analysis methods, such as linear regression, logistic regression, and testing the assumptions of linear regression. The leading causes of death in the United States were determined to be heart disease, lung cancer, and diabetes. Using linear regression models, we investigated possible underlying causes and risk factors for these three diseases. The most significant model found lung cancer to be highly correlated with percent of adults who currently smoke, with an adjusted R-squared value of .812. We further investigated risk factors using logistic regression, creating a variable that represented whether a mortality rate was one standard deviation above the mean for that indicator. The logistic regression model for heart disease was the most significant, and it showed Chicago, Long Beach, Philadelphia, and Sacramento as being the cities most at risk. The report will present visualizations that will more clearly demonstrate the relationships discovered, such as scatter plots and maps of the cities with the various indicators.

*Keywords*: mortality, health, linear regression, logistic regression

**Introduction**

Our team is writing this report to advise the U.S. Department of Health and Human Services on how to allocate their resources in order to improve health and welfare across the biggest cities of America. Our analysis is based on the data from the Health Indicators in Big Cities dataset (Nursnaaz, 2019), which contains 34 health and demographic indicators on 26 of the largest cities in America. This data is provided by the U.S. Centers for Disease Control and Prevention. Using the machine learning techniques taught in ORIE 3120, we plan on generating data visualizations and using statistical modeling to gather insights and eventually present our conclusions to your department. The U.S. Department of Health and Human Services is an executive branch department that aims to protect the health of all Americans by providing essential health services. These services include health insurance, social services, prevention and wellness services, as well as public health and safety services. Their mission is to make healthcare equitable and accessible, especially for those who are most vulnerable.

The physical health and well-being of the citizens of the United States is our number one priority and we feel that by analyzing this data, we will be able to better understand the risks and issues associated with each city. Given that larger cities tend to also have higher population densities, assessing data from this pool can help us brainstorm ideas for improving our healthcare system that can eventually be implemented throughout the country. As one of the wealthiest countries in the world, the United States ought to be able to provide for its citizens at the most basic level, which is that of physical health and safety.

The constraints of our dataset include the issue between the industrial nature of big cities that is unfortunately related to many major health risks, but is also what makes these big cities profitable for many large companies. The U.S. Department of Health and Human Services will have to come to some form of agreement between protecting the health and safety of the citizens while working with the big companies that keep the big cities booming. In addition, due to the health crisis within the past two years caused by COVID-19, many other health issues have fallen under the radar. However, this shouldn’t diminish the importance of our findings. As we begin to mitigate the spread of the pandemic, we must now turn to addressing other diseases and health problems that afflict those in the United States. Furthermore, certain government officials may be motivated to maintain their position in office. Thus, if a policy that directs tax dollars towards healthcare is unpopular with a government official’s constituents, they are much less likely to support it, even if it would benefit the health of the citizens.

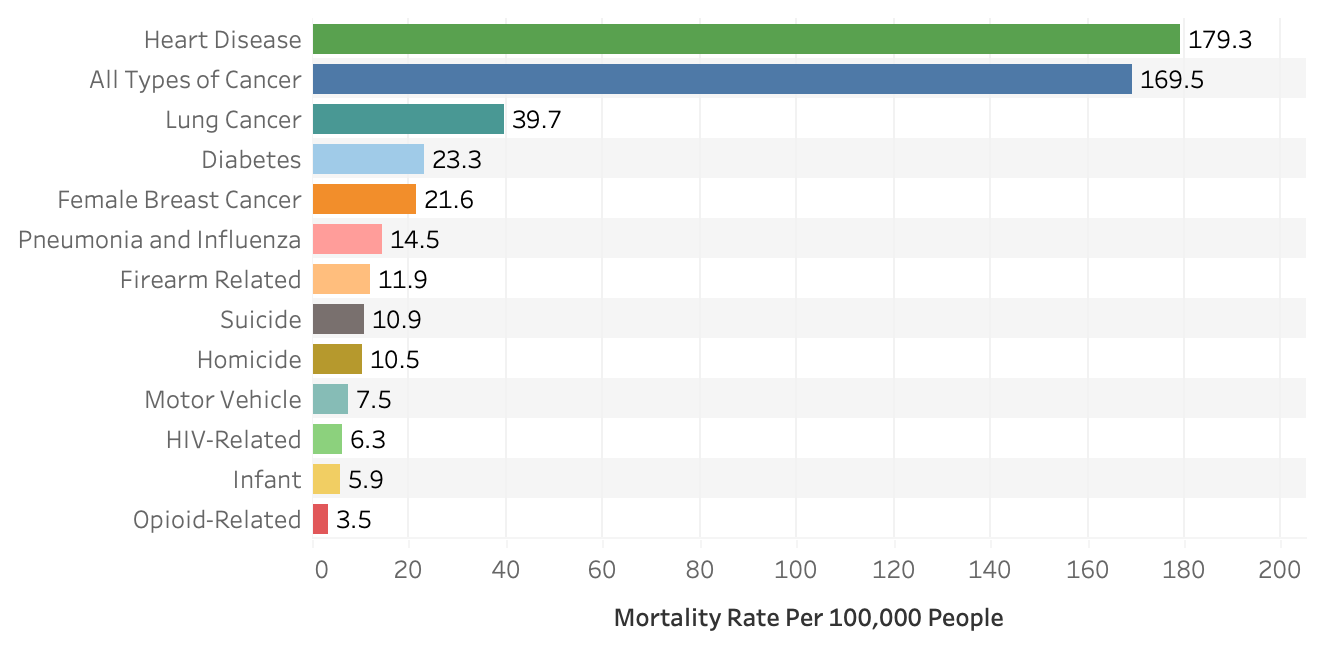
To tackle this complicated issue, we will begin by identifying the most pressing health ailments by determining the diseases that have the highest mortality rates across all the cities. Then, we will investigate any correlations between these diseases and underlying issues, such as obesity, poverty, and education. After that, we will look into other demographic information that might be relevant to mortality rates, such as city, race, and gender. We will use linear and logistic regression to find these correlations, and then test our models using sum of squares, variances, and means. Collecting and analyzing all of our findings, we will then devise a recommendation for the U.S. Department of Health.

## **Research Questions**

1. Which indicator caused the most deaths across all cities in 2012 (heart disease, cancer, homicide, etc)?
2. What are the underlying causes of these illnesses and based on these factors, which city will have the highest overall mortality rate within the next few years?
3. What preventive programs should we invest in to minimize loss of life (tobacco, nutrition, etc)?

**Question 1: Indicator that caused the most deaths across all cities.**

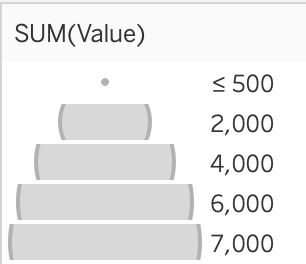
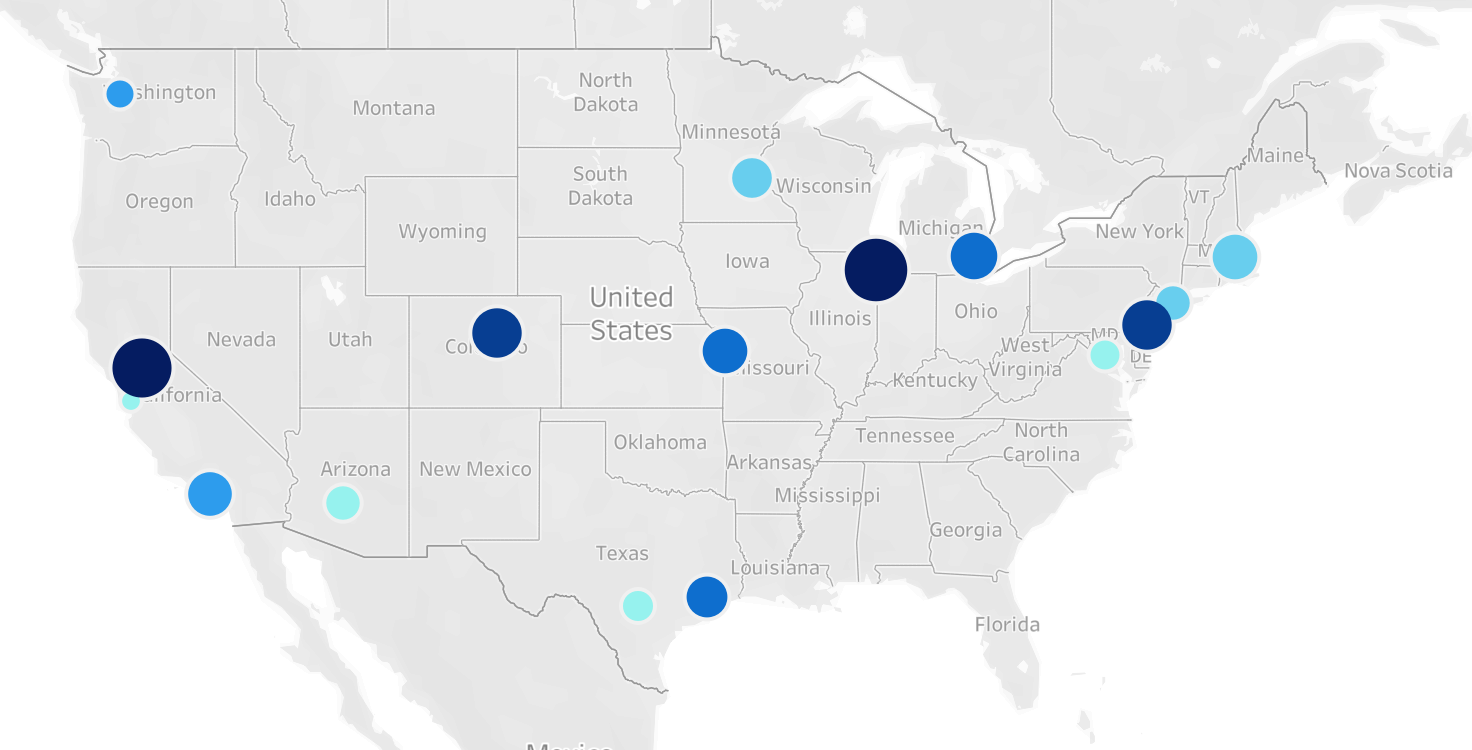
The dataset (Nursnaaz, 2019) includes mortality/infection rates and other statistics for a variety of indicators ranging from cancer to homicide and HIV. The overarching goal of this report is to pinpoint the factors and diseases most correlated with mortality and subsequently devise plans to address them. Thus, the first step in our analysis is to identify the most pressing health ailments to focus on in the United States so we can eventually dive into demographic information and narrow down the focus of our project.

**Figure 1: Causes of death across big cities in 2012. The chart has a bar for each cause of death on the y-axis, and the mortality rate per 100,000 people across all the cities for this cause on the x-axis.**

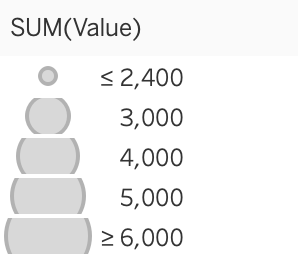
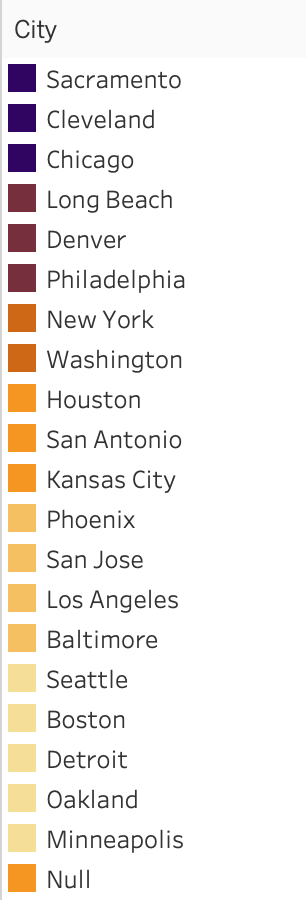
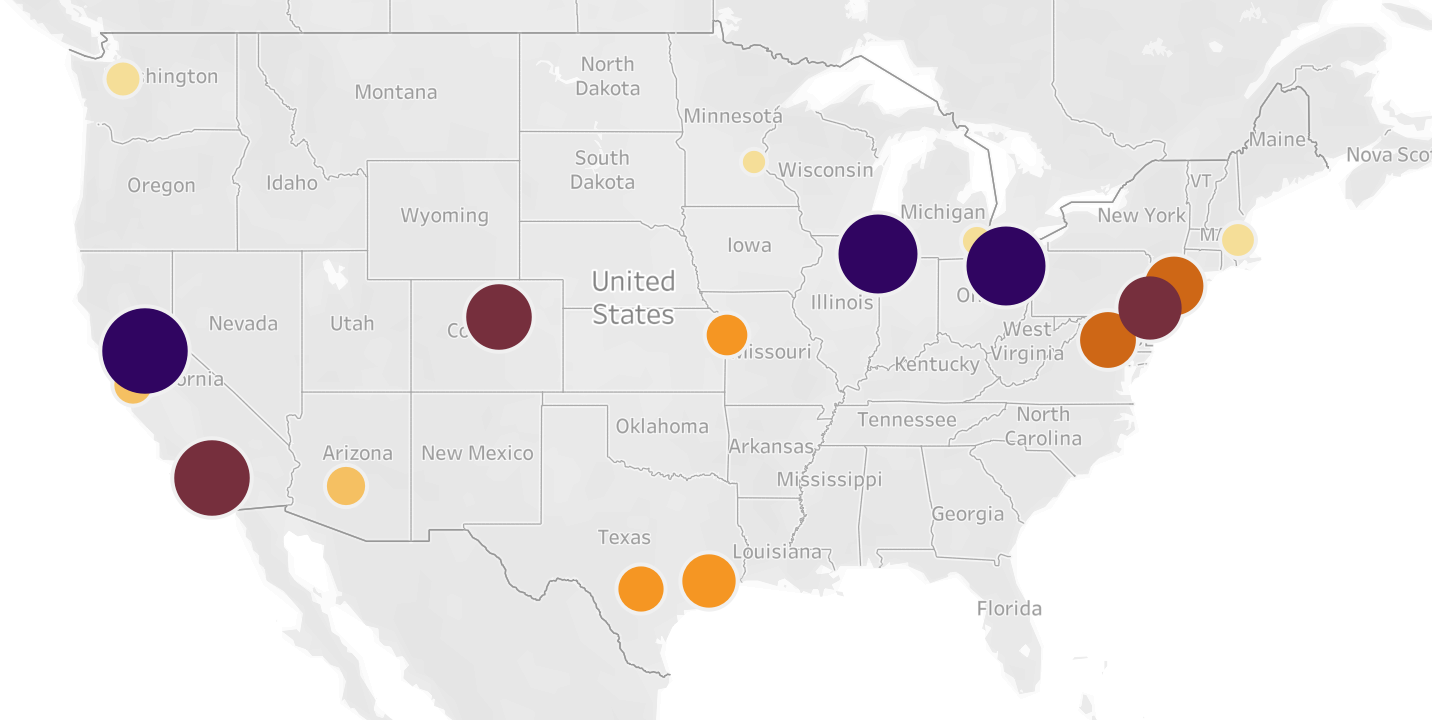
This visualization gives insight into the descriptive question of which causes of death occupy a larger proportion throughout all the big cities. The y-axis represents the various causes of death while the x-axis represents the total count of mortality rate per 100,000 people for that indicator. From this data, we can see that the main causes of death to mortality include heart disease, all types of cancer, diabetes. It relates to the U.S. Department of Health and Human Services, because it will allow them to determine which causes of death need the most support and resources by seeing the count in hundreds of thousands of the population size. By knowing which causes of death are at most risk, they can concentrate their efforts on helping people in these causes of death who are in most need.

## **Question 2: Underlying Causes of Mortality and Risk Factors.**

Hand-in-hand with identifying the culprit of highest mortality rates across all cities, we also want to explore the underlying causes of these diseases, as well as identify the population that’s most at risk based on city and demographic factors. Using linear and logistic regression methods, we hope to identify the correlations between mortality rates and other indicators besides those named as the top causes of deaths.



**Figure 2: Lung Cancer vs. Cities. The map shows a different colored circle for each city. The size of each circle represents the mortality rate of lung cancer. The shades are darker for those with higher mortality rates.**

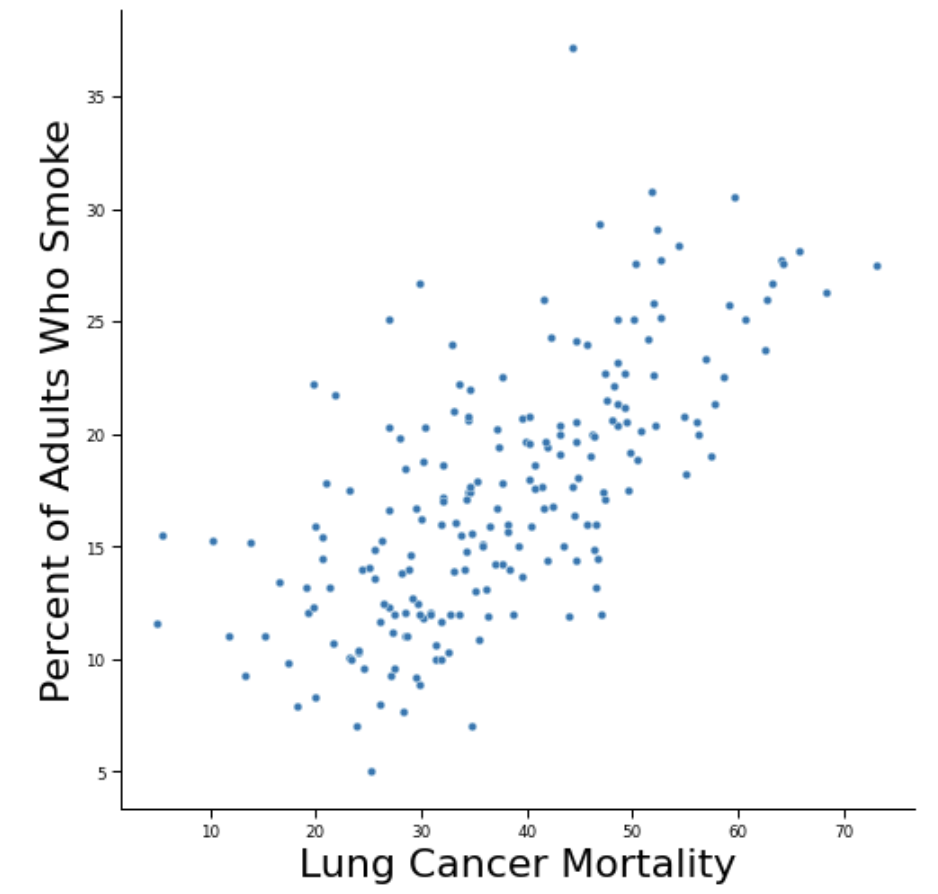


**Figure 3: Heart Disease vs. Cities. The map shows a different colored circle for each city. The size of each circle represents the mortality rate of heart disease. The shades are darker for those with higher mortality rates.**

In the descriptive question, we determined that the top three causes of death in the United States are heart disease, lung cancer, and diabetes. We looked into factors that might cause these, such as smoking for lung cancer, poverty for heart disease, and obesity for diabetes.

**Linear Regression**

We compared the outputs of a multitude of linear regression models in order to determine the indicators with the greatest predictive power on estimating lung cancer mortality rates. First, the dataset was featured to only include cities with at least three data points (for 2011, 2012, and 2013) for the dependent variable of interest. As a baseline, we used a model that only included location and year as features, which had an adjusted R-squared of only .358[A.0]. After also including gender and race as categorical variables, this increased significantly to .636[A.1]. In order to test the relation between rates of smoking and lung cancer in these cities, we performed a self-join on the dataset. The addition of “percent of adults who smoke” as a variable brought the adjusted R-squared value to .812, and the indicator itself had a p-value of 0[A.2].



**Figure 4: Lung Cancer Mortality vs. Percent of Adults Who Smoke. This pairplot shows that there is a positive correlation between these two indicators.**

Looking at the summary of this final model, gender and race also have a significant impact on lung cancer mortality rates. Nationwide, there are disparities in the incidence of lung cancer, with men being diagnosed with lung cancer more frequently than women (Lung cancer fact sheet, 2020), which is reflected in the negative and positive coefficients of the variables, respectively. Furthermore, the corresponding coefficients are only positive for C(Race)[T.Black] and C(Race)[T.White], with the magnitude of the coefficient for C(Race)[T.Black] being almost double that for C(Race)[T.White]. This is corroborated by real-world findings (Schabath et al., 2016), as lung cancer rates are highest among Black Americans, followed by White Americans. Lastly, the cities that were of statistical significance included Las Vegas, Phoenix, and Chicago. Specifically, Las Vegas and Chicago both had positive coefficients, suggesting that these two locations are associated with a higher lung cancer mortality rate.

**Logistic Regression**

Using logistic regression, we also determined how likely these diseases are to be deadly given city, gender, and race. In order to simulate a binary outcome, we created an additional column to our dataset, called “deadly.” The value of this column was 1 if the mortality rate of the disease for that city and demographic was greater than the average plus the standard deviation of the mortality rate. Otherwise, it was set to 0. The logistic regression for lung cancer had a pseudo R-squared value of 0.67 and a log-likelihood of approximately -27,000[B.0]; the regression for diabetes had a pseudo R-squared of 0.62 and a log-likelihood of approximately -33,000[B.1]; and the regression for heart disease had R-squared of 0.797 and log-likelihood of -30,000[B.2].

Based on these results, our model for heart disease was the most accurate. The p-values for almost all of the variables were significant (< 0.05), except for some of the cities as well as the multi-racial category, which we assume is a result of too few data points. According to the coefficients of these variables, heart disease is likely to have a higher mortality rate for those who are male and black, as well as those who live in Chicago, Long Beach, Philadelphia, and Sacramento. These are the same cities that show the highest rates of mortality of heart disease in Figure 3. Those less at risk include females, hispanics, and those living in Boston, Seattle, and Minneapolis.

**Testing Assumptions**

To test the validity of the assumptions made when constructing the linear regression model, we created a normal QQ-plot of the residuals. The figure depicts a straight line at approximately a 45 degree angle, supporting the statement that the residuals are normally distributed. The autocorrelation plot tests the mutual independence of errors. With no lags present, there is sufficient evidence supporting the assumption that the residuals are independent. The residual plots are difficult to interpret because many categorical variables were used and there is more noise due to a smaller sample. While the LOWESS line isn’t completely straight, there isn’t enough variation to completely reject the assumptions about the constant mean and variance of the residuals[A.3].

To test for overfitting in our logistic regression models and evaluate causality, we recalculated each model using a test, train, split procedure. We then compared the sum of squares for the predictive model on the test data with that of a simple prediction of 0. Each model had a lower sum of squares than the simple prediction, showing that there was no overfitting[B.3].

**Constraints**

Our models for both the logistic and linear regression could not determine any significant features for diabetes. According to our research, diabetes is caused by genetic as well as environmental factors, which are not indicators included in our dataset. Although we cannot use this in our evaluation of risk, diabetes was the lowest in terms of mortality rate for all cities, and therefore not as pressing to address. Furthermore, certain cities could not be included in our analysis due to a limited amount of data (for example, some cities only had values for one year, or did not have values at all for the indicators we were looking at). In order to further explore diabetes as an indicator, more data on confounding factors and diabetes mortality needs to be added. Due to a limited amount of data, it’s likely the models would’ve been subject to overfitting.

**Discussion**

The cities we determined to have the highest mortality rates in the future are Chicago, Sacramento, and Las Vegas. Looking at Figure 2 and Figure 3, it is evident that Chicago and Sacramento have some of the highest rates of lung cancer and heart disease. This was confirmed by our logistic regression. Las Vegas was not included in these figures, but our linear regression model determined that it had the highest positive correlation with mortality from lung cancer.

## **Question 3: Preventative Programs to Minimize Death.**

Given the scope of this dataset, it can be difficult to pinpoint how to best distribute resources targeting each of the indicator categories. Spanning from tobacco to underage drinking and nutrition, the expansive nature of the data can be overwhelming. Thus, we seek to address the question of which preventive programs the government should invest in to minimize loss of life and improve health in the long-term. Treating acute illness once it’s already developed not only means the U.S. lags behind other high-income measures such as life expectancy but also leads to a sharp increase in medical spending in proportion to GDP that is unsustainable. With our insights, we hope to advise the U.S. Department of Health and Human Services on how to mitigate the aforementioned issues plaguing the U.S. healthcare system.

Based on our data analysis, we discovered that the main underlying causes included obesity, smoking and poverty. We decided to focus our preventative program efforts towards three main areas including Vaccines & Immunizations, Nutrition & Fitness, and Mental Health & Substance Abuse. We strongly recommend supporting people to get annual check-ups where tests and screenings can be done to check on certain health conditions such as blood pressure, cholesterol levels and whether or not someone has diabetes. Vaccines and immunizations can also be conducted on an annual basis to protect people against potential health ailments. Cancer screenings are recommended to find cancer at an early stage before symptoms appear and the effects are irreversible.

We found that substance abuse is a leading cause of lung cancer. Due to the large positive correlation between smoking and lung cancer, we should advocate for campaigns against smoking, specifically targeted toward Black Americans in Denver, Washington DC, Las Vegas, and Baltimore. We advise increasing funding for campaigns against smoking to target young adults. The average adult smoker starts smoking at about the age of 18. Preventative programs such as anti-smoking campaigns like the Real Cost discourages young adults from starting to smoke in the first place by targeting these audiences through social media and creating a support system for them. While not necessarily fatal, obesity is also a precursor to many other health issues and has been linked to drinking and smoking in low-income populations. Thus, our final suggestion is to increase funding for national food campaigns because they bring awareness and provide resources for healthier nutrition options. Nutrition is important for combating many health ailments, but is especially useful against obesity. Programs such as these can help support people in poverty to have the proper nutrition they need. Some examples include the Child and Adult Care Food program, Community Food Systems, and Fresh Fruit and Vegetable Program which are all funded by the USDA.

## **Conclusion**

After analyzing all data pertaining to fatal diseases, it seems the three most pressing ailments to address are heart disease, lung cancer, and diabetes. These three indicators have the highest rates of incidence in the country overall and subsequently will affect a large subset of those living in the U.S. in the future. Building off of this information, we conducted linear and logistic regression to pinpoint the contributing factors to these diseases. Specifically, the population at the biggest risk for deteriorating health at the hands of these conditions are those living in Chicago, Sacramento, and Las Vegas. After assessing demographic information within this model, we identified male Black Americans within these areas as being the most susceptible. The validity of these conclusions was upheld through causality and testing the assumptions of our models.

We hope to guide The U.S. Department of Health and Human Services toward policies that can target the issues we identified. With a budget of almost over a trillion dollars (Assistant Secretary for Financial Resources, 2021), the U.S. Department of Health and Human Services has the potential to drastically improve the wellbeing of citizens of the United States. Allocating resources in accordance with our findings would allow the government to maximize impact, and this can be accomplished in two ways.

Firstly, the U.S. Department of Health and Human Services provides support to biomedical research and the development of health information technologies. By identifying the diseases and health risks that will proportionally afflict the highest number of individuals in the United States, we can determine where funds for research should be allocated. Then, by doing a more granular analysis of cities at risk for each disease, we can pinpoint locations where trial medications should be launched, such as Chicago, Las Vegas, and Sacramento for new lung cancer treatments. Similarly, we should pay special attention to patients with heart disease in Chicago, Sacramento, and Long Beach in the context of new heart disease treatments due to high mortality rates in those areas.

Despite the U.S. being a hub for innovation, especially in the realm of medical technology, 10% of the United States is uninsured (Cha & Cohen, 2022) and we lag behind other high-income countries in life expectancy. Furthermore, pervasive inequities embedded in the healthcare system lead to disparities in care between low-income populations, especially non-white populations, and the wealthy. In order to prevent medical spending from increasing without a paralleled increase in population health, our second recommendation is to redirect attention to the root of the problem through the implementation of preventive programs. Specifically, due to the large positive correlation between smoking and lung cancer mortality, we advocate for increased funding for campaigns against smoking. These initiatives should be concentrated in the cities and populations within them determined to be most at risk: male Black Americans living in Las Vegas, Sacramento, and Chicago.

Ultimately, while we were able to extract many insights from this dataset, we’ve only scratched the surface of how to tackle the issue of healthcare in the United States. Due to the limited time period represented in this data and small sample size, there’s potential for further exploration with more data and forecasting techniques. As mortality rates continue to rise and we suffer from health crises and pandemics, it’s more urgent than ever that we start to prioritize the health of U.S. citizens.

## **References**

Assistant Secretary for Financial Resources. (2021, May 28). *HHS FY 2022 Budget in Brief*. HHS.gov. Retrieved May 10, 2022, from https://www.hhs.gov/about/budget/fy2022/index.html

Cha, A., & Cohen , R. (2022, February 11). *Demographic Variation in Health Insurance Coverage: United States, 2020*. Retrieved May 16, 2022, from https://www.cdc.gov/nchs/data/nhsr/nhsr169.pdf

Lung cancer fact sheet. American Lung Association. (2020). Retrieved May 11, 2022, from

https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/resource-library/lung

-cancer-fact-sheet

Nursnaaz. (2019, June 16). Big City Health Data. Kaggle. Retrieved May 4, 2022, from

https://www.kaggle.com/datasets/noordeen/big-city-health-data

Schabath, M. B., Cress, D., Munoz-Antonia, T. (2016, October). Racial and ethnic differences in

the epidemiology and genomics of Lung Cancer. Cancer control : journal of the Moffitt Cancer

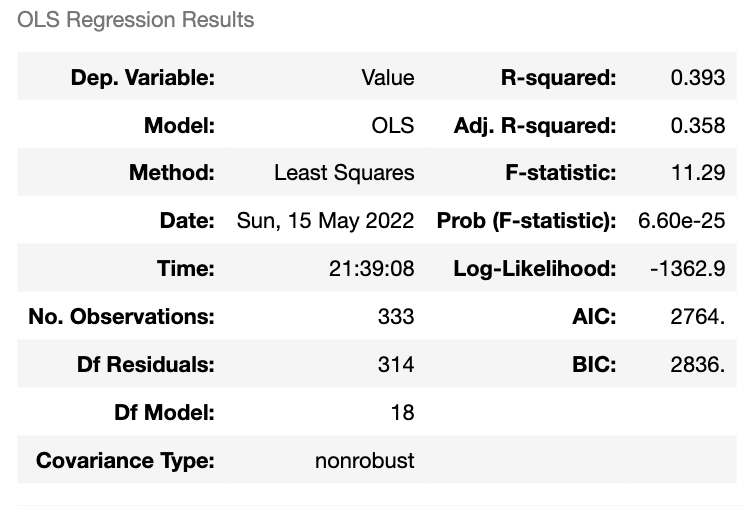
Center. Retrieved May 12, 2022, from

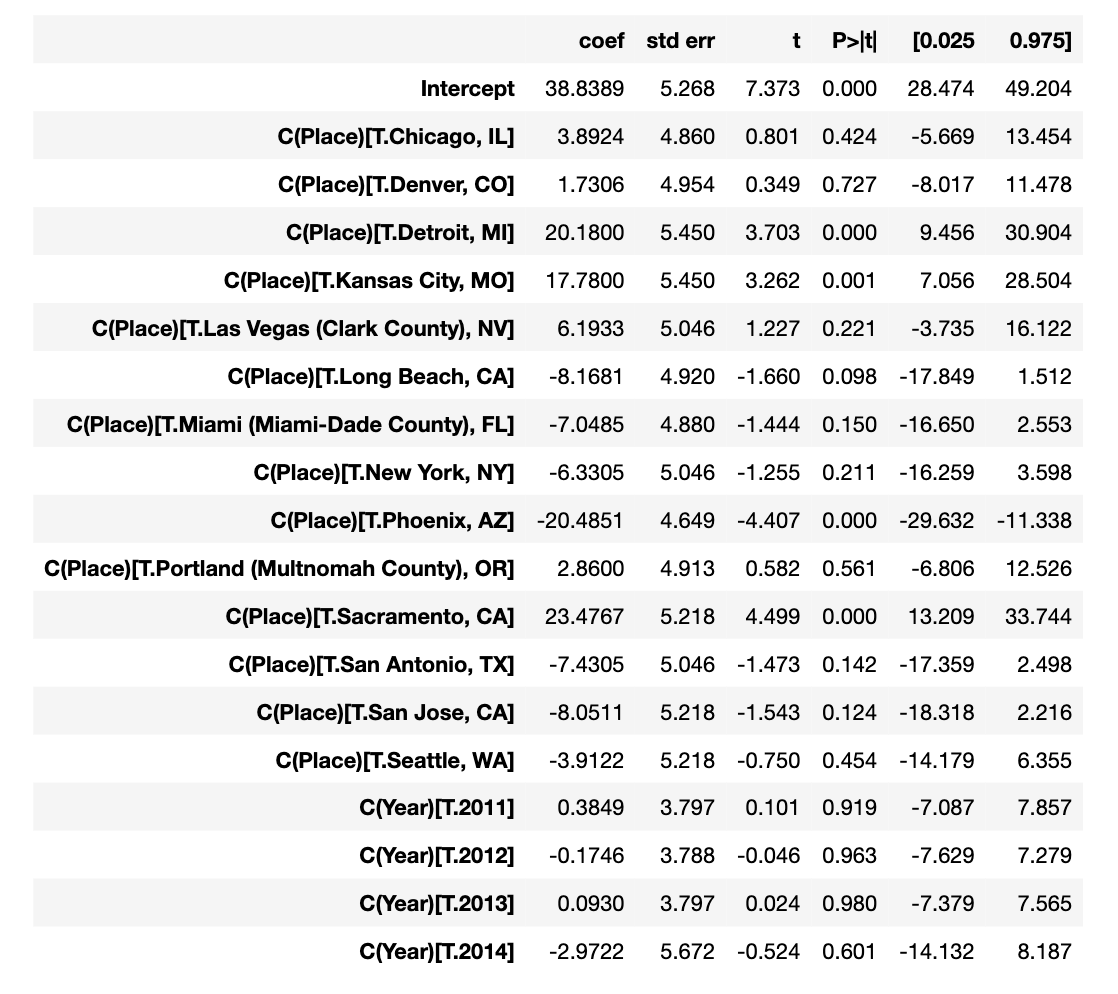
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5340153/#:~:text=In%20this%20report%2C%20

the%20annual,Islanders%20(38.4%20per%20100%2C000)

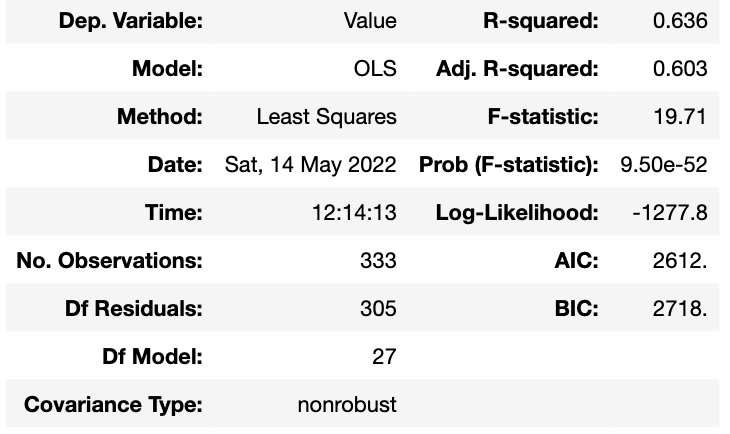
**Appendix A**

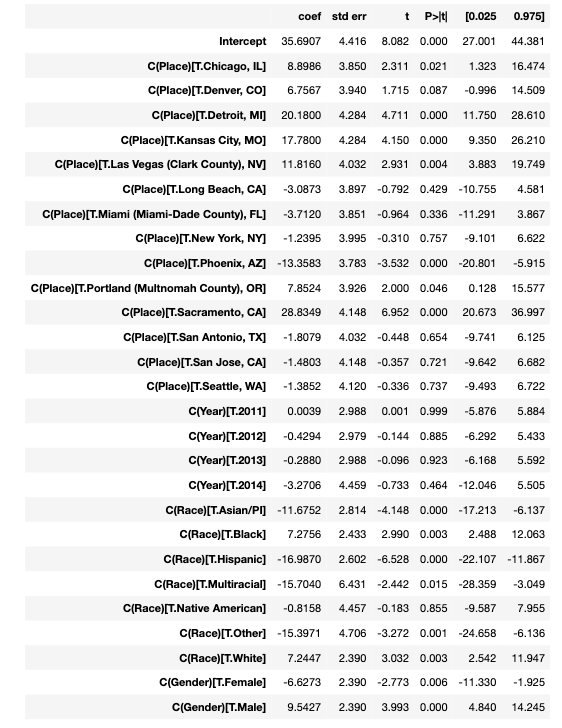
[A.0] Lung Cancer linear regression model 1– using just city and year as categorical variables:

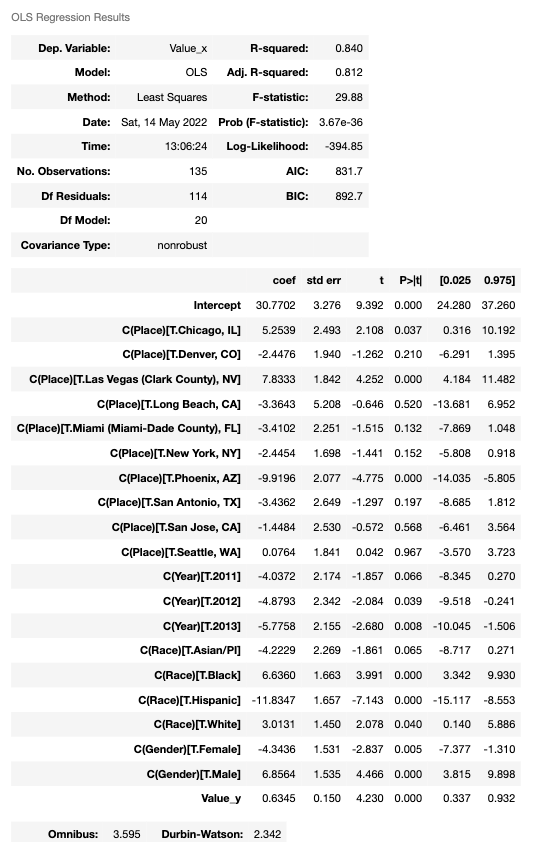


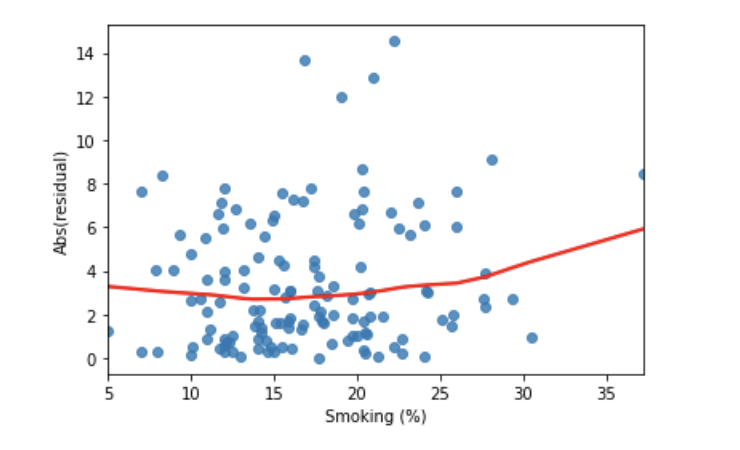
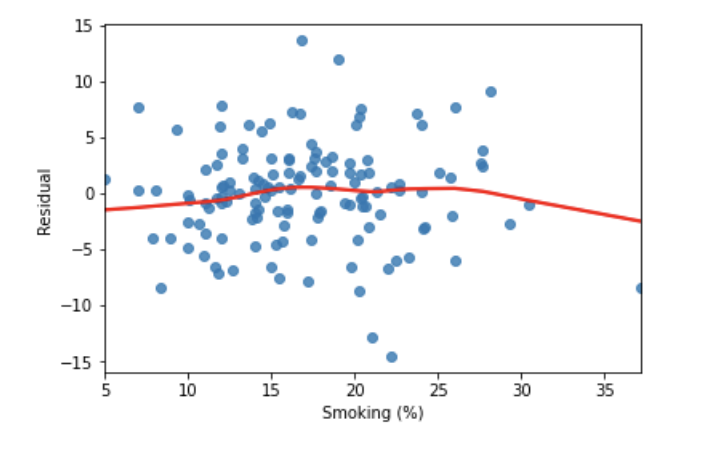


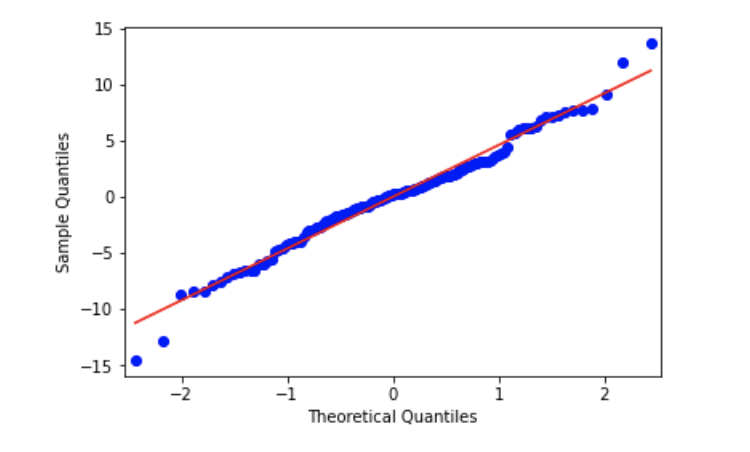
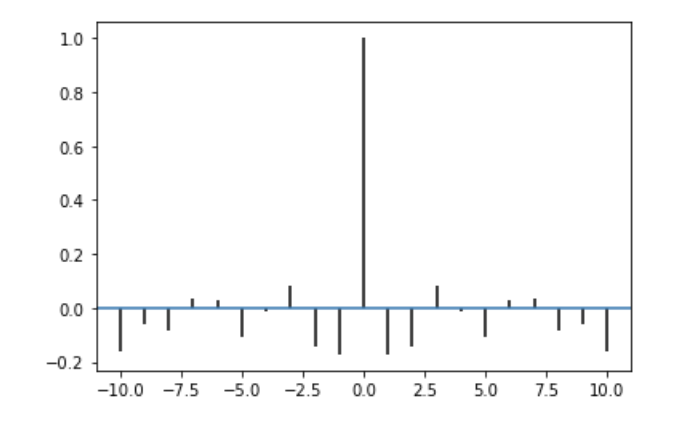
[A.1] Lung Cancer model 2– using city, year, gender, and race as features :





[A.2] Lung Cancer linear regression model with smoking (Value\_y): 

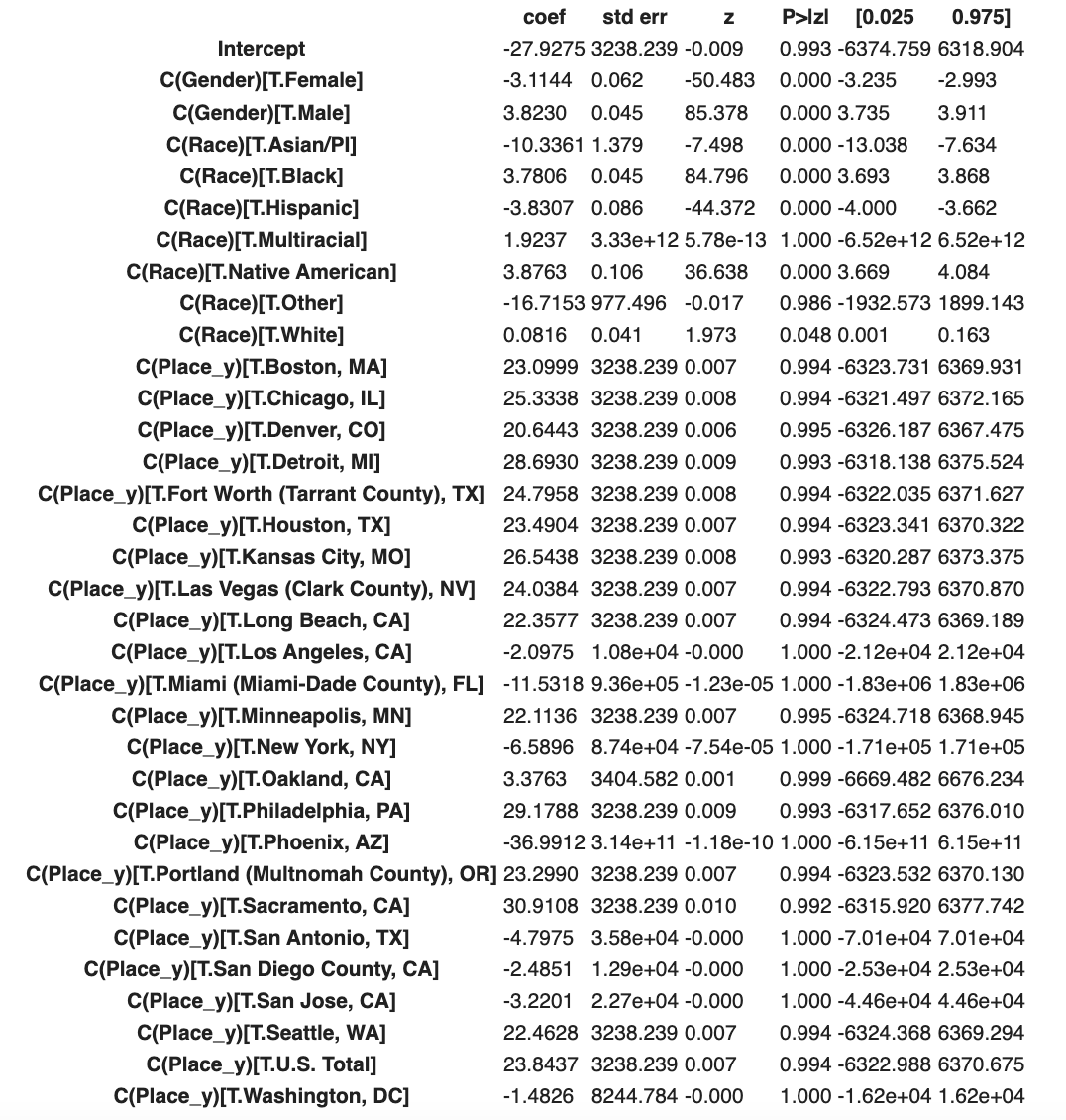
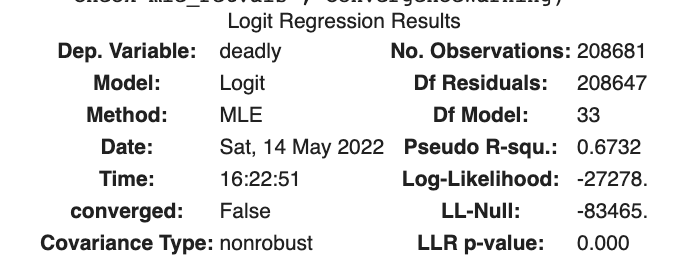
[A.3] Testing assumptions of the final linear model chosen (model 3)



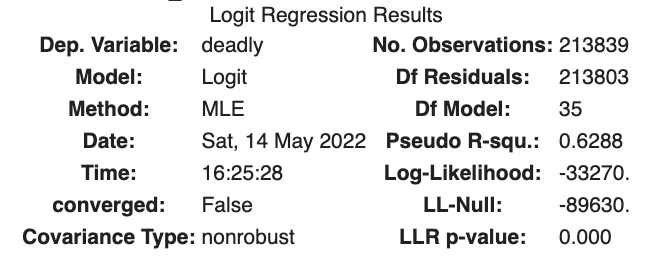
**Appendix B**

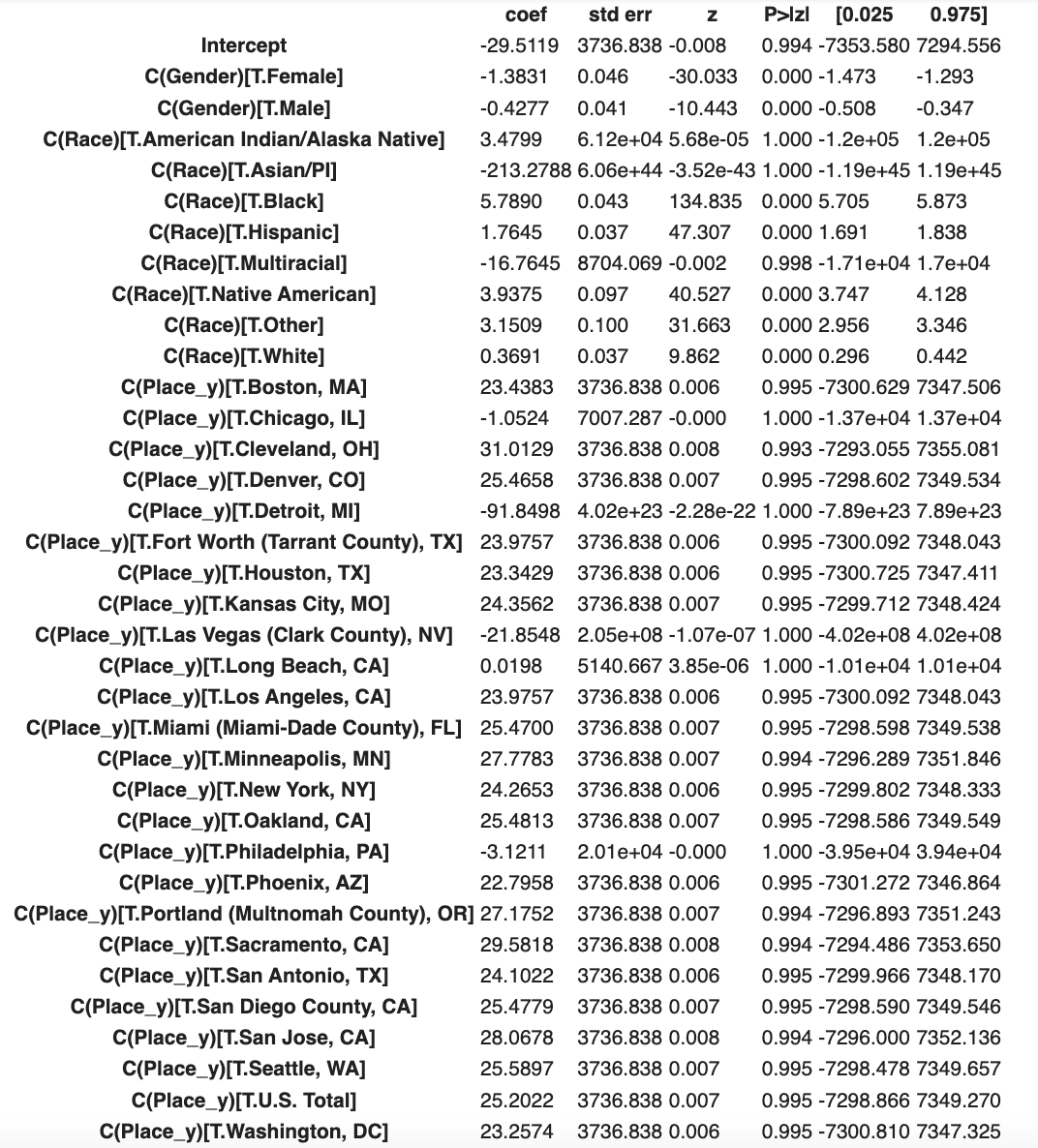
**Results of Logistic Regression Models**

[B.0] Lung cancer:

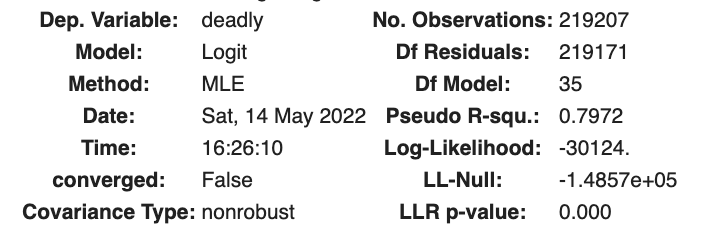


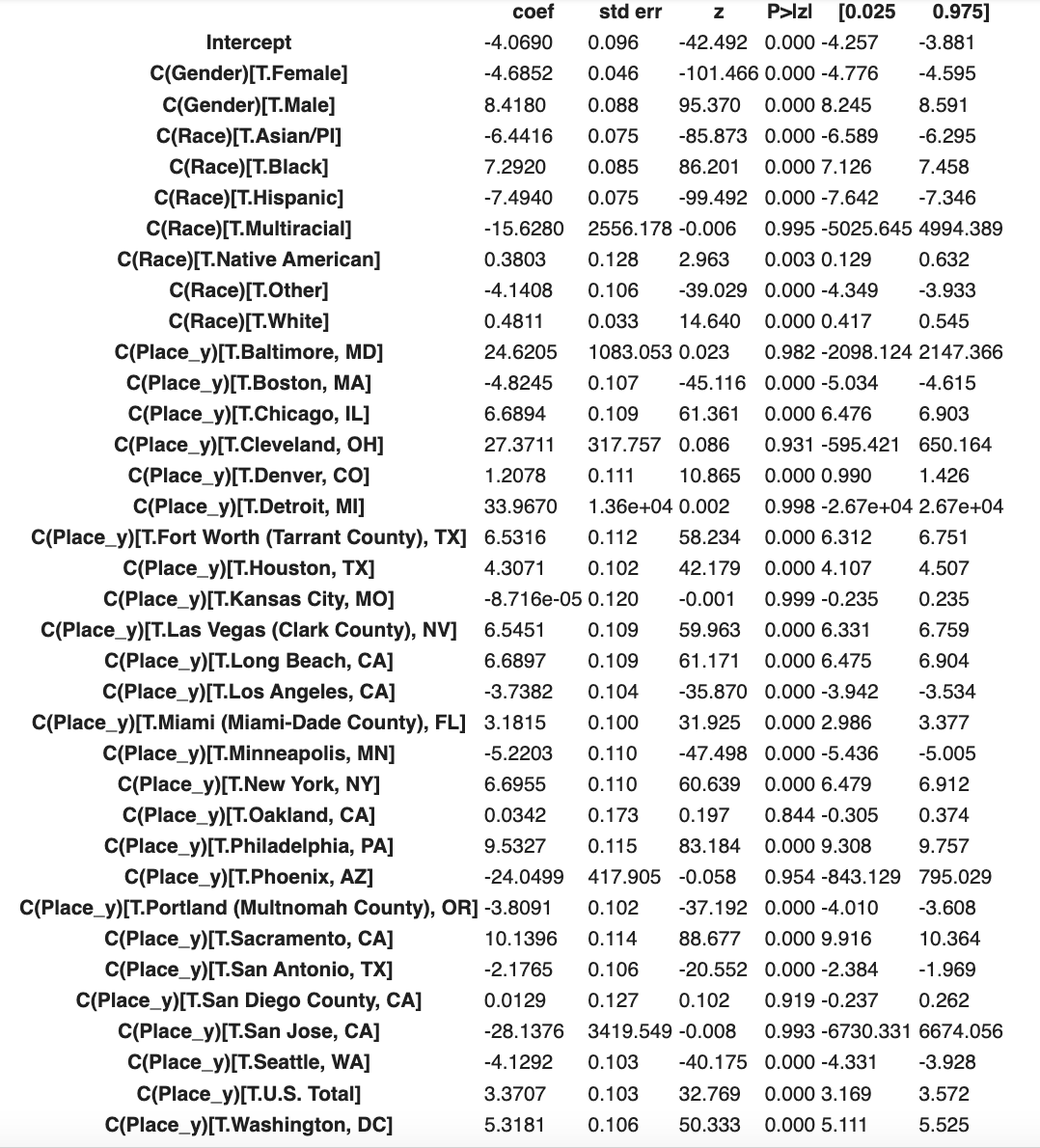
[B.1] Diabetes:





[B.2] Heart disease:

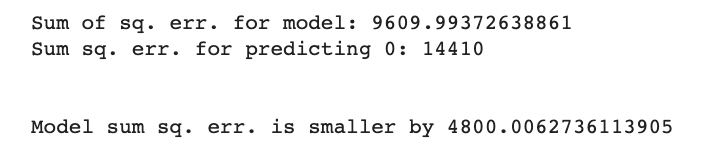




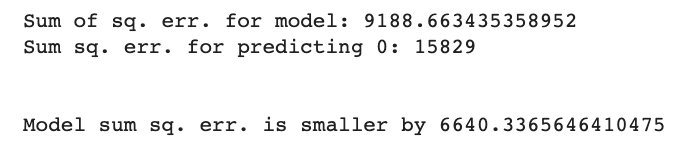
**Sum of Squares for Logistic Regression Models**

[B.3]

Lung Cancer:



Diabetes:



Heart Disease:

