**Introduction**

A growing concern with the increasing degradation of the ozone layer, is the effect it will have on skin cancer increase. Ozone molecules are composed of three oxygen atoms, and they reside in the ozone layer of the stratosphere[1]. As the ozone layer depletes, more UV rays pass through the stratosphere. Skin cancer is one of the more common types of cancer in the United States, commonly due to overexposure to UV light, a weak immune system, and family history [2]. Melanoma is not the most common type of cancer, but it is the type most likely to spread [2]. Knowing this, it is critical to notice the possible relationship of the diagnoses of skin cancer with the depletion of the ozone layer. The land in the United States is not uniform. The northeast region is more mountainous, while the mideast consists more of flat terrain. This can pertain to the relationship of UV exposure due to the proximity that the land is to the stratosphere since the closer to the stratosphere the land is, the higher the probability of getting more UV ray exposure. The problem that will be explored is the relationship between melanoma and ozone degradation with specificity to region.

**Data Source**

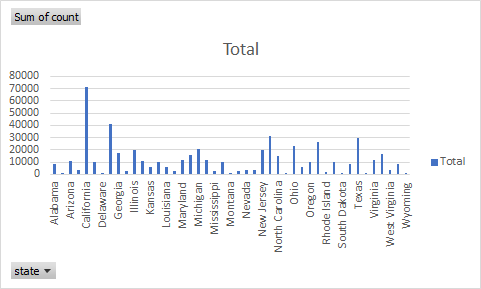
The dataset was supplied by the Centers for Disease Control and Prevention(CDC). The sample of data was chosen from people who had skin cancer, and lived in  U.S. between 1999 and 2009. The data was provided by the National Center for Health Statistics (NCHS), which we assume to be a reliable source. Due to the *Data Documentation* link being private for CDC members only, specific details on how the sampling was conducted was limited [3].

**Data Description**

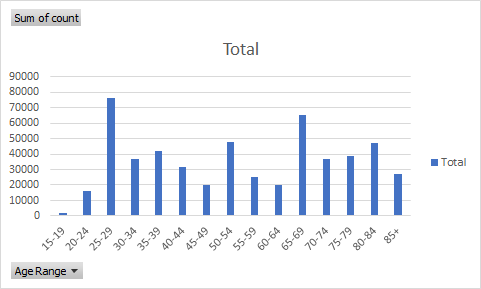
Five variables were present in this study:

* Age: 15-85+, split into 15 distinct age groups of five year intervals (15-20, 21-25, etc.)
* State: 44 different states
* Count: Number of people who died from skin cancer in each age or state
* Population: The total population of each state
* Rate: The rate of skin cancer per 100,000 people
  + (Rate = Count/Population \* 100,000)

The categorical predictors are age and state because they allowed to measure rate as the response variable against each age group and state. The count variable varied based on whether the count measured the number of people in age group or in each state. With defined variables, the model was generated. Below are the histograms of the variables. Figure 1 is State vs. Count, and Figure 2 is the Age Range vs. Count.



***Figure 1.*** *Histogram of State vs. Count*



***Figure 2.*** *Histogram of Age Range vs. Count*

**Modeling Methods**

**Diagnostics:**

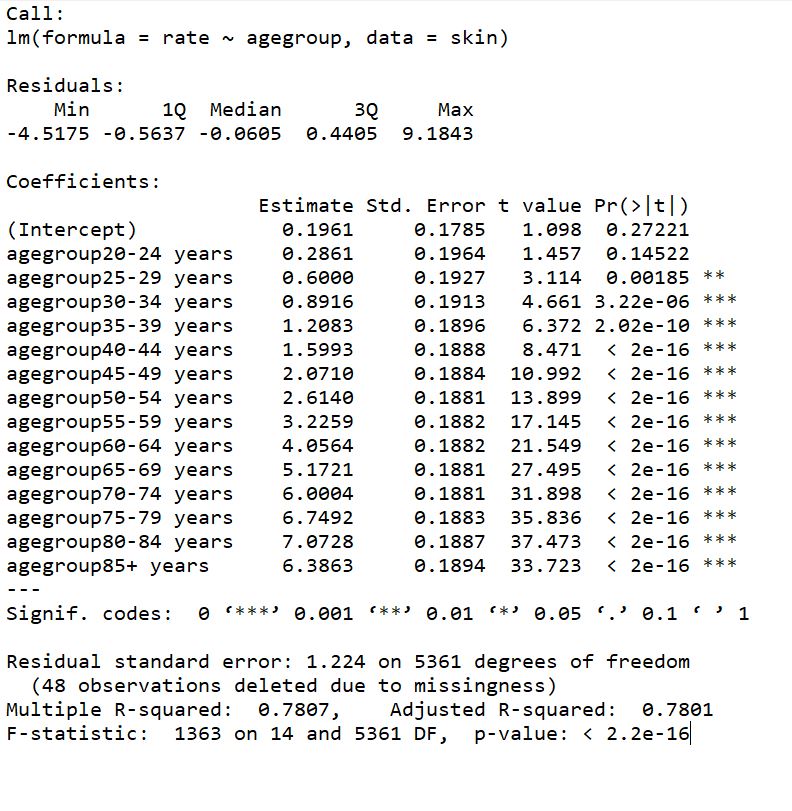
To test the Diagnostics of the model, a qq plot was ran using regression functions to see if the data followed a normal distribution or not. The qq plots and box plots were used for the categorical variables to see whether or not they had a “slightly normal” distribution

**Inferential test:**

After the model was built in R, a regression analysis was ran on the data. When a regression is run in R, the program automatically outputs the significance levels of the data using a T-test and an F test.

* + T-Test - do any of the independent variables have a linear correlation with the dependent variables?
  + F-Test - do any of the coefficients in the regression differ from 0? (i.e does it have a meaningful effect on the dependent variable)

**Modeling Results**

****

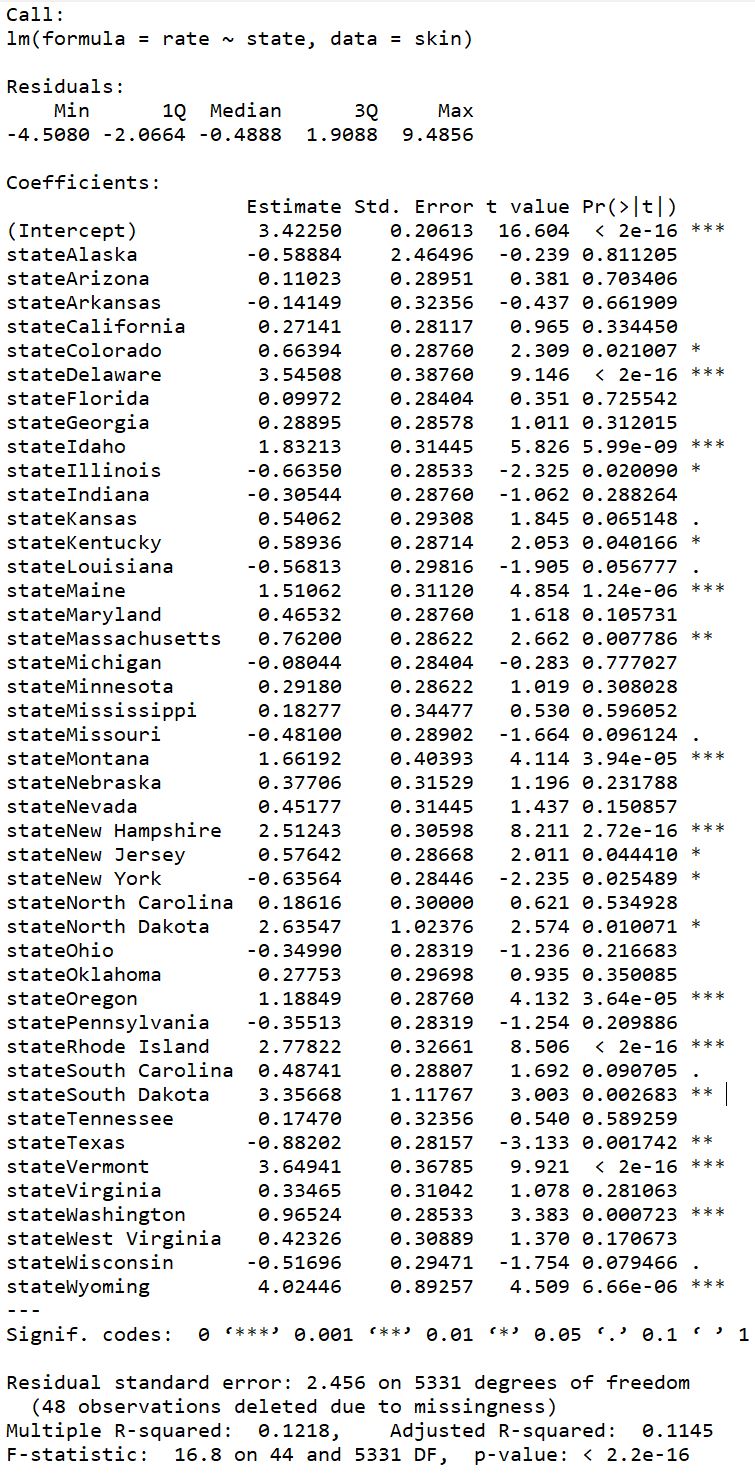
***Figure 3.*** *R output for the linear model of melanoma rate vs age group.*

Figures 3 and 4 demonstrate the F-test and T-test outputs for melanoma rate vs age group and melanoma vs state, respectively.

The T-test outputs for both give differing interpretations for each of the linear models. In age group, the T-value is generally significant, with a steady increase based on age group. This indicates that the independent variables demonstrate some type of linear correlation with the categorical predictor of age group against the response rate of melanoma. For states, the T value can be significant for some of the independent variable groups, such as Vermont (t value = 9.921). In general, values do not show the same trend for significance as the age groups did, which demonstrates that a linear correlation does not necessarily exist between melanoma rate and specific location in the United States.

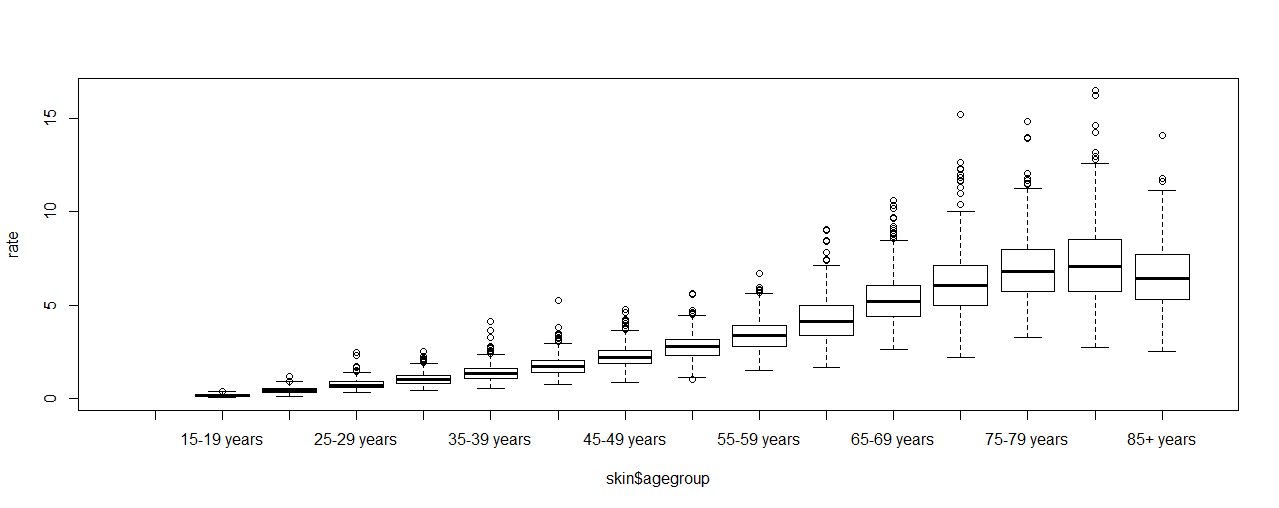
The F-test statistic outputs provide context for whether any of the coefficients differ from zero and consequently whether any coefficients have a meaningful effect on the response rate. For both age and states, the F-test p-values are less than any reasonable significance level due to how (2.2e^-16) which gives us enough evidence to reject the null hypothesis and accept that the results are not significant and the coefficients do not have a meaningful effect on the dependent variable for both state and age.

We also notice in Figure 3, the R squared value came out to be about .78, which indicates that about 78% of the data variability can be explained by the model for age. While in Figure 4 the R squared value is about .12, which indicates that only about 12% of the data variability can be explained by the model for state. These outcomes were consistent with our findings.

****

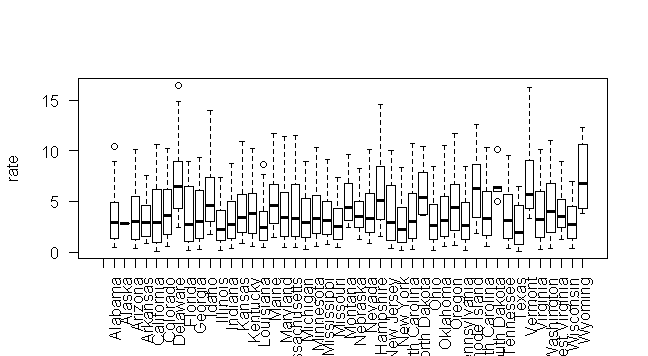
**Figure 4.** *R output for the linear model of melanoma rate vs state.*

Figure 5 describes the data in the form of a box plot after running and graphing the regression model for age. After analyzing this graph, we can deduce there is a trend between age and death rate. As age increases, the average death rate also increases, which is expected in this situation. Generally, it is more likely for an 80 year old to die from cancer than it is for a 15 year old.

****

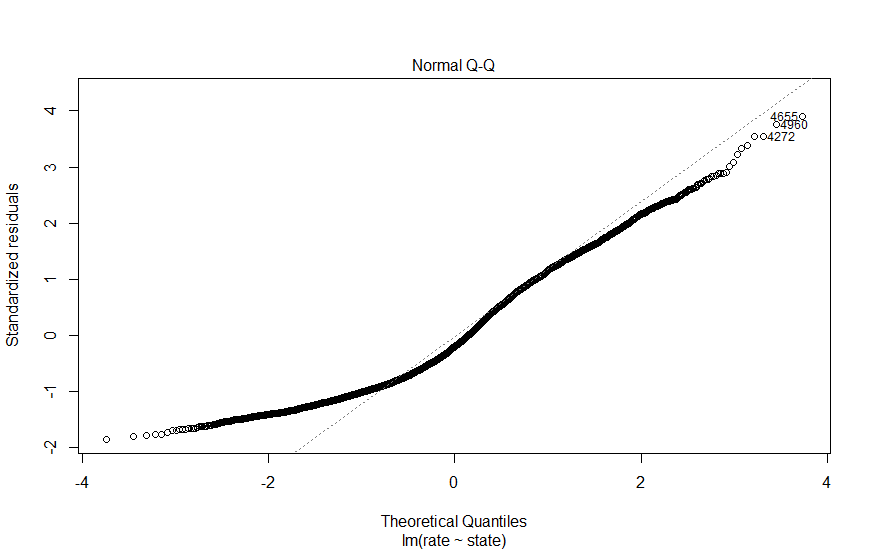
***Figure 5.*** *Box plot for age.*

After running the regression model for state, and graphing the results, a series of box plots describes the data. After analyzing this graph we can see that there is no distinguishable trend. It is very bi-modular with random high and low points. From only looking at the graph, it is difficult to draw a conclusion since there is no distinct pattern.

****

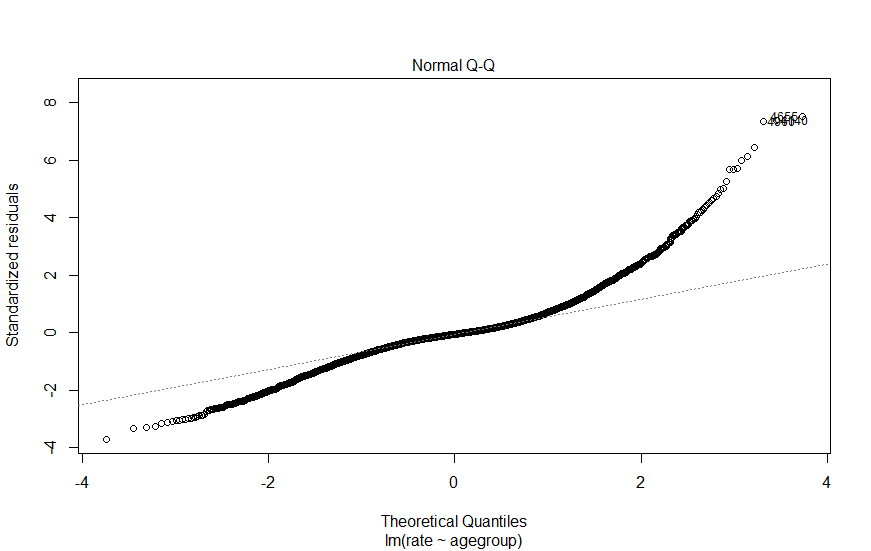
***Figure 6.*** *Box Plot for State.*

In addition to the two box plot graphs, QQ plots were also created based off the model in order to view the quantiles of melanoma rate and state. Looking at Figure 7, the data is not distributed normally. The graph is skewed at both the top and bottom. This is an outcome that was expected due to the bi-modular nature of the box plot for state.

****

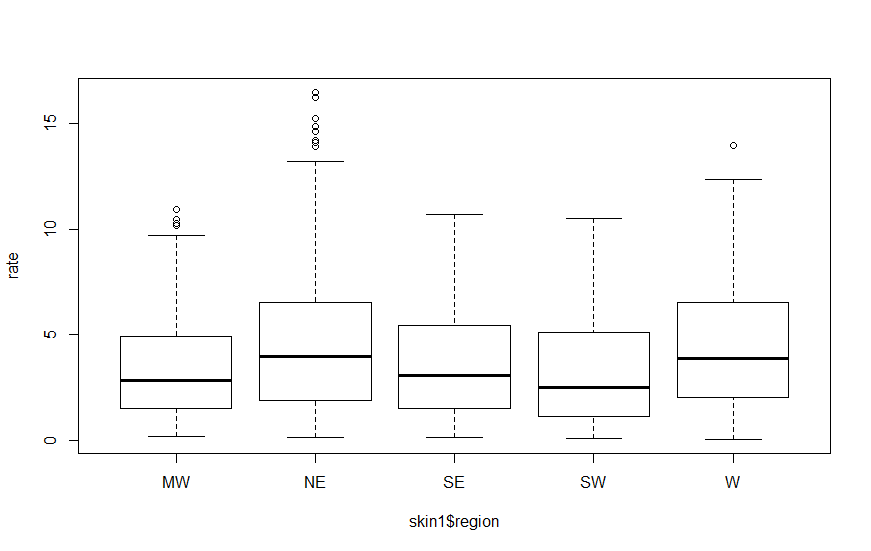
***Figure 7.*** *QQ Plot for State.*

Figure 8 displays the QQ plot for age based off the model mentioned afore. This graph shows that the dates are not distributed normally, as the graph is heavily skewed at the top and bottom. This outcome was also expected based off the information gathered from the box plot for age. Referring to Figure 6, the data is heavily skewed to the left, indicating the lack of a normal distribution.

****

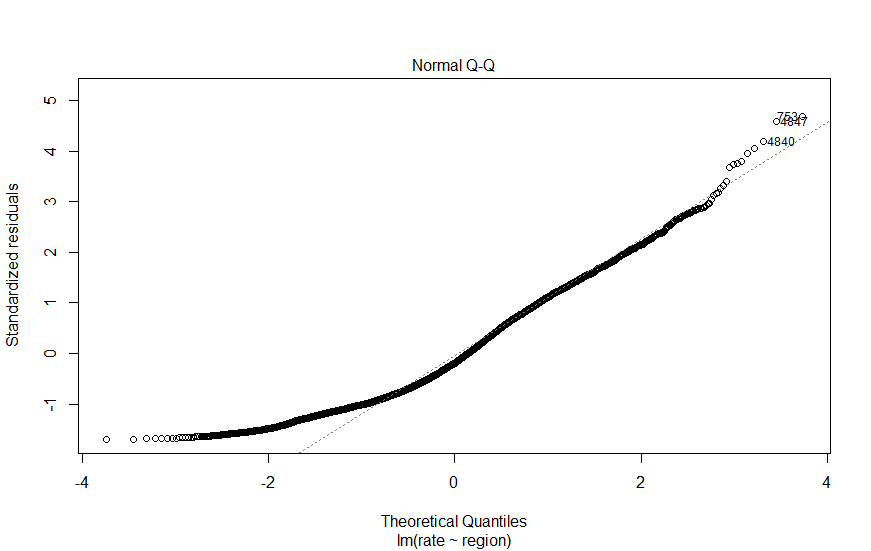
***Figure 8.*** *QQ Plot for Age.*

Figure 9 shows a box plot based off of US regions. Instead of separating the states individually, which was originally shown above, the states were separated into 5 different regions: midwest, northeast, southeast, southwest, and west. Although there is a trend that was gathered from this plot, there is a notable outlier in the northeast region. It was interesting to see that the northeast region had the highest rate because people relate skin cancer to the amount of sun exposure, but the northeast region of the US does not have much sun exposure.

****

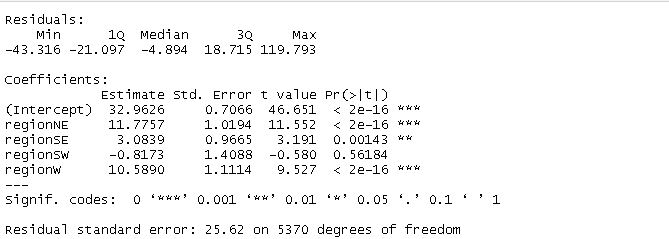
***Figure 9.*** *Box Plot for US Regions.*

Figure 10 is the corresponding QQ plot of the box plot in Figure 9. When comparing the QQ plots in Figures 5 and 6, this plot contains a more normal distribution. It is only heavily skewed at the bottom. This result was anticipated given the nature of the regional box plot.

****

***Figure 10.*** *QQ Plot for US Regions.*

Figure 11 shows the regression table for the US Regions. From this table we can see that when compared to the Midwest, every region has higher levels of skin cancer rates aside from the southwest region. Moreover, we can also see that each region is significant at least the 99% region or higher, aside from the SW region.



***Figure 11.*** *Regression table for US Regions.*

**Discussion**

The data indicates there exists a significant correlation between getting older and developing skin cancer. Beyond the past 30 age group, all results were significant at the 99% level which confirmed the general trend that people tend to suffer from skin cancer as they grow older. Despite correlation in age groups, regression analysis across the states failed to yield similar patterns. For example, the rates of Vermont, Wyoming, and Delaware were all significant at the 99% level while Arizona, Kansas, and Nebraska demonstrated no significance at all. A trend discovered was that states with significant results were mountainous. Nevertheless, states did not reveal any clear underlying trends in our regression models. Our regional analysis yielded mixed results. The Northeast region of the United States was the most significant while the Southwest demonstrated no significance.

A possible factor for these results could be the topography of the land in the United States because the northeast yielded higher melanoma rates and the northeast has more mountainous land. In extension, the southwest, which consists of flatter land, had lesser melanoma rates. As for age, it is intuitive that the older a person grows, the more at risk they are to getting cancer, thus melanoma. Therefore, a positive correlation between age and melanoma was expected.

It is also important to discuss the non-normality of the data. From the QQ plots, we can see that the data was not normal at all, which is what was expected. Cancer diagnosis does not follow a normal distribution because there are numerous risk factors associated with developing any form of cancer, including skin cancer. Therefore, we cannot expect normality in our data, since there are many factors developing cancer which means different groups may experience different rates of cancer, particularly skin cancers.

There are a few limitations to the model. Not every US state was included in the model because some states did not provide information on their skin cancer rates, or the researchers decided not to include the states in their data collection. This is also true for the regional analysis, since it was based off the state data collected. This may skew our data if the states that were not included in the data had high levels of skin cancer rates, like Utah. Another limitation of the model was that we looked at age group and state when determining skin cancer rates, even though cancer is an extremely complex disease. There are many other risk factors that are associated with cancer rates including race, diet, etc. This may explain why the R2 values were low for the models; there is more information needed to accurately predict skin cancer rates.

In the future, this analysis can be used to predict which states and regions need more funding to combat skin cancer. From the analysis, it does not make sense for the government to provide the same levels of support to fight skin cancer in the southwest as it does in the northeast, since the northeast experiences much higher levels of skin cancer that in the southeast. This is also true when developing state-based plans to fight skin cancer. From the analysis we can see that not every state has the same risk factors when it comes to its residents developing skin cancer. Therefore, the government should focus more time and resources on these states that are most at risk.

**References**

[1] O. US EPA, “Basic Ozone Layer Science,” US EPA, 16-Jul-2015. [Online]. Available: https://www.epa.gov/ozone-layer-protection/basic-ozone-layer-science. [Accessed: 21-Mar-2018].

[2] “Melanoma Skin Cancer.” [Online]. Available: https://www.cancer.org/cancer/melanoma-skin-cancer.html. [Accessed: 21-Mar-2018].

[3] C. Tiwari, K. Beyer, and G. Rushton, “The Impact of Data Suppression on Local Mortality Rates: The Case of CDC WONDER,” American Journal of Public Health, vol. 104, no. 8, pp. 1386–1388, Aug. 2014.