**An Analysis of U.S. Lung and Bronchus Cancer Rates since 1975**

Kirby Leo

May 7, 2018

Introduction to Computing

**Abstract**

Lung and bronchus cancer causes the most deaths out of all cancers in not just the United States, but all over the world. To further understand its effects on the population, a numerical analysis from an NCI-provided dataset ranging from years 1975 till 2015 was performed. The year at which there was a maximum of total lung-cancer affected rate was identified at 1992. The data was then split at this critical year and analyzed with a parabolic fit. The derivative of the fit was calculated to compare the rate of increase and decrease. The difference between these two rates was –0.0126. This means that the rate of increase was faster than the rate of decrease A discussion of data limitation and major historic events were provided to better contextualize this trend.

**Background**

Lung and bronchus cancer can be characterized by malignant, uncontrollable cell growth in the areas of the lung and bronchus (Cruz, C. et al 2011). There are many ways for diagnosing lung cancer currently – some of which include X-ray imaging, sputum cytology, and a lung biopsy where a small amount of tissue is removed and tested (Mayo Clinic). This aggressive cell growth while often causing mortality, can also be survived. Available treatments in this age include tumor resection, radiation therapy, chemotherapy, radiosurgery, and targeted drug delivery and immunotherapy. These diagnoses and treatments have heavily evolved in the last few decades, and have taken us bounds closer to preventing this disease. However, a way to also understand this disease more deeply is to take a look at past data and try to learn about the trends of lung and bronchus cancer.

The National Cancer Institute’s Surveillance, Epidemiology, and End Results Program provides public access to SEER\*Explorer, an interactive tool for US cancer statistics. *Incidence* and *Mortality* are two rate values provided. The *Incidence* value is the number of newly diagnosed cancers of a specific cancer per 100,000 population at risk. The primary site only includes the site of origin. On the other hand, cancer *Mortality* is the number of reported cancer deaths for a specific cancer per 100,000 population at risk. This data is provided unanalyzed, but with adjustable parameters. With this, a characterization of the interesting data points and the surrounding area around it can be made to further understand how lung and bronchus cancer rates may be depressed in the future.

**Materials and Methods**

*Source of dataset*

Data was sourced from the National Cancer Institute under the Surveillance, Epidemiology, and End Results Program. SEER Incidence and U.S. Mortality series were chosen for Lung and Bronchus Cancer Sites from 1975-2015 and included both sexes, all ages, and all races:

<https://seer.cancer.gov/faststats/selections.php?#Output>

*Software*

MATLAB\_R2017b was used for data analysis, table formation, and figure creation.

*Equations*

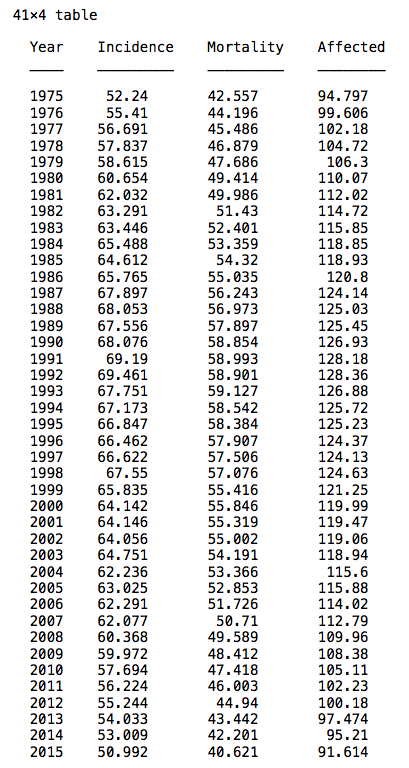
General formula for average:

Sum of Squares Error:

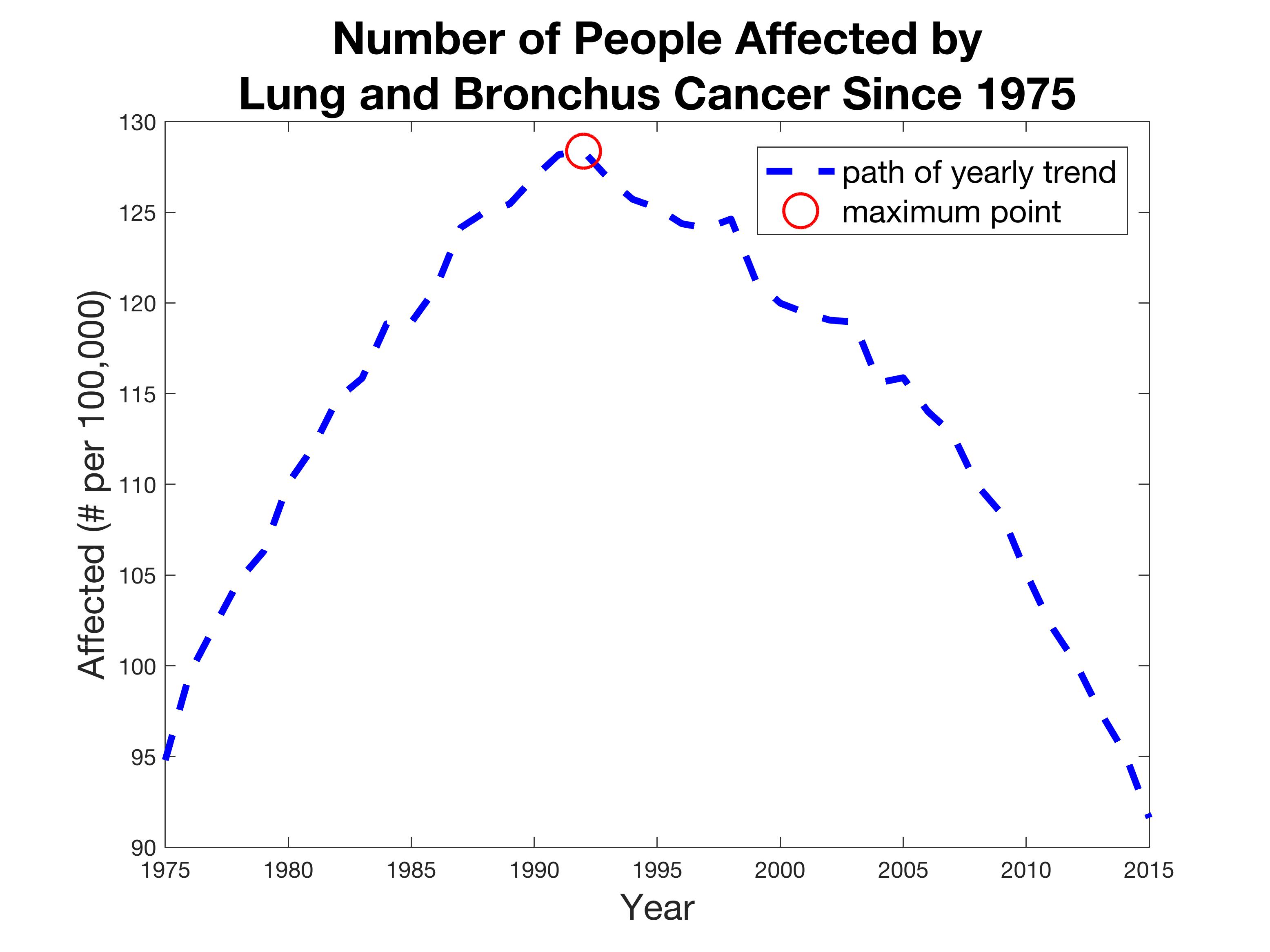
Standard deviation of the distance between data and fit:

Rules for 2nd order polynomial derivatives:

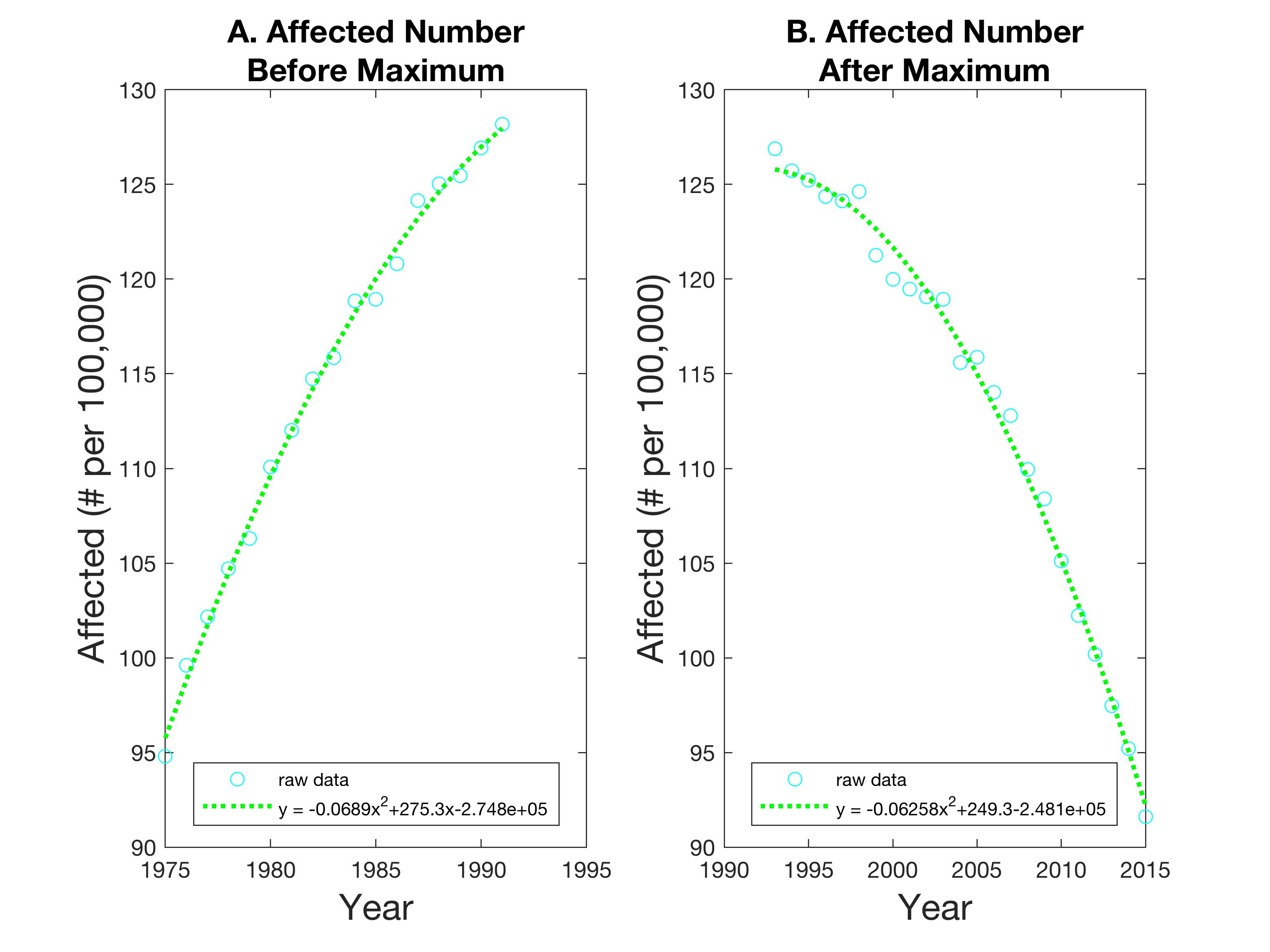
**Analysis**



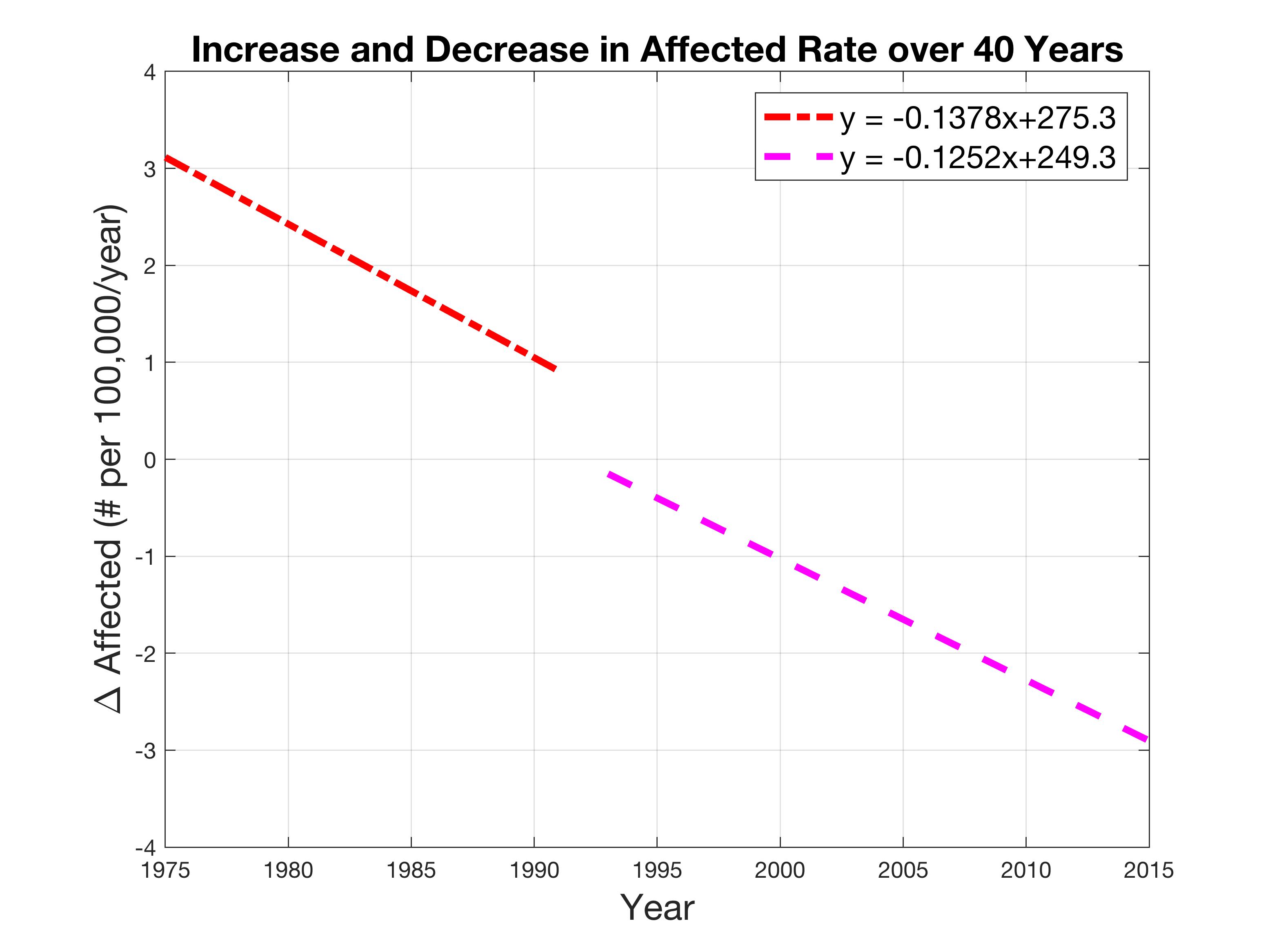
**Table 1.** Raw data was organized chronologically. Lung and bronchus cancer *Incidence* was added to *Mortality* to generate the *Affected* column.



**Figure 1.** Graph of all affected people (per 100,000) with respect to each year is shown by the blue dotted line. The maximum number of affected people in one year is highlighted by the red circle.



**Figure 2. (A)** Raw data before 1992 was plotted and subsequently fitted with a quadratic model represented by the green dotted line. The affected number is increasing, but with a concave down curvature. **(B)** Raw data after 1992 was plotted and subsequently fitted with a quadratic model represented by the green dotted line. The affected number is decreasing, also with a concave down curvature.



**Figure 3.** Derivatives of the two fits were plotted to generate the change in affected numbers over the years, with two distinct equations representing the derivative of the data for before and after the maximum point at 1992. Both slopes were negative.

Raw data was first organized in a table (Table 1). *Incidence* and *Mortality* rate (per 100,000) were added to create the *Affected* array. This array was used for all proceeding analysis.

This critical year (maximum point) is **1992** (Fig. 1) and its value is **128.4** people (per 100,000), which calculates to a lung and bronchus cancer appearance of about 0.13%. The average incidence number is 62.4, the average mortality number is 52.1, and the average affected number is 114.5 (data shown in code). While these numbers seem not very large, they have huge implications when applied beyond the *per 100,000* parameter to the whole population of the United States.

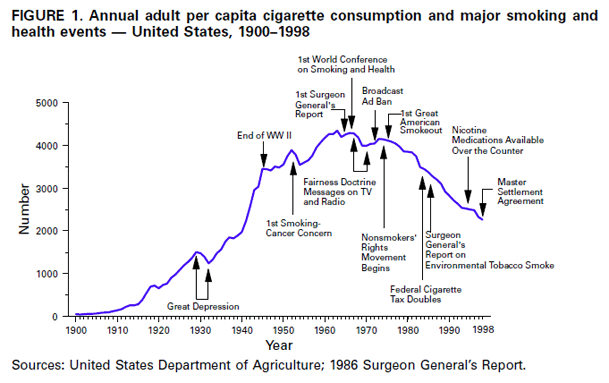
The two graphs (Fig. 2 A & B) are a replot of the raw data on either side of the critical year, but with a quadratic fit overlaid. Visually, the two sides of the data appeared to follow a quadratic fit, so it was chosen over a linear for more accurate analysis of rate of change. The sum of squares error for Graph A was **6.788** and Graph B was **15.903**. The standard deviation of the distance between the fit and data was **0.163** for Graph A and **0.181** for Graph B.Graph A shows a large boom in lung cancer that heavily increases to the maximum point. However, the rate itself levels off – hence the parabolic nature. Graph B shows the opposite: the lung cancer rates appear to decrease, at first quite slowly and then at a much faster rate. This leads into an analysis of the derivatives to quantify this.

The two lines plotted from the derivatives of the fits for before and after 1992 demonstrate the differences in the increase and decrease around the maximum point (Fig. 3). The slope for Graph A was -0.138 and for Graph B was -0.125. The difference between the two rates of change (rate before – rate after) was **-0.0126**.

**Discussion**

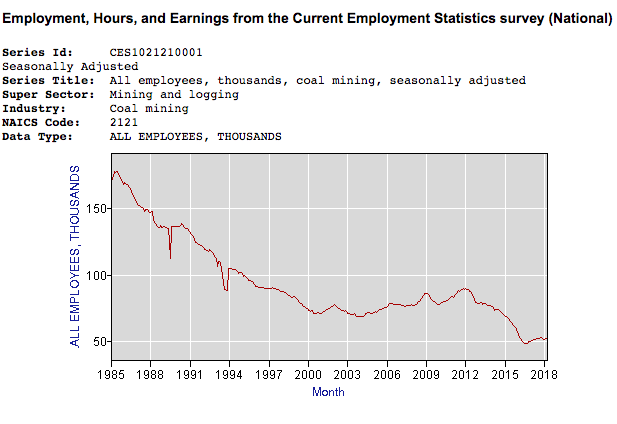
Since R-squared values are insufficient for goodness of fit analysis in nonlinear models (Spiess, A. N. et al, 2010), other methods must be used to evaluate the quadratic fitting to the two areas of the raw data before and after the maximum point. A simple look at the sum squares error and standard deviation let us know the relative and concrete deviations of the fit. Graph B has more variation than Graph A, but it’s standard deviation is an extremely small fraction of the smallest value, so the parabolic fit can be considered a good match. The difference between the two rates is negative, showing that the rate of increase to the maximum is going faster than the rate of decrease after. The reason or correlation for this cannot be exactly identified, so literature must provide back up.

To further understand this data, historical contextualization must occur. From what is known about lung cancer, some contradicting but also convincing correlations can be linked to the tobacco industry. The following graph from the CDC sourced from the United States Department of Agriculture shows a decreasing number of adult per capita from 1975 onwards.



While the decrease contradicts the increase in lung cancer data pre-1992, what strikes as interesting is the point at which nicotine medication became available over the counter. This point lies about right after the maximum year, 1992, we identified previously. The decrease then correlates with the decrease in lung cancer post-1992. The slower decrease might be related to the slow acceptance of these nicotine medications by everyone.

Another similar trend is found in the coal mining industry. Coal-mining has been notorious for causing lung cancer due to the silica dust and organic compounds involved (Jenkins, W.D. et al 2013). The following graph was taken from the US Bureau of Labor Statistics:



Like the tobacco data, the data pre-1992 does not correlate with the lung cancer data, however, post 1992 does generally decrease, like the lung data.

There are some limitations of the data that must be observed. It is important to note the possibility of cancer incidence and cancer mortality having a data overlap. This point may cause discrepancy in the true value of the number of people affected, but the overall trend would not be altered much. Additionally, understanding the situations at which data was collected would lead us to be aware of any noticeable biases that may skew the data in a certain way. It may be that lung cancer was not as thoroughly recorded and diagnosed in the earlier years of 1975-1992, creating lower numbers that increased to the maximum. Meanwhile, the consumer and occupational data can be quite accurate since methods of determining a worker, or sales of a product, has not changed much over the years.

**Conclusions**

Although NCI does not provide direct analysis of its own to go along with the dataset, with comparison to correlations to historical statistics of the tobacco and coal-mining industry, areas after the maximum point of 1992 agree. This confirms the scientific links already found between lung cancer and these two points. The parabolic fit used matches well to the data according to the SSE and standard deviation analysis. Some additional information gathered includes the rate of increase and decrease close to the maximum point. Reading into this, some inferences can be made about sudden market incidences and adoption of treatment. Limitations of the data were also acknowledged.

A number of factors could be explored to expand the information garnered from this particular dataset. For example, racial and gender disparities in lung cancer rates are stark (Meza, R. et al, 2015). A histological analysis can also reveal more about the nature of the disease spread. The four major types are squamous cell carcinomas, adenocarcinomas, neuroendocrine carcinomas, and large cell carcinomas (Novaes, F. T., et al 2008). Clustering these categories could provide more insight into lung and bronchus cancer trends for future treatment.

**References**

1. SEER\*Explorer: An interactive website for SEER cancer statistics [Internet]. Beta Version. Surveillance Research Program, National Cancer Institute. [Cited 2018 May 4]. Available from <https://seer.cancer.gov/explorer/>.
2. Lung Cancer: Diagnosis. [Internet]. Mayo Clinic. [Cited 2018 May 4]. Retrieved from: <https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627>
3. Cruz, C. S. D., Tanoue, L. T., & Matthay, R. A. (2011). Lung cancer: epidemiology, etiology, and prevention. Clinics in chest medicine, 32(4), 605-644.
4. Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC pharmacology, 10(1), 6.
5. Flaherty, L. (2000). Achievements in public health, 1900-1999: tobacco use-United States, 1900-1999. Journal of Emergency Nursing, 26(2), 150-151.
6. Index, C. P. (2011). US Bureau of labor statistics. Washington, DC.
7. Jenkins, W. D., Christian, W. J., Mueller, G., & Robbins, K. T. (2013). Population cancer risks associated with coal mining: a systematic review. PloS one, 8(8), e71312.
8. Meza, R., Meernik, C., Jeon, J., & Cote, M. L. (2015). Lung cancer incidence trends by gender, race and histology in the United States, 1973–2010. PloS one, 10(3), e0121323.
9. Novaes, F. T., Cataneo, D. C., Ruiz Junior, R. L., Defaveri, J., Michelin, O. C., & Cataneo, A. J. M. (2008). Lung cancer: histology, staging, treatment and survival. Jornal Brasileiro de Pneumologia, 34(8), 595-600.