



# **Statistical Modeling to Predict the Trend in Lung Cancer Data in Utah using Joinpoint Regression Analysis**

**Riley Morgan, Gregory Schmidt**

Advisors: Dr. Md Sazib Hasan & Dr. Vinodh Chellamuthu



# Outline

---

Our goals

Brief background on Lung Cancer in Utah

Statistical models

Results per category

Conclusion

Future work



# Our Goals

---

- We obtained data from the Surveillance Epidemiology and End Results Program website (SEER)
- We used the statistical methods:
  1. Linear Regression
  2. Time Series Analysis-ARIMA
  3. Joinpoint Regression
- Our simulation results will predict the trend of new total deaths



# Lung Cancer

---

- Lung and bronchus cancer is the leading cause of cancer related deaths in Utah and in the United States
- In 2021, it is estimated that there was 131,880 deaths from lung cancer in the United States
- Nearly 90% of lung cancer deaths can be attributed to smoking



# Model Advantages

---

## Linear Regression:

- Simple
- Visually easy to understand

## Time Series Analysis-ARIMA:

- Identifies and forecasts patterns over time
- Predicts future trends within a confidence interval

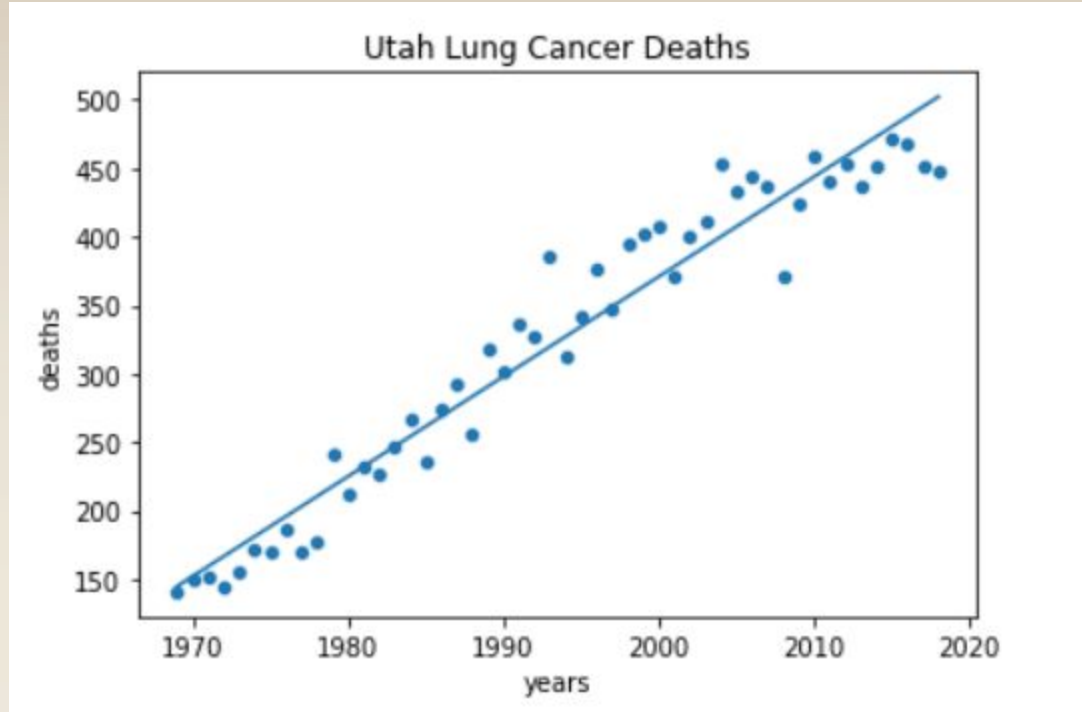
## Joinpoint Regression:

- Splits data into trend intervals



# Results: Linear Regression of Lung Cancer Deaths

---



- By using linear regression, we get one simple line to fit the model
- In 2030, this model predicts 590 deaths

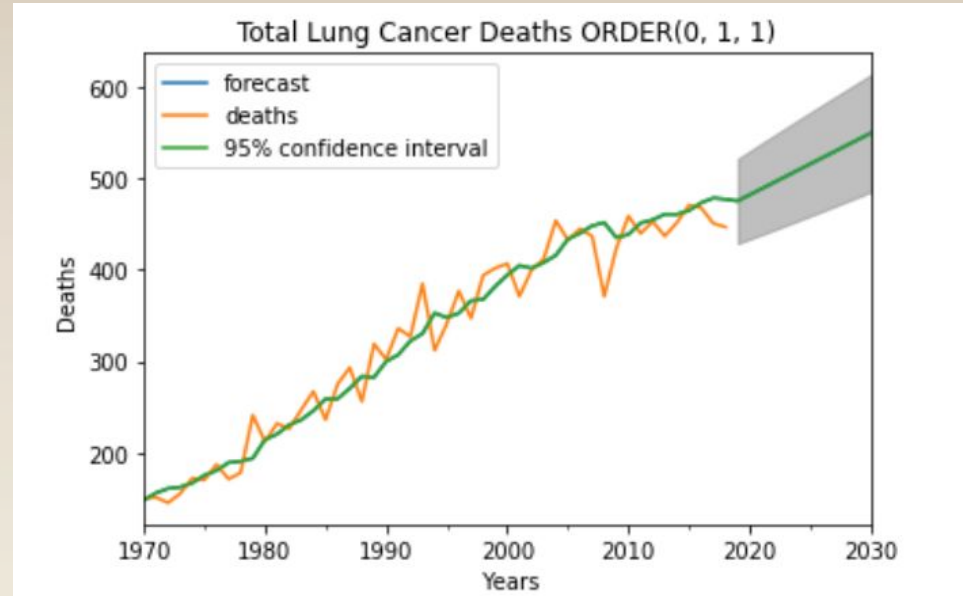


# Time Series Analysis-ARIMA

ARIMA(p, d, q) can be used to make a future prediction

- p for the AR [Auto Regressive] term
- d for the I [Integrated] term
- q for the MA [Moving Average] term

Est. deaths by 2030: 550





# What is Joinpoint? Why Joinpoint?

---

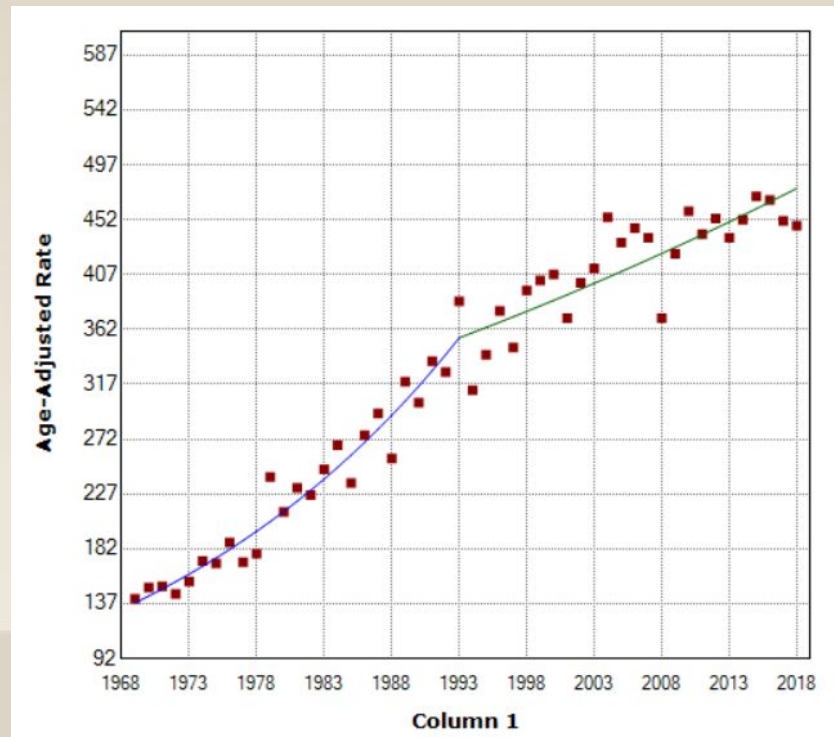
- Piecewise linear regression model
- Calculates and searches for peak points in trends
- Great for analyzing trends over time



# Joinpoint Regression

Joinpoint Regression takes data and fits a model for a specific number of joinpoints

By using one joinpoint, a clear upward trend can be observed



# How to Determine the Number of Joinpoints

Model Selection Method							
Model Selection Method							
Permutation Test							
Test For Number of Joinpoints							
Test Number	Null Hypothesis	Alternate Hypothesis	Numerator Degrees of Freedom	Denominator Degrees of Freedom	Number of Permutations	P-Value	Significance Level~
#1	0 Joinpoint(s)	7 Joinpoint(s) ^	14	34	4500	0.0002222	0.0071429
#2	1 Joinpoint(s) ^	7 Joinpoint(s)	12	34	4500	0.7404444	0.0083333
#3	1 Joinpoint(s) ^	6 Joinpoint(s)	10	36	4500	0.7328889	0.0083333
#4	1 Joinpoint(s) ^	5 Joinpoint(s)	8	38	4500	0.7768889	0.0083333
#5	1 Joinpoint(s) ^	4 Joinpoint(s)	6	40	4500	0.6551111	0.0083333
#6	1 Joinpoint(s) ^	3 Joinpoint(s)	4	42	4500	0.6048889	0.0083333
#7	1 Joinpoint(s) ^	2 Joinpoint(s)	2	44	4500	0.3046667	0.0083333

^ Selected Model  
 ~ Significance level for individual test (alpha = 0.05).  
 Final Selected Model: 1 Joinpoint(s)

Uses the Permutation Test to compare each model

One Joinpoint was calculated to be the best fit for this model

# What is APC?

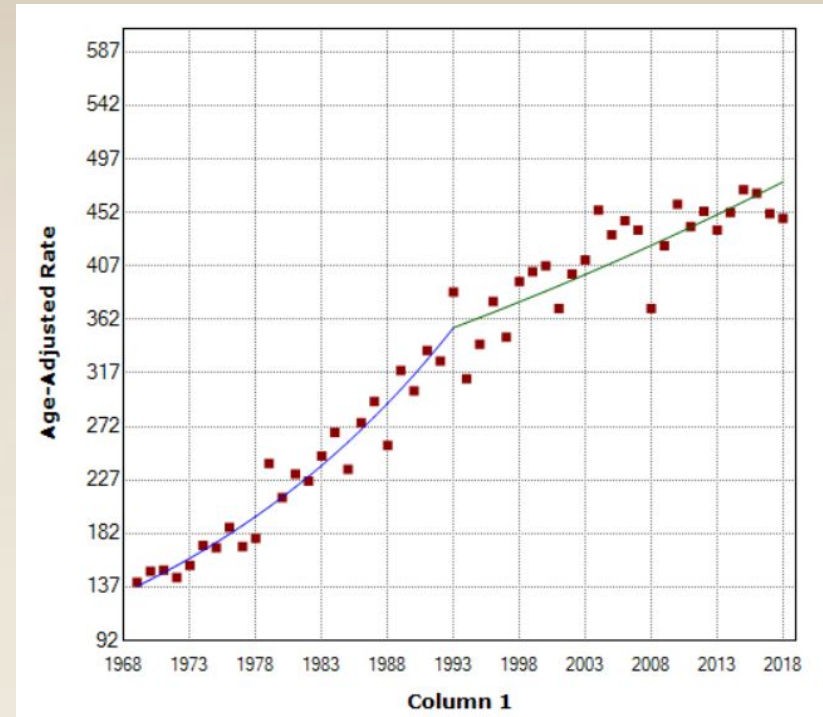
---

- Annual Percentage Change (APC) represents the degree of change over time
- Used for monitoring the growth rates of a specific set of data
- Positive values indicate a growth
- Negative values indicate a decay

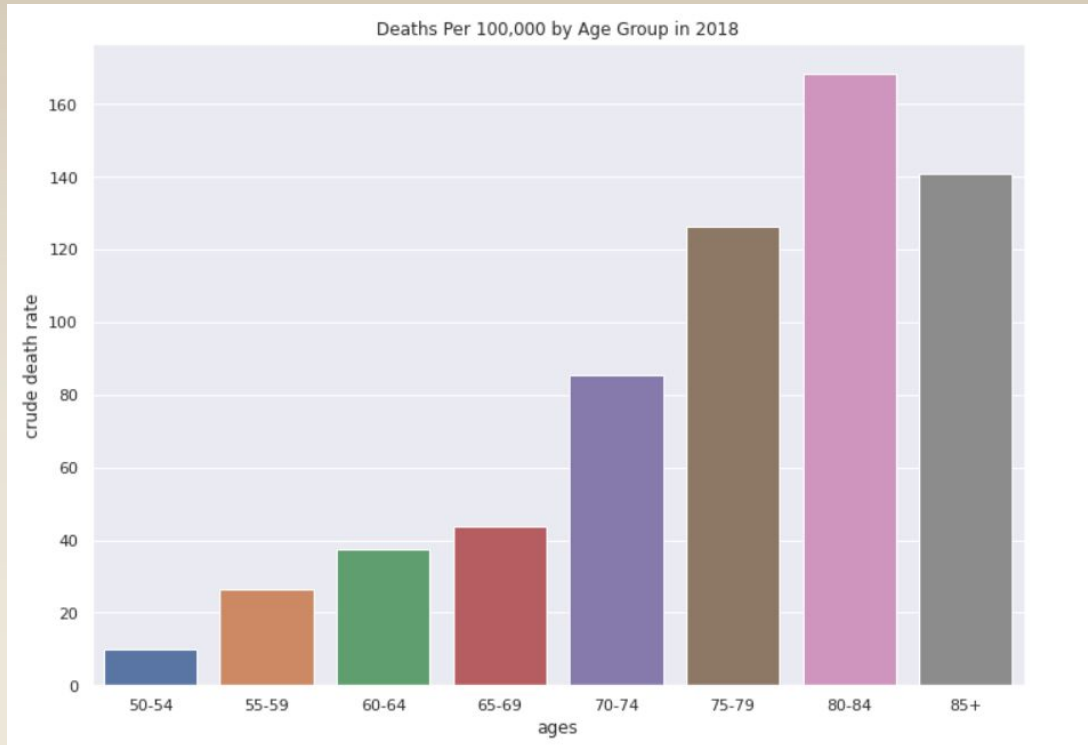


# Joinpoint - Annual Percent Change

- Our model is split into two trends
- Model has an average APC of 2.6
- The APC from 1969-1993 is 4.0
- The APC from 1993-2018 is 1.2
- Est. deaths in 2030: 551

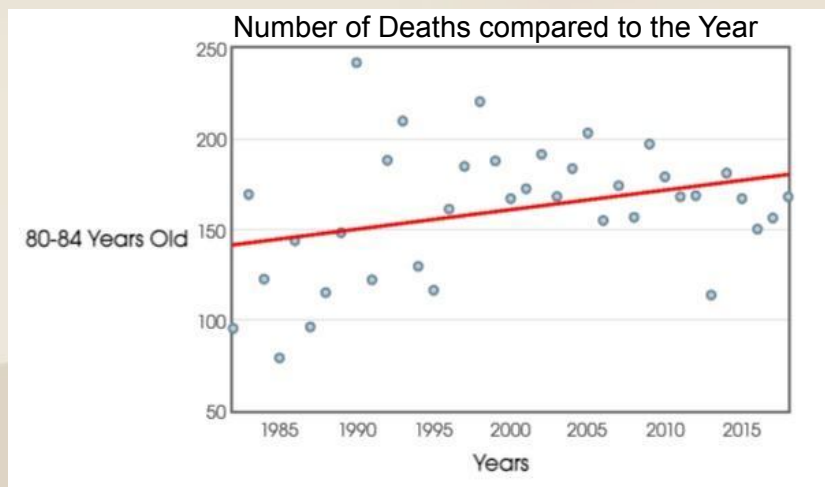
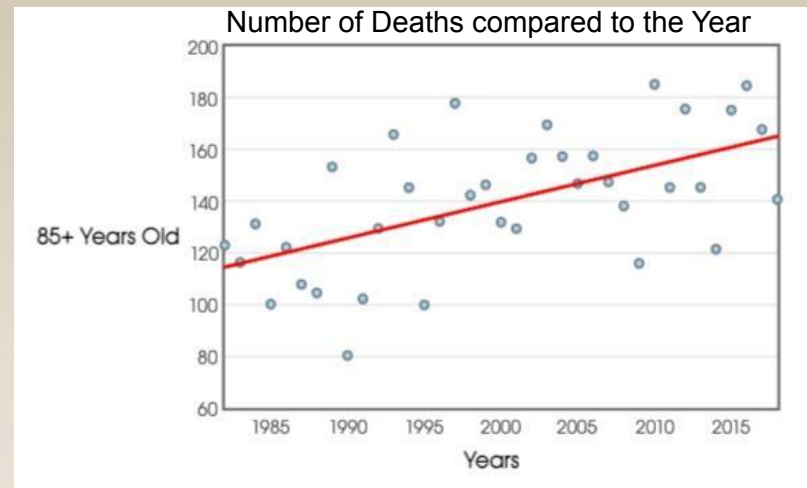
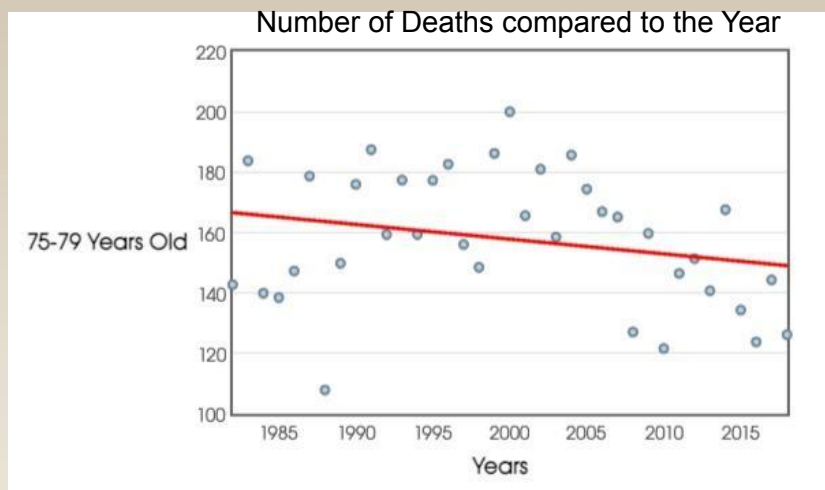


# Age Groups at Most Risk



As age increases, the risk of dying from lung cancer dramatically increases

# Age: Linear Regression (per 100k)



75-79 Years Old:

Slope =  $-0.49$

Correlation:  $r = -0.24$

Total by 2030 = 143

80-84 Years Old:

Slope =  $1.08$

Correlation:  $r = 0.32$

Total by 2030 = 193

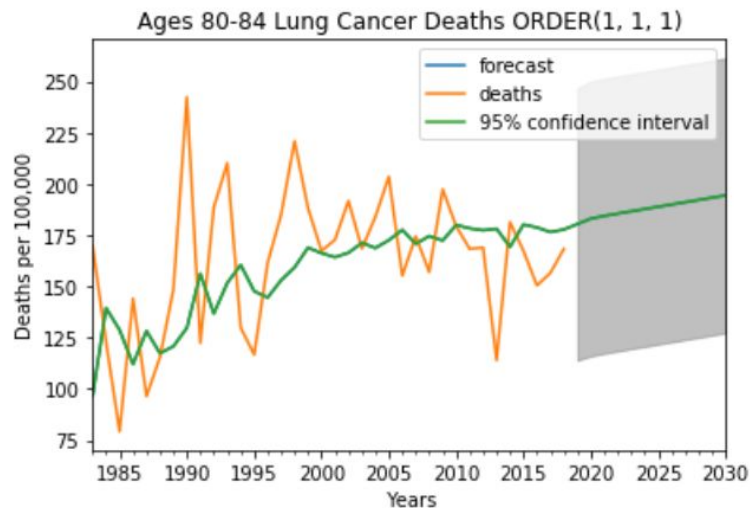
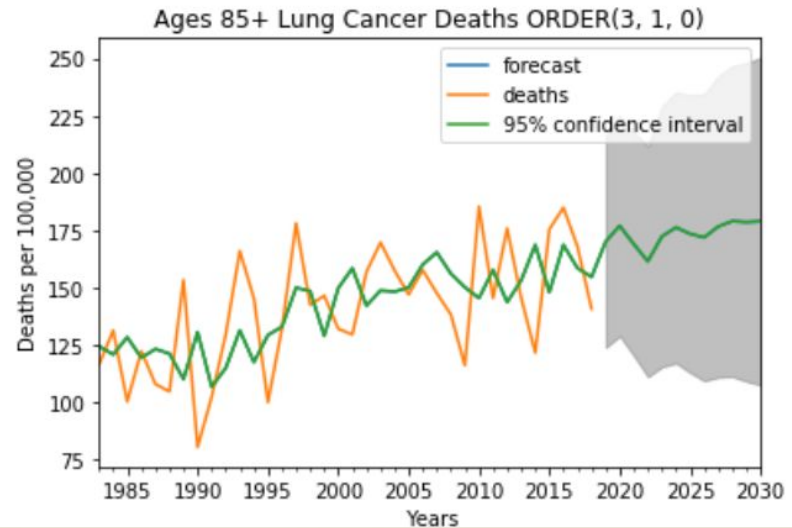
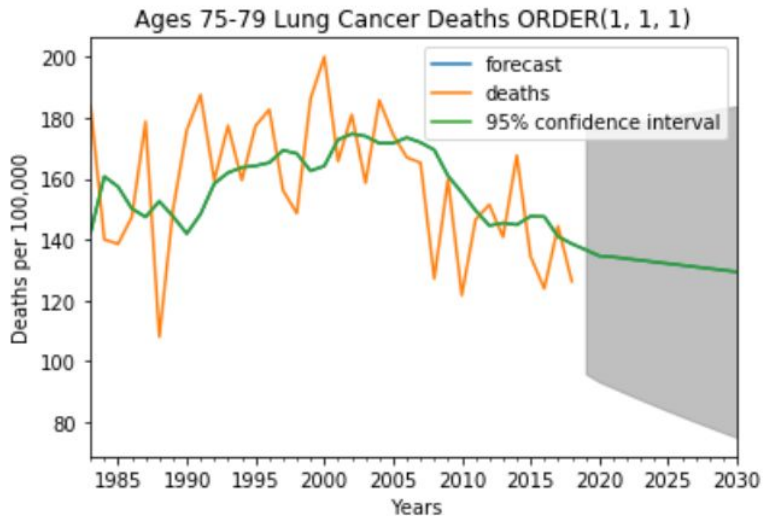
85+ Years Old:

Slope =  $1.41$

Correlation:  $r = 0.58$

Total by 2030 = 182

# Age: ARIMA (Deaths per 100k)

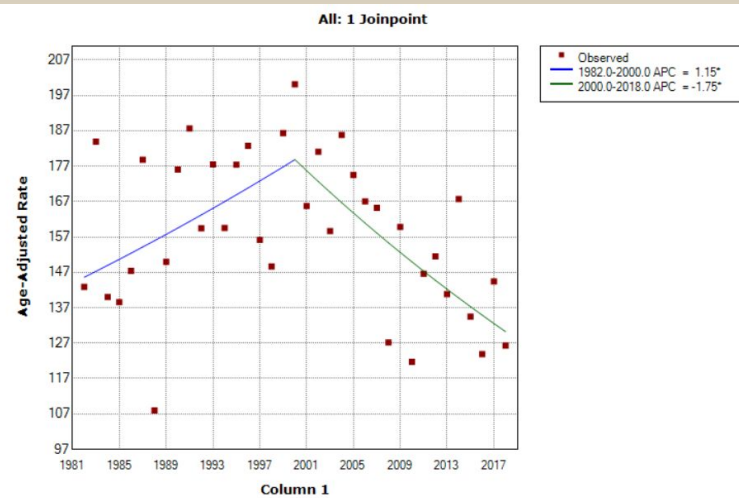


- A decrease for ages 75-79
- A steady increase for ages 80-84 and 85+

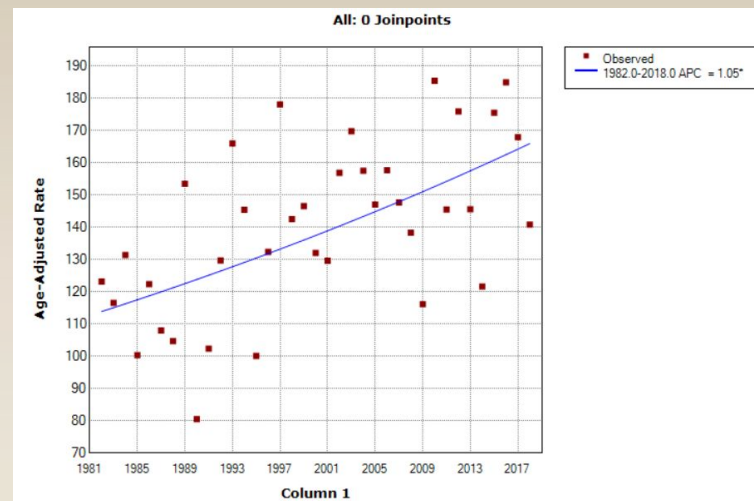


# Age: Joinpoint (Deaths per 100k)

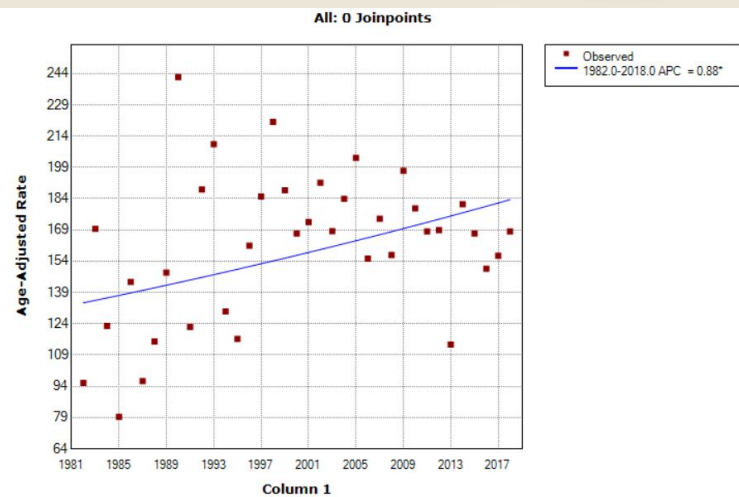
## Ages 75-79



## Ages 85+



## Ages 80-84

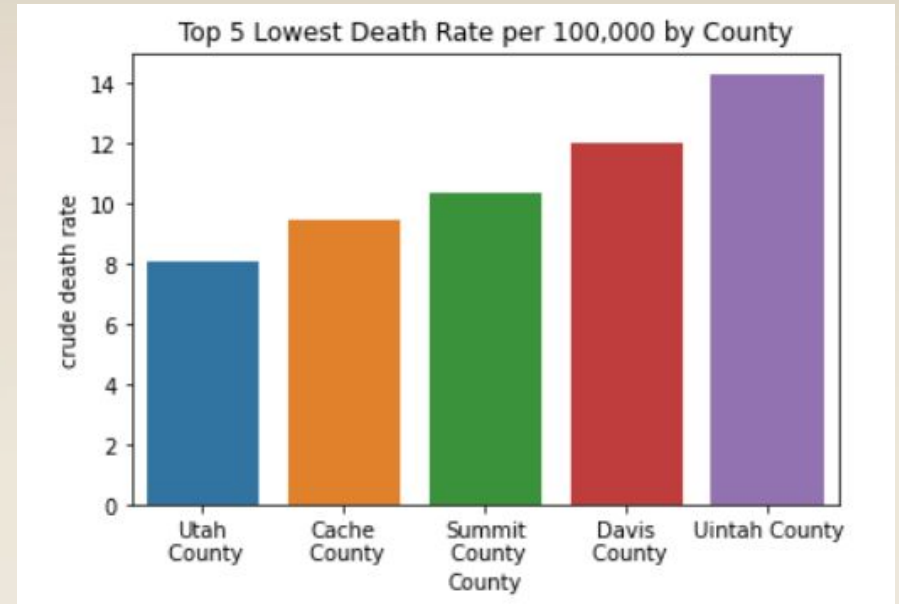
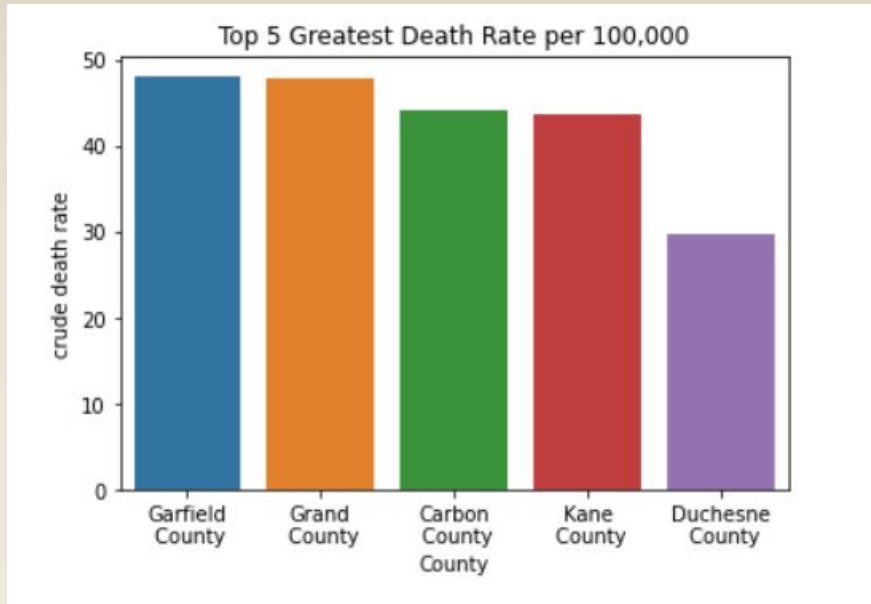


## In Year 2030

- Ages 75-79: 105 deaths
- Ages 80-84: 204 deaths
- Ages 85+: 188 deaths

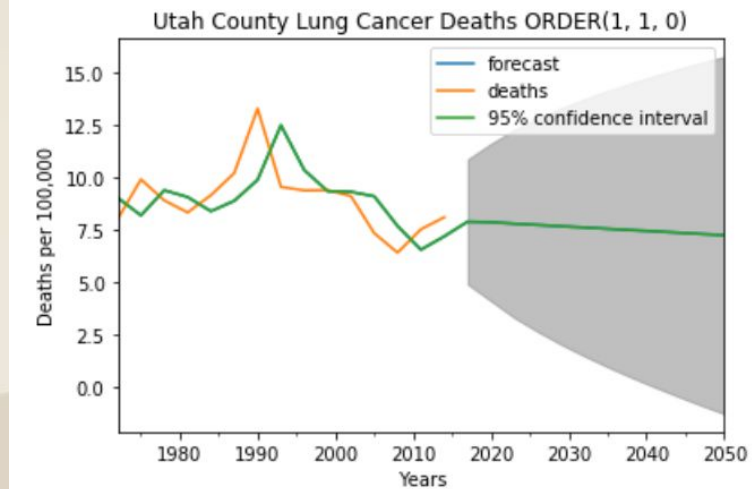
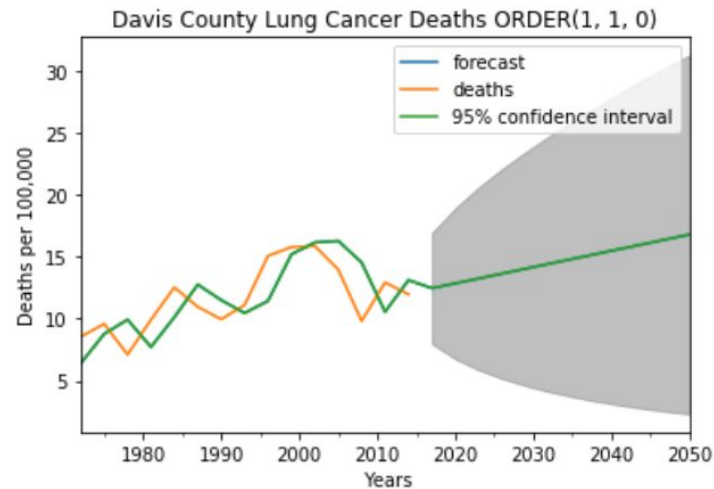
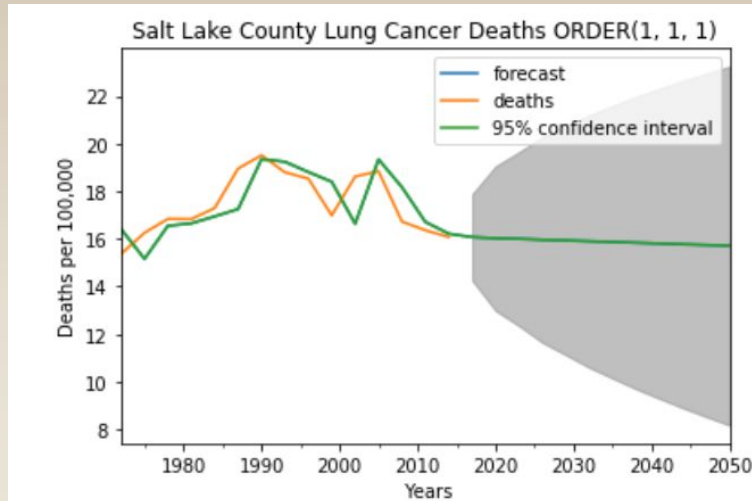
# Utah Counties at Most Risk

Average deaths per 100,000 by county with at least 10 deaths from 2014-2018



Counties with larger populations usually have smaller mortality rates

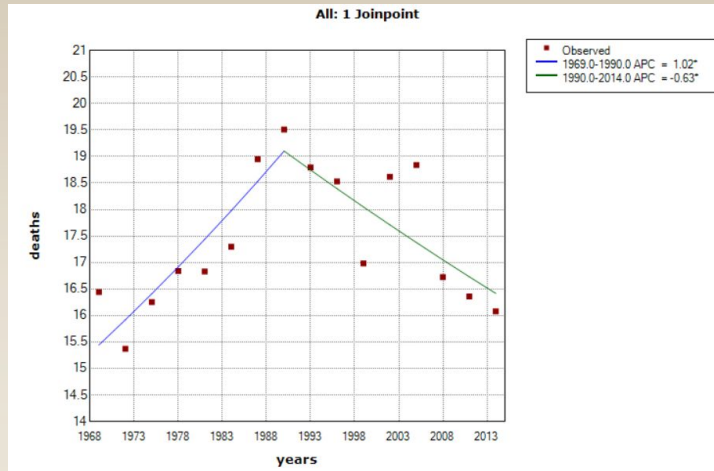
# County: ARIMA (Deaths Per 100k)



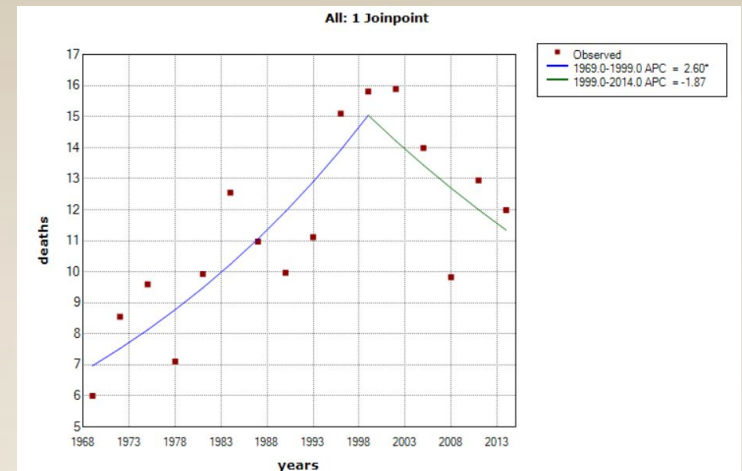
- A slight downward trend for Salt Lake County and Utah County
- Steady upward trend for Davis County

# County: Joinpoint (Deaths Per 100k)

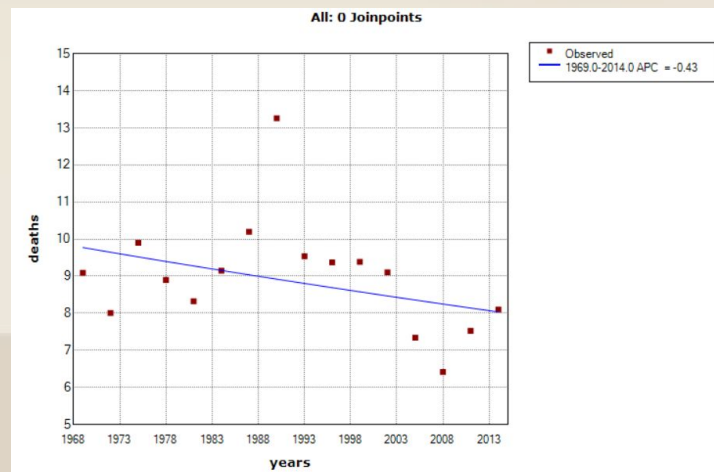
## Salt Lake County



## Davis County



## Utah County



## In Year 2030

- Salt Lake County: 16 deaths
- Davis County: 11 deaths
- Utah County: 8 deaths

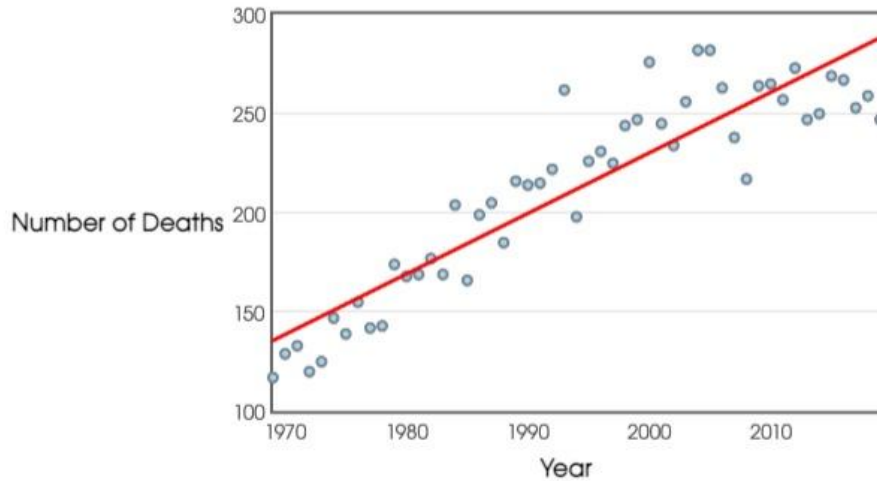
# Gender

---

- 56.7% of Utah lung cancer deaths were males in 2018
- A 121% increase from 1969 to 2018 for male lung cancer deaths
- A 662% increase for female lung cancer deaths



# Gender: Linear Regression

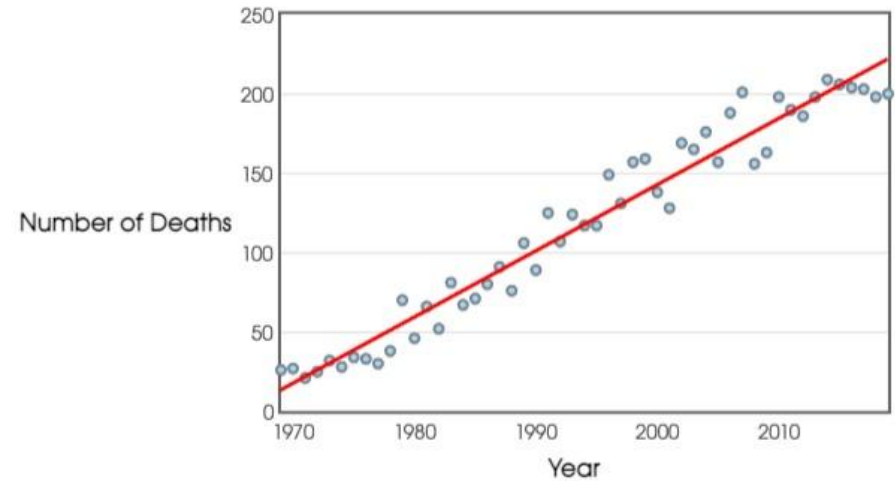


## Men

Slope = 3.1

Correlation:  $r = 0.91$

Total Deaths in the year 2030 = 322



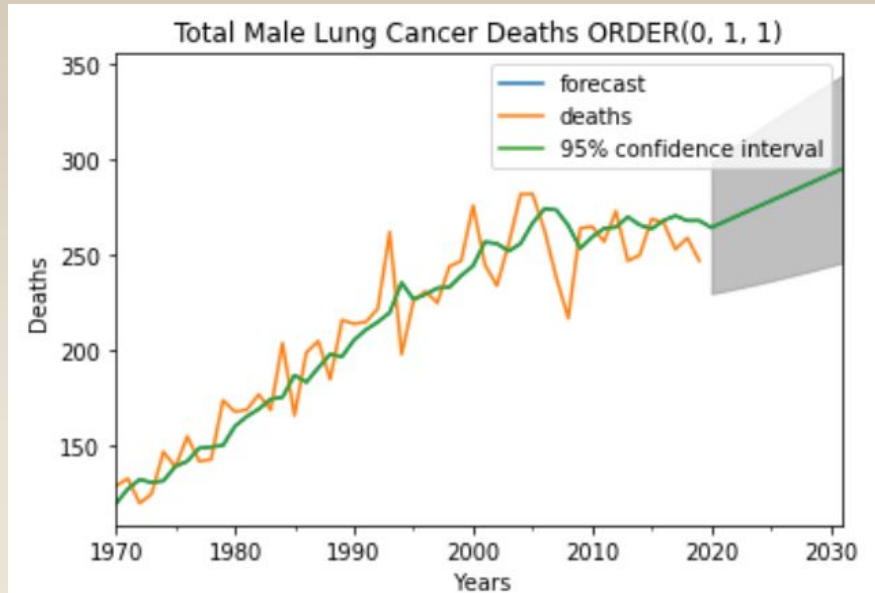
## Women

Slope = 4.2

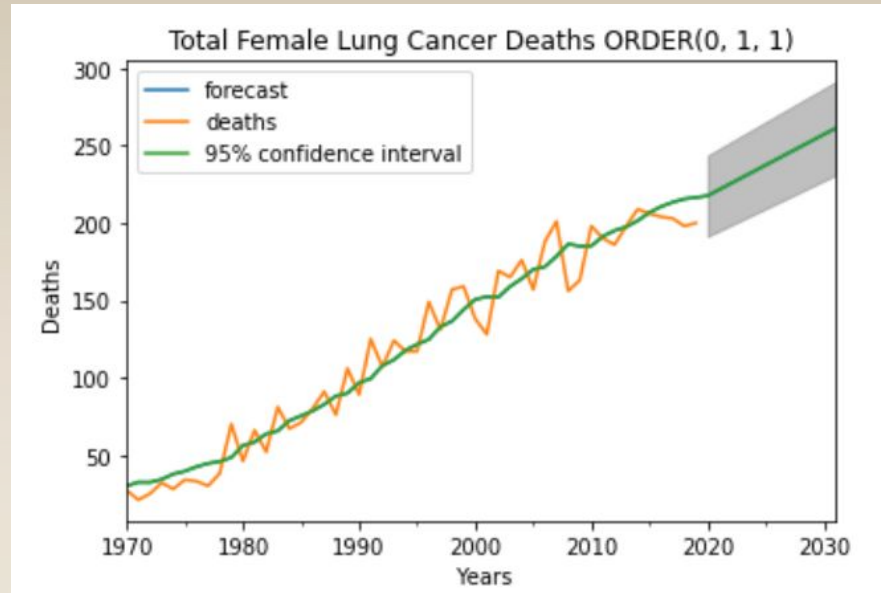
Correlation:  $r = 0.98$

Total Deaths in the year 2030 = 268

# Gender: ARIMA



Est. Deaths by 2030: 292



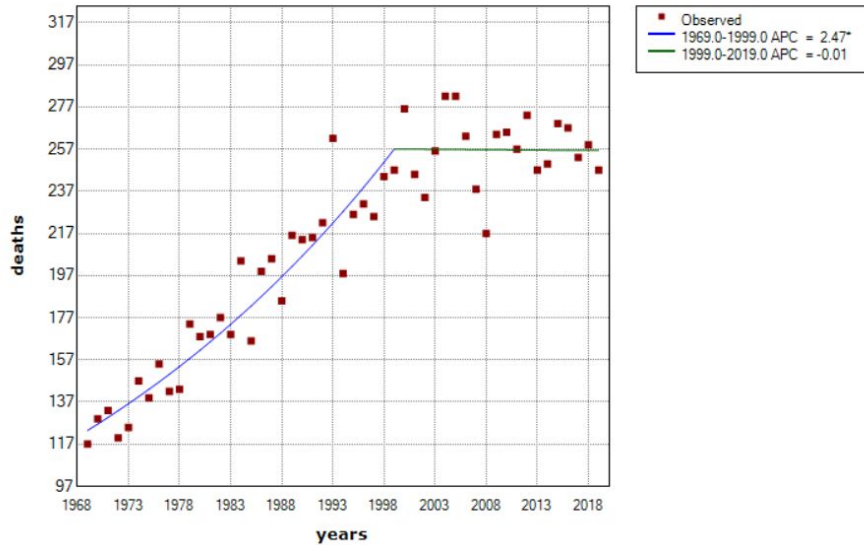
Est. Deaths by 2030: 257



# Gender: Joinpoint

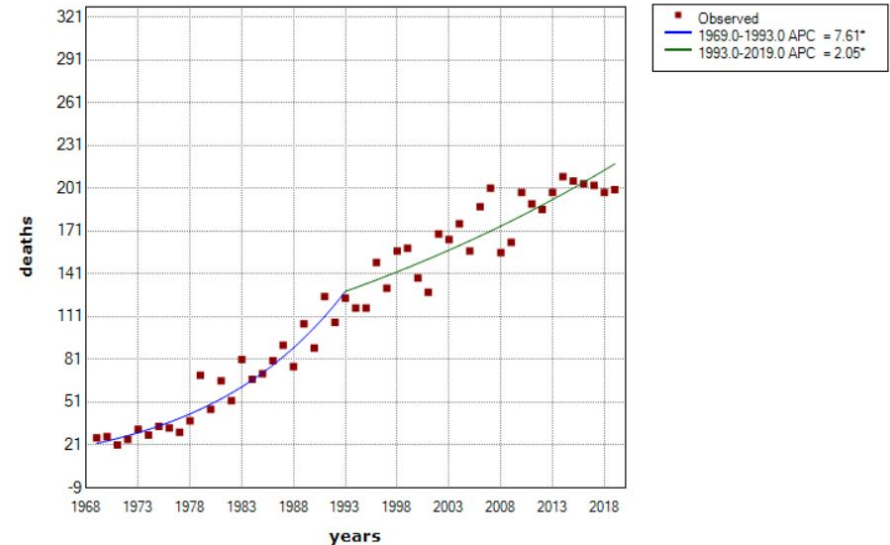
Male

All: 1 Joinpoint



Female

All: 1 Joinpoint



Year 2030:

Males: 256 deaths

Females: 273 deaths



# Conclusion

---

- Almost all models show an increase in total deaths
- The most populated counties show a decreasing trend in deaths per 100k
- Male deaths indicate a stable trend
- Female deaths indicate an increasing trend



# Future Work

---

- Look into incidence rates
- Compare incidence rates to mortality rates
- Explore trends within ethnicity demographics
- Use LSTM Model as a comparison
- Publish research paper



# Acknowledgement

---

Funding: DSU Office of Undergraduate Research

