Comparing Interpolation and Regression Techniques in Python Using Alzheimer's Patient Health Data from the CDC

Willow Leach

April 23 2023

Introduction

Prediction is a common goal in medical studies and is often motivated by gaps in current research. Alzheimer's disease — a progressive neurodegenerative disease that leads to poor quality of life and an inevitable death — is full of such gaps in data as the target population of study (patients experiencing cognitive decline) is often difficult for researchers to contact (Field et al., 2019). While missing data is may not present significant issues in large samples, smaller data sets like those associated with Alzheimer's disease can magnify problems in testing efficiency, accuracy, and precision (Miao et al., 2021). These problems can lead to a domino-effect for many other areas of the research process including but not limited to "information loss" and "result bias" (Miao et al., 2021).

A method often used to counteract these gaps in research is interpolation, a mathematical technique that utilizes various polynomial functions to estimate missing values in a bounded data set. However, interpolation alone is not sufficient to draw conclusions about trends in data sets; long-term predictions regarding the relationship between independent and dependent variables do not spontaneously appear just because there is more data to work with. Thus, another technique (e.g. least-squares regression) must be used as a supplement to interpolation. Least-squares regression will minimize error between original data set values and values estimated with interpolation to ensure accuracy in trend predictions. However, this method cannot be used in place of interpolation as the mechanics of regression can lack the precision interpolation brings to the table.

This paper aims to make use of both interpolation and least-squares regression techniques to fill gaps in Alzheimer's data gathered by the Center for Disease Control (CDC) to estimate trends between poor physical health and cognitive decline among older Americans (ages 65+).

Theory

Interpolation

Interpolation is a general technique used on numerical data to estimate potential points a data set. While not in the given set of values, these point estimations are values that would likely be a part of that data pool given a larger sample size.

Linear interpolation estimates points in a given set's domain by (in an intuitive sense) drawing straight lines between adjacent points and seeing what other points lie on those lines. This notion is represented by the following equation:

$$\hat{y} = y_i + \frac{(y_{i+1} - y_i)(x - x_i)}{x_{i+1} - x_i},$$

where x represents the independent (x) value from the original data set's domain that we want to estimate

a potential dependent (y) value for (given that one does not already exist), x_{i+1} and y_{i+1} being the right-adjacent x and y values from the original data set, x_{i-1} and y_{i-1} being the left-adjacent values, and $\hat{y}$ representing the final estimated y value.

Cubic interpolation follows the same concept as linear interpolation, with the exception of what kind of function is "drawn" between each point from the sample. As the name suggests, cubic interpolation uses a third-degree polynomial rather than a linear equation to estimate potential values to add to the raw data. These cubic functions are collected into a single piece-wise function (cubic *spline*).

To make this piece-wise cubic curve, four equations are required:

$$S_i(x_i) = y_i, \quad i = 1, ..., n-1 \tag{1}$$

$$S_i(x_{i+1}) = y_{i+1}, \quad i = 1, ..., n-1 \tag{2}$$

$$S_i^{'}(x+1) = S_i^{'}(x+1), \quad i = 1, ..., n-2 \tag{3}$$

$$S_i^{''}(x+1) = S_i^{''}(x+1), \quad i = 1, ..., n-2, \tag{4}$$

where i is the index of a point from the original data set and S is the cubic spline function.

Equations 1 and 2 represent the intersection of each two adjacent data points as parameters for the estimations. Equations 3 and 4 give S cubic behavior, requiring the first and second derivatives of S to be continuous at each data point $i = 2, ..., n-1$, defining the second derivative to be zero at each original data point (i.e. creating the maxima of the cubic spline curve).

Least Squares Regression

Least squares regression is a statistical method used to best represent trends in data.

While both interpolation and regression create functions to represent data points, regression builds functions to model relationships between independent and dependent variables rather than to estimate dependent values given the independent variable.

The following equations represent the least squares regression method of modelling a set of points:

$$\hat{y}(x_m) = \alpha_1 f_1(x_m) + ... + \alpha_m f_m(x_m)$$

$$\beta = (A^T A)^{-1} A^T Y,$$

where $\hat{y}$ represents the estimation of data at each point in the set and β represents the minimization of error of the estimations from $\hat{y}$.

Given the equations for least squares regression, an important difference to note between interpolation and regression is regression considers all points from a data set to make a single function rather than adjacent points to make a piece-wise function.

Implementation

The independent and dependent variables to be statistically analyzed in this paper as previously described are time in years (x) and percentage of Americans ages 65 and above (y) experiencing cognitive decline and/or poor physical health, respectively. In other words, the relationship between poor physical health and cognitive decline will be explored from a sample of "older" Americans over time by means of interpolation and least squares regression. This implementation is described below:

```
Data Cleaning
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.interpolate import interp1d, CubicSpline
from scipy import optimize

# Read in data
df = pd.read_csv('/content/Alzheimers.csv')

# Combine 'YearStart' and 'YearEnd' column
df_long = pd.melt(df, id_vars = ['Data_Value', 'Data_Value_Type', 'Topic', 'LocationDesc'],
    value_vars = ['YearStart', 'YearEnd'], var_name = 'Start/End ID', value_name = 'Year')

# Sort 'Year' in ascending order
df_sorted = df_long.sort_values(by = 'Year')

# Filter out NaN values
df_sorted_NaN = df_sorted.dropna(subset=['Data_Value'])

# Filter for percent data type
df_sorted_NaN_percent = df_sorted_NaN[df_sorted_NaN['Data_Value_Type'] == 'Percentage']

# Filter rows for physical and cognitive health
df_physical = df_sorted_NaN_percent[df_sorted_NaN_percent['Topic'].str.contains('(fair to poor
    health)|unhealthy')];
df_cognitive = df_sorted_NaN_percent[df_sorted_NaN_percent['Topic'].str.contains('cognitive
    decline')];

# Group by states for comparison of trends
#df_physical_states = df_physical.groupby(['LocationDesc'])
#df_cognitive_states = df_cognitive.groupby(['LocationDesc'])

# Extract independent and dependent variables for simplicity
years_physical = np.array((df_physical['Year']))
years_cognitive = np.array((df_cognitive['Year']))

percent_physical = np.array((df_physical['Data_Value']))
percent_cognitive = np.array((df_cognitive['Data_Value']))

"""Interpolation Calculation"""

# Years for Interpolation
years_interpol = np.arange(2015, 2020, 1 / 12)
                    # from 2015 to 2020 w/ 1 month intervals

# Linear Interpolation
lin_interpol_physical_func = interp1d(years_physical, percent_physical)
                    # PHYSICAL health linear interpolation
lin_interpol_cognitive_func = interp1d(years_cognitive, percent_cognitive)
                    # COGNITIVE health linear interpolation

percent_physical_lin = lin_interpol_physical_func(years_interpol)
                    # PHYSICAL health linear calc
percent_cognitive_lin = lin_interpol_cognitive_func(years_interpol)
                    # COGNITIVE health linear calc
```

```
51
# Find unique values of years and corresponding percents for Cubic Spline requirements
53 unique_years_physical, index = np.unique(years_physical, return_index=True)
                    # PHYSICAL health unique years
   unique_percent_physical = percent_physical[index]
                    # PHYSICAL health corresponding percent
55
   unique_years_cognitive, index = np.unique(years_cognitive, return_index=True)
                    # COGNITIVE health unique years
57 unique_percent_cognitive = percent_cognitive[index]
                    # COGNITIVE health corresponding percent

59 # Cubic Spline Interpolation
   cubic_interpol_physical_func = CubicSpline(unique_years_physical, unique_percent_physical,
       bc_type='natural')     # PHYSICAL health cubic spline interpolation
61 cubic_interpol_cognitive_func = CubicSpline(unique_years_cognitive, unique_percent_cognitive,
       bc_type='natural')  # COGNITIVE health cubic spline interpolation

63 percent_physical_cubic = cubic_interpol_physical_func(years_interpol)
                    # PHYSICAL health cubic spline calc
   percent_cognitive_cubic = cubic_interpol_cognitive_func(years_interpol)
                    # COGNITIVE health cubic spline calc
65
   """Regression Calculation"""
67
   # Define linear function for regression w/ linear interpolation dataset
69 def linear_func(x, a, b):
     return a * x + b
71
   # Define quadratic function for regression w/ linear interpolation dataset
73 def quadratic_func(x, a, b, c):
     return a * x**2 + b * x + c
75
   # Define cubic function for regression w/ cubic spline dataset
77 def cubic_func(x, a, b, c, d):
     return a * x**3 + b * x**2 + c * x + d
79
   # Fit linear regression for linear interpolation dataset
81 popt_physical_lin_linreg = optimize.curve_fit(linear_func, years_interpol, percent_physical_lin)
       [0]
   popt_cognitive_lin_linreg = optimize.curve_fit(linear_func, years_interpol, percent_cognitive_lin
       )[0]
83
   # Fit quadratic regression for linear interpolation dataset
85 popt_physical_lin_quadreg = optimize.curve_fit(quadratic_func, years_interpol,
       percent_physical_lin)[0]
   popt_cognitive_lin_quadreg = optimize.curve_fit(quadratic_func, years_interpol,
       percent_cognitive_lin)[0]
87
   # Fit cubic regression for cubic interpolation dataset
89 popt_physical_cubic_cubicreg = optimize.curve_fit(cubic_func, years_interpol,
       percent_physical_cubic)[0]
   popt_cognitive_cubic_cubicreg = optimize.curve_fit(cubic_func, years_interpol,
       percent_cognitive_cubic)[0]
91
   """Interpolation & Regression Plotting"""
93
   # Plotting Params
95 alpha = 0.3
                # transparency for scatter plot
   s = 0.5
                # size of points for scatter plot
97 figsize = (6.4, 4.8)
                # default value for figure size
   marker = 'o'
                # shape of points
99 linestyle_interpol = ''
               # line style for interpolated data
```

```
linestyle_reg = '-'
                # line style for regression curves
color_physical = 'k'
                # black coloring for physical health
color_cognitive = 'b'
                # blue coloring for cognitive health


# Raw Data Jittering
jitter = 0.1
years_physical_jittered = years_physical + jitter * np.random.randn(len(years_physical))
                # jitter physical year values
percent_physical_jittered = percent_physical + jitter * np.random.randn(len(percent_physical))
                # jitter physical percent values
years_cognitive_jittered = years_cognitive + jitter * np.random.randn(len(years_cognitive)) + 0.5
                # jitter cognitive year values and add an offset of 0.5
percent_cognitive_jittered = percent_cognitive + jitter * np.random.randn(len(percent_cognitive))
                # jitter cognitive percent values


# Raw Data Plotting
plt.figure(figsize = figsize)
plt.scatter(years_physical_jittered, percent_physical_jittered, marker = marker, c =
    color_physical, s = s, alpha = alpha, label="Poor Physical Health")
plt.scatter(years_cognitive_jittered, percent_cognitive_jittered, marker = marker, c =
    color_cognitive, s = s, alpha = alpha, label="Cognitive Decline")
plt.xlabel('Years')
plt.ylabel('Percentage')
plt.title('Raw Data of Relationship Between Poor Physical Health and Cognitive Decline in Older
    Americans')
plt.legend(markerscale = 10)
plt.savefig('raw.png', bbox_inches='tight')


# Linear Interpolation & Linear Regression Plotting
plt.figure(figsize = figsize)

plt.plot(years_interpol, percent_physical_lin,
         color=color_physical, marker=marker, linestyle=linestyle_interpol, label='Poor Physical
    Health')
plt.plot(years_interpol, linear_func(years_interpol, *popt_physical_lin_linreg),
         color=color_physical, linestyle=linestyle_reg, label='Corresponding Linear Regression')

plt.plot(years_interpol, percent_cognitive_lin,
         color=color_cognitive, marker=marker, linestyle=linestyle_interpol, label='Cognitive
    Decline')
plt.plot(years_interpol, linear_func(years_interpol, *popt_cognitive_lin_linreg),
         color=color_cognitive, linestyle=linestyle_reg, label='Corresponding Linear Regression')

plt.xlabel('Years')
plt.ylabel('Percentage of Older Americans')
plt.title('Linear Interpolation with Linear Regression Models')
plt.legend(loc='lower left')
plt.savefig('linear_interpol_linear_reg.png', bbox_inches='tight')


# Linear Interpolation & Quadratic Regression Plotting
plt.figure(figsize = figsize)

plt.plot(years_interpol, percent_physical_lin,
         color=color_physical, marker=marker, linestyle=linestyle_interpol, label='Poor Physical
    Health')
plt.plot(years_interpol, quadratic_func(years_interpol, *popt_physical_lin_quadreg),
         color=color_physical, linestyle=linestyle_reg, label='Corresponding Quadratic Regression
    ')

plt.plot(years_interpol, percent_cognitive_lin,
         color=color_cognitive, marker=marker, linestyle=linestyle_interpol, label='Cognitive
```

```
    Decline')
plt.plot(years_interpol, quadratic_func(years_interpol, *popt_cognitive_lin_quadreg),
         color=color_cognitive, linestyle=linestyle_reg, label='Corresponding Quadratic
    Regression')

plt.xlabel('Years')
plt.ylabel('Percentage of Older Americans')
plt.title('Linear Interpolation with Quadratic Regression Models')
plt.legend(loc='lower left')
plt.savefig('linear_interpol_quadratic_reg.png', bbox_inches='tight')


# Cubic Interpolation & Regression Plotting
plt.figure(figsize = figsize)

plt.plot(years_interpol, percent_physical_cubic,
         color=color_physical, marker=marker, linestyle=linestyle_interpol, label="Poor Physical
    Health")
plt.plot(years_interpol, cubic_func(years_interpol, *popt_physical_cubic_cubicreg),
         color=color_physical, linestyle=linestyle_reg, label='Corresponding Cubic Regression')

plt.plot(years_interpol, percent_cognitive_cubic,
         color=color_cognitive, marker=marker, linestyle=linestyle_interpol, label="Cognitive
    Decline")
plt.plot(years_interpol, cubic_func(years_interpol, *popt_cognitive_cubic_cubicreg),
         color=color_cognitive, linestyle=linestyle_reg, label='Corresponding Cubic Regression')

plt.xlabel('Years')
```

paper.py

Prior to beginning any statistical analysis or manipulation, the raw Alzheimer's data from the CDC was tidied in Python `pandas` for ease of access and processing. The data was converted to a long format, sorted by time in descending order, `NaN` values were removed, and percentages were filtered by cognitive decline and physical decline and stored in `numpy` arrays for manipulation.

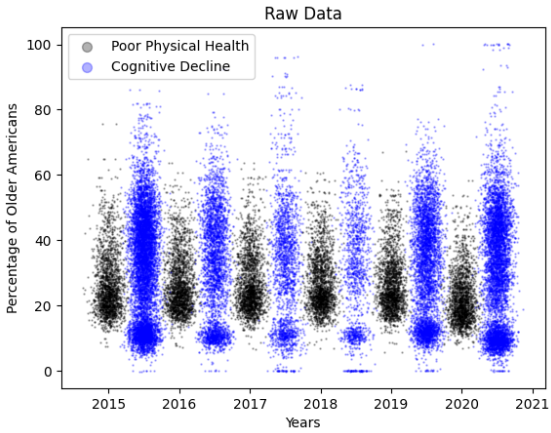
Using `scipy`'s `interp1d` and `CubicSpline` functions, the tidied Alzheimer's data was processed into respective linear interpolation and cubic interpolation functions. These functions created from the tidied data then estimated percentages of older adults experiencing either cognitive decline or poor physical health in one month intervals between 2015 and 2020. Note that to produce a cubic interpolation function with `CubicSpline` from `scipy`, the tidied CDC data set had to be filtered to remove duplicate years as this is a requirement of the cubic spline equations. This non-duplicate data was used only for this function.

Following the respective linear and cubic interpolations of the tidied data, `scipy`'s `optimize.curve_fit` function was used to fit varying degrees of regression curves to the interpolated data. The linearly interpolated data was fit with both linear and quadratic regression functions, while the cubic data had a cubic line-of-best-fit made to match the degree of interpolation used and improve accuracy.

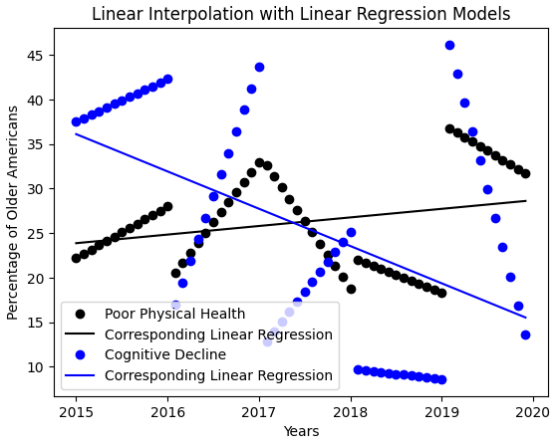
Given the small size of the domain of the (tidied) raw data set, significant overlap was present in visualization. To combat this, a scatter plot was produced with each point jittered by 10% and plotted with only 30% transparency to see the densities of each percentage. Additionally, the subset of cognitive decline data was dodged with the poor health data subset for a clearer comparison. This style of manipulation was not required for the subsequent interpolated data sets as there were not as many points present at this point in the process given the equal step size required by interpolation methods. In subsequent plots, each interpolation method was visualized against its respective regression curves (e.g. linear interpolation data vs linear regression curves). The cognitive decline subsets were labeled with blue visuals, the physical decline subsets were labeled with black, interpolated data was plotted with individual set points, and regression curves were plotted with solid lines. Each plot was produced using `matplotlib`.

Data

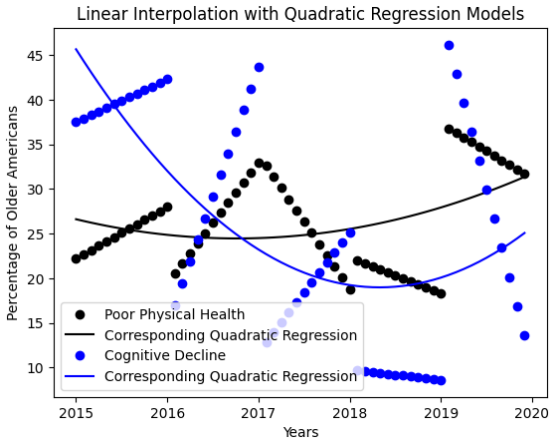
The following figures are the product of running `paper.py`:



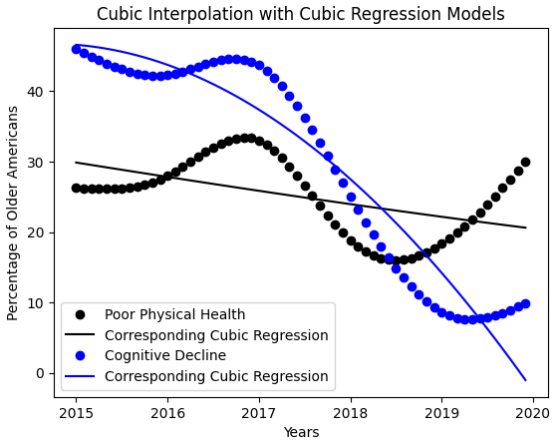
(a) Raw data of relationship between poor physical health and cognitive decline in Older Americans over a 5-year period.



(b) Linearly-interpolated data of poor physical health and cognitive decline over a 5-year period fit with linear regression.



(c) Linearly-interpolated data of poor physical health and cognitive decline over a 5-year period fit with quadratic regression.



(d) Cubicly-interpolated data of poor physical health and cognitive decline over a 5-year period fit with cubic regression.

Figure 1: Comparison of linear interpolation with linear and quadratic regression models and cubic interpolation with cubic spline regression models.

Figure 1(a) does not show any particular relationship between cognitive decline and poor physical health other than there are clearly more individuals experiencing mental decline than individuals experiencing physical decline. **Figure 1(b)** and **Figure1(c)** show implications of a relationship between each dependent variable given both the interpolated data and the regression curves. For the interpolated data, each one-year interval expect for 2017 to 2018 implies a direct relationship between physical and cognitive decline; 2017 to 2018 implies an inverse relationship. However, both the linear and quadratic regression curves imply and inverse relationship between the variables. However, given the small domain of the overall data set, the lack of agreeance is to be expected. For instance, consider the interval 2015 to 2016. Each non-boundary point is an estimate based on the end points. Therefore, each estimated point will be drawn on the same straight

line and mislead audiences with its trends.

The cubic spline interpolation, however, appears to be much more accurate. Both the cubicly interpolated data and the cubic agree a direct relationship exists between poor physical health and cognitive decline. While the regression curves do not align to a high degree of precision to the interpolated data, the curves still follow the same pattern and are relatively accurate to their respective data sets. If the regression curves were to align perfectly to the data points, this would imply over fitting has occurred and thus question whether their respective functions are misleading. Note that the regression curves in `Figure 1(b)` and `Figure 1(c)` are under fit to their respective data.

Improvements to this implementation would be to test the accuracy of higher degrees of regression for each interpolated data set and possibly extrapolate the data sets (rather than interpolate) to expand the domain. Additionally, excluding time from future exploration would allow physical health and cognitive health to be plotted on opposing axes to better estimate their correlation to one another. As both appear to follow similar trends over time, eliminating time from the equations would not have a significant effect on describing the relationship between cognitive decline and physical health.

References

Department of Health and Human Services, Alzheimer's Disease and Healthy Aging Data (2022). Centers for Disease Control and Prevention. Retrieved February 26, 2023, from https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data.

Field, B., Mountain, G., Burgess, J., Di Bona, L., Kelleher, D., Mundy, J., & Wenborn, J. (2019). Recruiting hard to reach populations to studies: Breaking the silence: An example from a study that recruited people with dementia. BMJ Open, 9(11). https://doi.org/10.1136/bmjopen-2019-030829

Miao, S.-di, Li, S.-qi, Zheng, X.-yang, Wang, R.-tao, Li, J., Ding, S.-si, & Ma, J.-feng. (2021). Missing data interpolation of alzheimer's disease based on column-by-column mixed mode. Complexity, 2021. https://doi.org/10.1155/2021/3541516