

Group Projects on Newton-Rapson Optimization.

P8160 Advanced Statistical Computing

Project 1: Analyses of daily COVID-19 cases, hospitalization, death in NYC

Background:

The COVID-19 pandemic has affected many aspects of our lives. One positive change is the push for an open data policy. It has become a new norm for governments, healthcare facilities, and academic institutes to collect data in a timely manner and share them publicly. These publicly available data at all levels empowered COVID-19 research and helped policymakers make informed decisions.

The New York Department of Health (NYDOH) published citywide and borough-specific daily counts of COVID cases, hospitalizations, and deaths since February 29, 2020, when the Health Department classified the start of the COVID-19 outbreak in NYC (i.e., the date of the first laboratory-confirmed case). A subset of its data up to December 11, 2020, was included in the attached data by day. csv, which recorded the following variables:

```
##
##
## varnames          vardefs
## -----
## DATE_OF_INTEREST   Date
## CASE_COUNT         Count of confirmed cases citywide
## HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED citywide
## DEATH_COUNT        Count of confirmed deaths citywide
## BX_CASE_COUNT      Count of confirmed cases in the Bronx
## BX_HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED in the Bronx
## BX_DEATH_COUNT     Count of confirmed deaths in the Bronx
## BK_CASE_COUNT      Count of confirmed cases in the Brooklyn
## BK_HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED in the Brooklyn
## BK_DEATH_COUNT     Count of confirmed deaths in the Brooklyn
## MN_CASE_COUNT      Count of confirmed cases in the Manhattan
## MN_HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED in the Manhattan
## MN_DEATH_COUNT     Count of confirmed deaths in the Manhattan
## QN_CASE_COUNT      Count of confirmed cases in the Queens
## QN_HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED in the Queens
## QN_DEATH_COUNT     Count of confirmed deaths in the Queens
## SI_CASE_COUNT      Count of confirmed cases in the Staten Island
## SI_HOSPITALIZED_COUNT Count of confirmed HOSPITALIZED in the Staten Island
## SI_DEATH_COUNT     Count of confirmed deaths in the Staten Island
```

Richard growth curve

Richard's growth function is a four-parameter nonlinear S -shaped function that has been commonly used in biology to model the growth of a population; Let $N(t)$ be a population at time t , Richard's growth function takes the form

$$N(t) = \frac{a}{\{1 + d \exp\{-k(t - t_0)\}\}^{1/d}},$$

where t is the time since the beginning of a population, $c(a, k, d, t_0)$ are shape parameters with specific geometric meanings. The parameter a is the upper bound of the function ($a = \lim_{t \rightarrow \infty} N(t)$, i.e., the largest population it could reach); The parameter k is the growth rate, which controls the slope at an inflection; The parameter t_0 is the time at an inflection, where the curve changes from convex to concave; and finally, the parameter d is another shape parameter. The four-parameter Richard function is a generalization of the logistic curve since it does not enforce symmetry before and after the inflection point.

Model the number of cumulative cases in a pandemic wave by the Richard growth curve

Let $(y_i, t_i)_{i=1, \dots, n}$ be a sequence of the observed daily cases at time t_i , where t_i is the number of days since the beginning of a pandemic wave. Let $Y_i = \sum_{k=1}^i y_k$ be cumulative number of cases by time t_i , we assume that (Y_i, t_i) follows the following non-linear model

$$Y_i = N(t_i, \theta) + \epsilon_i$$

where $\theta = c(a, k, d, t_0)$ is the parameters in the Richard's function, and ϵ_i is the random error with mean zero.

Task 1.1: Develop a Newton-Raphson algorithm to fit a Richard curve to each NYC borough's cumulative number of cases in the first pandemic wave. The proposed algorithm should also include a strategy for choosing starting values $\theta = c(a, k, d, t_0)$ and state the rationale for choosing starting values.

Wrong starting values result in longer iterations, greater execution time, non-convergence, or incorrect convergence. Therefore, it is important to have a strategy to select the starting values sensibly. Can you "guess" good starting values of $c(a, k, d, t_0)$, using the observed daily cases and knowledge of the shape parameters in the Richard function? Propose and state your rationale for your choice of starting values.

Task 1.2: Present your fitted curves and estimated parameters from your algorithm, and compare them across different pandemic waves and five NYC boroughs. Apply the same optimization algorithms to model the cumulative hospitalizations and deaths and compare them across different pandemic waves and the five NYC boroughs. What do we learn?

Task 1.3:

On December 11, 2020, the FDA authorized the use of vaccination. The New York Department of Health (NYDOH) needed to develop a plan on how to distribute the vaccinations (once they arrived in NYC) to the five boroughs and how fast they needed to roll out the vaccinations given that the city was in the middle of the second wave.

You are one of the consultants to the NYDOH. Can you apply your algorithm to second-wave data to predict trends after Dec 11th? Based on what you learned from your model, together with the latest census data on Borough populations (see below), what would be your suggestions for NYDOH regarding the vaccination roll-out plan?

**** NYC borough populations**

Queens – 2,405,464

Manhattan – 1,694,251

The Bronx – 1,472,654

Brooklyn – 2,736,074

Staten Island – 495,747

#Includes your R codes

Project 2: Breast Cancer Diagnosis

Background

The data *breast-cancer.csv* have 569 row and 33 columns. The first column **ID** labels individual breast tissue images; The second column **Diagnosis** identifies if the image is coming from cancer tissue or benign cases (M=malignant, B = benign). There are 357 benign and 212 malignant cases. The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of the following 10 features computed for the cellnuclei;

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The goal of the exercise is to build a predictive model based on logistic regression to facilitate cancer diagnosis;

Tasks:

1. Build a logistic model to classify the images into malignant/benign, and write down your likelihood function, its gradient and Hessian matrix.
2. Develop a Newton-Raphson algorithm to estimate your model;
3. Build a logistic-LASSO model to select features, and implement a path-wise coordinate-wise optimization algorithm to obtain a path of solutions with a sequence of descending λ 's.
4. Use 5-fold cross-validation to select the best λ . Compare the prediction performance between the 'optimal' model and 'full' model
5. Write a report to summarize your findings.