Analysis of World COVID-19 Vaccination Progress

2021

Project Description

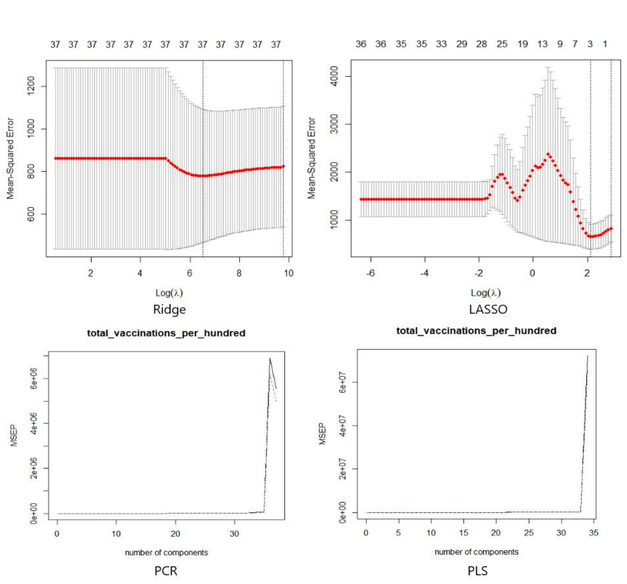
- This project is about finding the most significant factors that influenced the number of population vaccinated against the COVID-19 in various countries around the world. And demonstrate the vaccination progress by data visualization methods.
- Model used in the project: ridge, LASSO, PCR, PLS and best subset selection.
- Goals of this project:
  - [1]Tracking the progress of vaccination around the world by visualizations.
  - [2]Identify the key factors affecting the vaccination process.
- Use short sentences.

Data Source

- All datasets came from *Kaggle.com.*

Data Description

- Two datasets were used in the project:
  - [1]country_vaccinations
  - [2]country_profile_variables



MSEP plot for Ridge, LASSO, PCR, and PLS method, plotted with the software R.

Team Members:

Linpeng Sun
Priya Venkadesh

Methodology

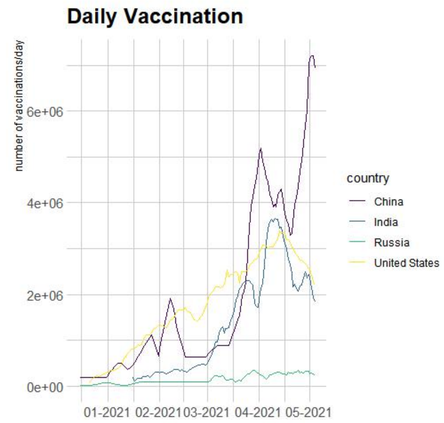
1. Clean and re-organized data to appropriate format for the usage of our model.
2. Fit the new dataset to different models with the methods we chose.
3. Draw MSEP plot for Ridge, LASSO, PCR, and PLS method and RSS, adjrsq, Cp, BIC vs. number of variables for best subset selection method.
4. Selected the best model of each methods, and by comparing the test MSE of those models to determine the best model for our dataset.
5. Use ggplot2 to draw the daily vaccination plot and total vaccination per hundred plot for countries with the largest number.

Results and Conclusions

1. After comparing the test MSE of all models, the principal components regression has the smallest test MSE value. Which is 320.5459 when 25 components are used.
2. The principal components regression is good for deriving a low-dimensional set of features from a large set of variables.
3. Gibraltar, Seychelles, Israel are the top-3 countries with the highest vaccination per hundred people.
4. China, United States, and India are the top-3 countries with the highest number of total vaccination of 270406000, 243463471, and 153626325 by 2021-05-01.

Implementation (Tuning, R Functions, Algorithm)

LASSO: Least absolute shrinkage and selection operator.
PCR: Principal components regression.
PLS: Partial Least Squares.



Daily vaccinations in US, China, India, and Russia, plotted with the ggplot2 package in R.

Country	Gibraltar	Seychelles	Israel	United Arab Emirates	Cayman Islands
Region	SouthernEurope	EasternAfrica	WesternAsia	WesternAsia	Caribbean
Total Vaccination/100	211.08	129.88	121.03	108.99	99.96

Countries with highest total vaccination per hundred.

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R* (Springer Texts in Statistics) (1st ed. 2013, Corr. 7th printing 2017 ed.). Springer.
2. DATA *country_profile_variables* is an internet based data service launched by the United Nations Statistics Division (UNSD) of the Department of Economic and Social Affairs (DESA).
3. country_vaccinations data is collected from Our World in Data GitHub repository for covid-19.