

Yimeng Shang

+1 (646) 704-5390

ys3298@cumc.columbia.edu

<https://ys3298.github.io/>

EDUCATION

Columbia University, Mailman School of Public Health

- Master of Science in Biostatistics (GPA: 4.0/ 4.0)

New York City, NY, USA

08/2019-Expected in 05/2021

East China Normal University

- Bachelor of Science in Mathematics and Applied Mathematics

Shanghai, China

09/2015-06/2019

University of California, Berkeley

- Berkeley International Study Program

Berkeley, CA, USA

08/2017-05/2018

SKILLS & RELEVANT COURSES

- **Relevant Courses:** Biostatistical Methods, Statistical Inference, Statistical Computing, Data Science, Machine Learning, Epidemiology
- **Computer Skills:** R, SAS, Python, MATLAB, Microsoft Office Software.

INTERSHIPS

Eli Lilly China

Shanghai, China, 09/2018– 06/2019

- *Data Science & Solution Intern*

- Supported data management work in clinical trials including data cleaning and missing data query under supervision of China DSS team and participated in the discussion of statistical methods applied for process design of clinical trials.
- Constructed quantitative analysis of possible interference risks during clinical trials and organization operation, developed an automatic web page for reproducible monthly analysis to improve efficiency

RESEARCH/ PROJECTS EXPERIENCE

Feature selection and Prediction using EM algorithm base on Multiple label-noise response

05/2020-08/2020

- *Research assistant, supervised by Professor Hua Shen*

- Combined EM algorithm and Adaptive LASSO to estimate sensitivity and specificity of each surrogate, did feature selection and prediction based on this algorithm
- Constructed multiple simulation studies in different settings, including low dimension and high dimension under different assumptions
- Achieved improvement in AUC, ACC, TPR, TNR, PPV, NPV compared to Naive, Ad Hoc method
- An academic report is under draft

Metabolomics statistical analysis on Autism data

05/2020-08/2020

- *Research assistant, supervised by Professor Xiaoyu Che*

- Constructed both logistic regression and Cox hazard model to estimate the effect size and did power analysis to select the final model and tested interaction term
- Used Adaptive LASSO, Random Forest and XGBoosting to select features from high dimensional dataset and fit a final model with selected feature to do prediction
- Constructed conditional logistic regression, feature selection and prediction base on one to one matched dataset

COVID-19 Analysis via Logistic Curve and Clustering

04/2020

- *Course: The Advanced Topics in Statistical Computing*

- Estimated parameters in logistic curve using the combination of Ordinary Differential Equation method and Gradient Descent model. First used ODE to get the initial guess and plug into gradient descent to estimate the parameters (No package used).
- Implemented both Gaussian Mixture Model with EM algorithm and Kmeans algorithm to cluster the estimated parameters for each countries (No package used)
- Used the clustering results drawing world to visualize the distribution of clusters to see the geometric association

Implementation and optimization of algorithms on breast cancer diagnosis dataset

03/2020

- *Course: The Advanced Topics in Statistical Computing*

- Built a predictive model based on logistic regression to facilitate cancer diagnosis
- Implemented and trained logistic regression models with Newton Raphson, Gradient Decent algorithms and a logistic-LASSO regression model with Pathwise Coordinate Descent from scratch in R
- The misclassification rate of logistic-LASSO model is the lowest by comparing models' performance by Cross Validation in R

The Suicide Rate Project

11/2019-12/2019

- *Course: Data Science I*

- Used *R programming* to produce time-based graphics indicating number of suicide cases related to risk factors based on the gathered data from 13 different sources and realized multiple linear regression to find the main effect using *LASSO* and *BIC* criteria.
- Made a Crude Suicide Rate map allowing users to toggle between different years and choose gender, demonstrating the changes in the suicide rate over time and visualize regional differences across the U.S.

