Breast cancer outcome prediction using screening-based and prediction-based methods via the BCSC dataset

Biost 544 proposal presentation
Date: 12/09/2020
Team members: Jiawen Liu and Ziyuan Wang

Breast Cancer - why do we care?

- Breast cancer death is the second-common cause of cancer-associated death among women of all races and ethnicity.
- Successful cancer detection is an important aspect of cancer treatment, especially at an early stage.
- Various detection and prediction methods used
    - Mammography
    - Prediction algorithms

Data description

The Breast Cancer Surveillance Consortium (BCSC) collects data on breast cancer outcomes, potential cancer risk factors, and breast mammography performance.

We have ~15,000 observations with 10 potential covariates after removing missing data, the outcome of interest in cancer indicator.

Baseline age at time of mammogram, Radiologist's assessment, Availability of prior mammogram for comparison, Breast density, Family history of breast cancer, Current use of hormone therapy, Previous mammogram, Mammogram type, Previous biopsy and BMI

The scientific question of interest

In this project, we are interested in building building a **prediction model** using screen-based methods, best subset selection and prediction-based methods (Ridge and Lasso), to best predict breast cancer outcomes using the BCSC dataset.

Challenge

- Determine the measurement of association between continuous covariates and binary outcome
  - Previously we used
    - mean difference between binary covariate and continuous outcome
    - Measure of association like R2, kendall correlation between continuous covariates and continuous outcome

Method - screen-based

- Use **permutation** to measure association between potential predictors and outcome
  - We use logistics regression as we have binary outcome
  - Use McFadden’s R2 to evaluate regression
  - Use likelihood ratio test to compare model with covariates and null model

Method - Best subset selection

- Use **best subset** method and 5-fold **cross validation (CV)** to determine predictors included in regression model
  - 10 potential covariates and a total of 2^10 = 2048 possible subsets
  - First use regsubsets() function in R to identify the best model that contains 1:10 predictors
  - ~15,000 observations so 5-fold CV still have enough observations in each training/test set

Method - penalized models

General form: $avgL(outcome, \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k) + \lambda P(\beta_1, \ldots, \beta_k)$

Ridge regression

- Penalized term: $\lambda(\beta_1^2 + \cdots + \beta_k^2)$

Lasso

- Penalized term: $\lambda(|\beta_1| + \cdots + |\beta_k|)$
- Result in a sparse model and more computational friendly than best subset selection (especially for dataset with large number of covariates)

Key findings

- **Screen based**
  - **Permutation using McFadden's R2**: availability of prior mammogram for comparison, current use of hormone treatment, Prior mammogram, Prior biopsy, mammogram types and BMI
  - **Permutation using likelihood ratio test**: breast density, and prior biopsy
- **Best subset selection**: The final best model includes 4 features: radiologists' assessments, availability of prior mammogram for comparison, breast density and mammography type
- **Ridge**: When including all 10 covariates, the error is the smallest when lambda = 0
- **Lasso**: The optimal model we found contains five features: radiologists' assessments, breast density, family history of breast cancer, and current use of hormone treatment

Questions

Thank you