PH 240C Final Project

Instructor: Jingshen Wang

November 26, 2020

1 If you work on an assigned dataset:

1.1 Overview of the new dataset

In the final presentation, you are going to keep working with the UK Biobank data. The new dataset contains 2,000 individuals with the same biomarkers and lifestyle variables as the ones provided in the final presentation. The only difference is that half of the new sample contains incidence cases for CVD. We have again perturbed the original UK Biobank dataset to avoid privacy issues.

1.2 Data description

The dataset is included in the data file as *final-project-data.RData*. The variables include baseline biomarkers, lifestyle variables and living conditions. Please check UK Biobank data encoding for more details.

1. Biomarkers
    - *cad*: cardiovascular disease status (0: No; 1: Yes) (Use this variable as the label)
    - *PCA1, PCA2, PCA3*: the first three principle components of family ancestry
    - *ches pain discomfort, high blood pressure, high cholesterol, diabetes, chip*: (0: No; 1: Yes)
    - *age, gender*: (1:Male, 0: Female)
    - *Diastolic blood pressure, systolic blood pressure, pulse rate*: continuous variables
2. Lifestyle variables
    - *alcohol intake frequency*: 1: daily; 2: three or four times per week; 3: once or twice per week; 4: one to three times per month; 5: special occasions; 6: never.
    - *oily fish intake*: 0: never; 1: less than once per week; 2: once a week; 3: two to four times per week; 4: five to six times per week; 5: once or more daily.
    - *number of days/week walked 10 minutes, time spent watching television, sleep duration, sleep var, cooked veggie intake, break intake, tea intake*: continuous variables
3. Living condition variables
    - Close to major road (0:No; 1: Yes)

- *location, townsend deprivation index, inverse distance to the nearest major road, inverse distance to the nearest road, nitrogen dioxide air pollution 2005-2010, particulate matter air pollution, sum of road length, total traffic load, traffic intensity, average daytime sound level, average nighttime sound level*: continuous variables
- *number of vehicles in household*: 1:None; 2: One; 3: Two; 4: Three; 5: Four or more

1.3 Tasks

Please work on the following problems for your final presentation:

1. Introduce the overview of your new dataset. For example, number of observations, dimension of covariates, and covariance structure, etc.
2. Based on the new dataset, use the same methods you have used in the midterm to classify the cardiovascular disease status. What are the prediction accuracy based on different methods? Which method yields highest prediction accuracy? Please discuss your results.
3. Because we hope to further improve prediction accuracy, please introduce interaction terms into your model building process. How do your prediction results change? Please discuss these changes. Why or why not theses changes make sense to you?
4. If we could recommend each individual to change their lifestyles, for example (1) reducing alcohol intake or (2) increasing oily fish intake, what are the effect sizes of these lifestyle changes on lowering the CVD risk? Please discuss your results and methods.
5. Please answer the open-ended question you raised in the mid term presentation.

2 If you work on your own dataset:

Please work on the following for your final presentation:

1. Introduce the overview of your dataset. For example, number of observations, dimension of covariates, and covariance structure, etc.
2. Use the methods we have talked about in the second half of the term to analyze your dataset. How do your results differ from your mid-term results?
3. Please introduce interaction terms into your model building process. How do your prediction results change? Please discuss these changes. Why or why not theses changes make sense to you?
4. In the model with interaction terms, how predictive are your covariates? Can you quantify how predictive they are? If so, please give your quantification methods; if not, please explain why.
5. Please answer the question(s) the instructor raised in your midterm presentation.

3 Final presentation format

- Class on December 1st is office hour. Feel free to drop by if you have any questions and concerns about the final presentation through the usual class zoom link;
- We will have 8 final presentations: four on December 2nd from 4 pm and another four on December 3rd from 5 pm: link to schedule;
- Each group has 15 minutes, plus 2-3 minutes including Q&A;
- Each group cannot use more than 15 slides (excluding the title page and the thank you page), so you have approximately 1 minute for each slides;
- Your presentation will be graded based on (100 points in total):
    1. Use of allotted time. Does the talk run over or under? Was time used efficiently? (10/100)
    2. Logical progression. Is the talk easy to follow? (10/100)
    3. Completeness. Has the talk answered all the questions raised above (30/100);
    4. Clarity. Have the audience understood the talk? Has the presentation answered all the questions clearly? (20/100)
    5. Knowledge of the adopted method. Does the presenter seem to understand the statistical methodology used to analyze the dataset? (20/100)
    6. Creativity. Do you find the proposed open-ended questions interesting? (10/100)
- Each student needs to evaluate your peer's presentation based on the above rubric and send me the score for each presentation. The final grade of the final project will take these scores into account.