Final Project Proposal

Ruofan Chen 2862919

1 Background and Purpose

The CDC provides the public dataset Mortality Multiple Cause Files, which contains the record for mortality events and their corresponding information, including the demographical description, category of death, the diagnosis of the latent causes and other follow-up. Hence, an interesting point will be suicide risk screening based on the demographical and other info. On the first stage, a descriptive analysis will be carried out, the boxplot, distribution curve plot and other useful visualization will be used. To realize the analysis, a weighted logistic regression for the imbalanced data set classification will be established with other covariates (e.g. education status, gender, age, marital and etc.) With the established model, we can interpret the effect of each covariate to the odds ratio of the suicide risk. To have a more insightful conclusion with simplest form, the sparsity is encouraged by adding a lasso penalty. To verify the conclusion and validation of the model, other model selection tool such as Mallows's C_p, VIF and the residual analysis will be adopted. Besides, if time is sufficient, I will also involve the external data from Kansas Health Information Network (KHIN) after the data desensitization. T. Tony Cai et.al (2020) proposed a scheme for linear transfer learning that utilize the source dataset to improve the performance of estimator on the target dataset. In this project, we explore the way to extend the proposed scheme to the case of general linear model.

2 Data Source and Description

- Data source: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortality_Multiple
- Data description: https://www.cdc.gov/nchs/data/dvs/Multiple-Cause-Record-Layout-2019-508.pdf

3 Time Line and Statistical Method

- Data extraction and sorting (before March 25th)
- Descriptive analysis (before April 1st): Boxplot, distribution curve plot and scatter matrix plot for correlation analysis.
- Weighted logistic regression with L1 penalty (before April 10th): For the weighted logistic regression with L1 penalty, the predicting model is:

  $$P\left(y\,|x\right)=\frac{\exp\left(x^T\beta y\right)}{1+\exp\left(x^T\beta y\right)},$$

  where x is the covariate vector (after precessing such as dummy transformation or the normalization). The optimization task should be:

  $$\hat{\beta}\in\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i}\log\left(1+\exp\left(-{x_i}^T\beta y_i\right)\right)+\lambda\|\beta\|_1.$$

- Hypothesis test by deviance (before April 15th): To determine whether β_j is not 0, the statistics could be

  $$G^2=D\left(R\right)-D\left(F\right)\sim\chi^2\left(1\right),$$

  where $D\left(R\right)$ is the deviance for reduced model with $\beta_j=0$.

- Transfer learning method I'd like to play with (before April 30th): Step 1, train

  $$\hat{w}\in\arg\min_{w}\frac{1}{|I_S\cup I_T|}\sum_{i\in I_S\cup I_T}\log\left(1+\exp\left(-{x_i}^T w y_i\right)\right)+\lambda\|w\|_1,$$

  where I_S and I_T are sample index for source dataset and target dataset. Then step 2 will be the estimation of the "parameter gap"

  $$\hat{\delta}\in\arg\min_{\delta}\frac{1}{|I_S|}\sum_{i\in I_S}\log\left(1+\exp\left(-{x_i}^T\left(\hat{w}+\delta\right)y_i\right)\right)+\lambda\|\delta\|_1.$$

  Finally, output the improved estimator of β as $\hat{\beta}=\hat{w}+\hat{\delta}$.