US Suicide Risk Screening

Ruofan Chen

2021-05-02

Contents

Abstract

In 2019, suicide is the 10th leading cause of death in the US. This project uses Mortality Multiple Cause Files by CDC, conducting research in suicide risk screening based on 2019 demographic data related to the death. In the first stage, a descriptive analysis is carried out, distribution curve plot and tables are generated. To realize the analysis, logistic regression is established with other covariates (education status, gender, month, age, marital status, and race). After clarifying that the distribution of the response (manner of death is suicide) is unbalanced, two methods are utilized in the model, including down-sampling technique and weighted loss function. To have a more insightful conclusion with the simplest form, the sparsity is encouraged by adding a lasso penalty. To find the optimal penalized parameters, a 5-fold cross-validation grid search is performed on specific intervals. Finally, both down-sampled and weighted logistic regression models can attain the area under the ROC curve (AUC) greater than 0.87 and the corresponding confidence intervals of parameters are obtained via the Fisher information approach. Due to the large sample size, down-sampling is appropriate when considering computer calculation time, with an AUC of 0.870063. In the second stage, a framework for transfer learning of logistic regression is built up and is verified to be useful if the "informative" external data is accessible.

Key Words: GLM, logistic regression, regularization, down-sampling method, weighted loss, area under ROC curve (AUC), transfer learning

Section 1. Introduction

Suicide is defined as death caused by self-directed injurious behavior with intent to die as a result of the behavior. According to National Vital Statistics System - Mortality Data (2019) via CDC WONDER, suicide in 2019 was the tenth leading cause of death in the United States. More notably, it is estimated that there were 1.38 million suicide attempts in 2019. Studies have found that people who have attempted suicide in the past have a higher risk of suicide in the future.

The project uses the public data set 'Mortality Multiple Cause Files' provided by the CDC, which contains records of mortality events and their corresponding information, to screen for suicide risk based on demographic information and other information.

After clarifying the unbalanced distribution of the response variables, two methods were used in the logistic regression model with regularization, including the down-sampling technique and weighted loss function. The data preprocessing process includes selecting variables related to the response (suicide occurs), deleting observations with missing or undeclared values, and converting some categorical variables.

In the following sections of the report, the characteristics of the data set, the development process of the model, and the related final results will be discussed. Section two contains data descriptions and generates related tables and graphs. The third section 'Models and Methods' shows down-sampling method, logistic regression model, significance test for the estimation, K-fold Cross-Validation, and transfer learning. Section four explains the influence of covariates on the odds ratio of the suicide risk and discusses its performance, and validation for transfer learning. Section four and five illustrate the final results and release some comments on the model. The appendix contains definitions of variables, graphs, and model supplements related to the models.

Section 2. Data Description

The data set is the 2019 data extracted from the CDC's 'Mortality Multiple Cause Files', please refer to https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortality_Multiple. There are a total of 2,861,523 observations, and 10 death-related information is selected as relevant variables, including Education (2003 version), Month, Sex, Age Detailed, Marital Status, Day of Week, Manner of Death, Place of Injury, ICD Code (Version 10), and Bridged Race Recode 5. For this study, the variable 'Manner of Death' is set

as the dependent variable. Before deciding which variables to set as predictors, the data must be examined and described.

To prepare for analysis, delete observations, including unstated values, unspecified values, unknown values, or blank values that are not applicable to variables, but keep the Place of Injury and Day of Week as they are prepared for descriptive analysis. Second, adjust the data type of the variable according to its definition, and set all variables except 'Age Detailed' as factors. Then calculate the age in years. Another transformation of the dependent variable is to divide the response into two categories: suicide and non-suicide, labeled 'Y' and 'N' respectively. After applying these preprocessing, there are 2,517,393 observations.

The descriptive analysis is based on suicide cases, including all the variables that have been read. Observations with missing or unspecified values will be eliminated.

Suicide most often occurs in the 55-60 age group, and suicides are mainly concentrated in the 20-65 age group. The number of suicides from 0 to 20 years old rises rapidly, then begins to fluctuate and peaks at 55-60 years old, and then the number of suicides decreases with age. The total number of deaths shows a clear left-skewed distribution. The frequency of deaths increases with age from 0 to 90 years old, reaching a peak in the 85-90 year-old age group, and the number of deaths decreases after 90 years of age. Figure 1 shows the age distribution.

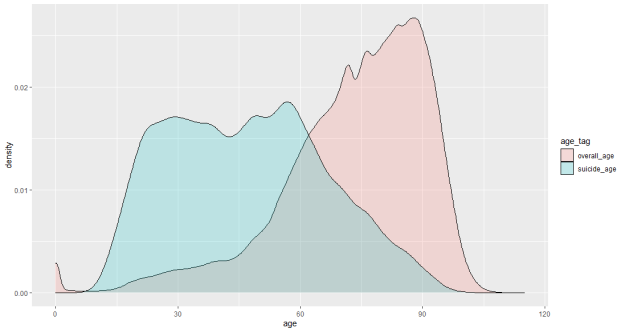

Figure 1: Age Distribution

Descriptive analysis is carried out from two aspects: comparing suicide cases with all causes of death, and the distribution of variables in suicide cases. After checking all relevant variables, most suicides occurred at home, accounting for 73.19% of all suicides, and deaths of all causes were consistent with this. Most suicide cases are high school graduates or who have completed GED, or obtained some college credits but no degree. They often occur in July, August, and September, which is different from the overall deaths that occur in December, January, and March. The proportion of male suicides is much higher than that of females, accounting for 78.5%, and the proportion of single persons is slightly higher than that of married persons. The suicide death toll of the days of the week is similar and does not seem to have any relationship with the overall death toll. The top three ICD codes are X74, X70, X72, which respectively represent intentional self-harm by other and unspecified firearm and gun discharge, intentional self-harm by hanging, strangulation and suffocation, intentional self-harm by handgun discharge respectively. White people account for the largest proportion, which is similar to the overall death proportion. Tables 1 to 3 are summary tables of age, frequency table of marital status, and frequency table of sex respectively. For other related tables please refer to Appendix A.2 Data Description Supplement.

Table 1: Summary Table of Age

Age	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Suicide	8.0	31.0	47.0	47.1	61.0	103.0
Overall	0.0	64.0	76.0	73.2	87.0	115.0

Table 2: Frequency Table of Marital Status

Marital Status	Value	D	M	S	W
Suicide	Number	9331	14471	17189	2629
Overall	Number	425329	926520	351488	814056

Table 3: Frequency Table of Sex

Sex	Value	M	F
Suicide	Number	34238	9382
Overall	Number	1299580	1217813

After checking all the variables, Education (2003 version), Month, Sex, Age Detailed, Marital Status, Day of Week, and Bridged Race Recode 5 are selected as predictors in the next section–modeling part.

Table 4 shows the first six rows of the data set used for modeling. For the definition of variable values, please refer to Appendix A.1.

Table 4: First Six Observations

edu	mon	sex	age	mar	race	manner
4	1	M	36	M	1	N
4	1	F	63	M	1	N
3	1	F	97	W	1	N
3	1	M	76	M	1	N
4	1	M	64	M	1	N
8	1	M	74	M	1	N

Section 3. Models and Methods

3.1 Downsampling Method

Down-Sampling method will randomly sample a data set so that the frequency of the majority class is the same as the frequency of the minority class. Due to the large sample size, down-sampling is appropriate when considering computer calculation time.

3.2 Logistic Regression with Regularization

Formula 1 and 2 are the logistic regression model. Here, 'Manner of Death' is set as the response variable, which is labeled Y, and follows the Bernoulli distribution with parameter p. In addition, each observation is independent. The notation p is a binomial parameter representing the probability of occurrence of the Suicide(Manner of Death equals to 'Y').

The hypothesis is probability p follows a logistic distribution. β_0 represents the intercept, β_j (j from 1 to 6) represents the partial coefficient of increasing 1 unit on X_j while holding all other predictors fixed, the change of log odds.

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

$$logit(p_i)=log(\frac{p_i}{1-p_i})=x_i'\beta=\beta_0+\beta_1X_{edu}+\beta_2X_{mon}+\beta_3X_{sex}+\beta_4X_{age}+\beta_5X_{mar}+\beta_6X_{race} \tag{2}$$

The objective function for the weighted logistic regression with L1 penalty lasso uses the negative binomial log-likelihood and is shown in Formula 3. The tuning parameter λ controls the overall strength of the penalty.

$$\hat{\beta}\in\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i}\log\left(1+\exp\left(-{x_i}^{T}\beta y_i\right)\right)+\lambda\|\beta\|_1 \tag{3}$$

3.3 Significance Test for the Estimation (Confidence Interval)

Taking the second-order derivative of minus of loss function (i.e. the likelihood function), Hessian matrix can be obtained:

$$H\left(\tilde{\beta}\right)_{jk}=\frac{\partial l\left(\tilde{\beta}\right)}{\partial\tilde{\beta}_j\tilde{\beta}_k}=-\sum_{i=1}^{n}\mu_i\left(1-\mu_i\right)x_{ij}x_{ik}$$

where $\mu_i=\exp\left(\tilde{x}_i^T\tilde{\beta}\right)/\left(1+\exp\left(\tilde{x}_i^T\tilde{\beta}\right)\right)$, $\tilde{\beta}^T=\left(\beta_0,\beta^T\right)$ and $\tilde{x}_i^T=\left(1,x_i^T\right)$. Then, the estimation of Hessian matrix can be obtained at the end of the optimization of loss by

$$\begin{aligned}\hat{H}\left(\tilde{\beta}\right)_{jk}&=-\sum_{i=1}^{n}\hat{\mu}_i\left(1-\hat{\mu}_i\right)x_{ij}x_{ik}\\ \hat{\mu}_i&=\exp\left(\tilde{x}_i^T\hat{\beta}\right)/\left(1+\exp\left(\tilde{x}_i^T\hat{\beta}\right)\right).\end{aligned}$$

Moreover, according to the relation of Hessian matrix with the Fisher information, the estimation of Fisher information will be

$$\hat{I}\left(\tilde{\beta}\right)_{jk}=-\hat{H}\left(\tilde{\beta}\right)_{ik}=\sum_{i=1}^{n}\hat{\mu}_i\left(1-\hat{\mu}_i\right)x_{ij}x_{ik}$$

With this observation, the estimation of standard error for $i-$th parameter $\tilde{\beta}_i$ will be

$$s.e.\left(\tilde{\beta}_i\right)=\sqrt{\left(\hat{I}\left(\tilde{\beta}\right)^{-1}\right)_{i,i}}.$$

Finally, the 95 percent confidence interval of estimation of $\tilde{\beta}_i$ will be carried out according to the large sample normal setting as

$$\hat{\beta}_i\pm1.96s.e.\left(\tilde{\beta}_i\right).$$

3.4 K-fold Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. The k-fold cross-validation method evaluates the model performance on a different subset of the training data and then calculates the average prediction score. Here, this method is used to find the optimized penalized parameters by picking the penalized parameters that have the highest AUC performance.

3.5 Transfer Learning

There are more and more people who agree with one statement: transfer learning (TL) is the next frontier of machine learning. The reason why transfer learning is so useful is based on a fact: more and more companies or governments realize the value of data, hence, they tend to create their private database. With the collaborations between them, a problem appears: how to take advantage of external data to improve the performance of local forecasting, whatever the regression or classification. Manifestly, stacking all the data together makes no sense, and can even generate ridiculous results as the distinction between the study cohorts for different companies. Therefore, a delicate design is needed for every different mission. For example, speech recognition , robot training , brain image diagnosis and so many industrial fields can benefit from TL.

Classification is a very common mission in statistical projects, and logistic regression is one of the most prevalent methods among all the classification tools because it can give the prediction of probability. Formally, a logistic regression suggests a target model as

$$P\left(y_i^{(0)}=1\right)=\frac{\exp\left(x_i^{(0)T}\beta\right)}{1+\exp\left(x_i^{(0)T}\beta\right)}, i=1,2,\ldots,n_0$$

where $\left\{\left(x_i^{(0)},y_i^{(0)}\right)\right\}$ are i.i.d samples and β is the true parameter to be estimated. Meanwhile, assuming that other K auxiliary logistic models are also in our interest, they can be described as

$$P\left(y_i^{(k)}=1\right)=\frac{\exp\left(x_i^{(k)T}w^{(k)}\right)}{1+\exp\left(x_i^{(k)T}w^{(k)}\right)}, i=1,2,\ldots,n_k; k=1,2,\ldots K$$

where $w^{(k)}$ plays the same rules as β in the target model. However, just as mentioned earlier, as the difference of study cohorts, every $w^{(k)}$ is assumed to be distinct from β. Therefore, this relation can be described by $\delta^{(k)}=\beta-w^{(k)}$. Besides, K is the total number of auxiliary models and data set. With this decomposition of parameters, the "informative" data sets can be defined. By defining

$$\mathcal{A}(h)=\left\{1\le k\le K:\left\|\delta^{(k)}\right\|_1\le h\right\},$$

the purpose of this session is to use $\left\{k\in A(h)\cup\{0\}:\left(x_i^{(k)},y_i^{(k)}\right)\right\}$ to give a better estimation of β than only using the sample in the target.

3.5.1 Trans-logistic Regression Algorithm

Before carrying out the algorithm, the following lemma gives an alternative loss of logistic regression with the one in session 3.2.

Lemma (Alternative Loss)

$$\sum_{i=1}^{n}\left[y_i\left(\beta_0+x_i^T\beta\right)-\log\left(1+\exp\left(\beta_0+x_i^T\beta\right)\right)\right]=\sum_{i=1}^{n}\log\left(1+\exp\left(-y_i\left(\beta_0+x_i^T\beta\right)\right)\right)$$

Proof

$$\begin{aligned}
&f(z)=\tfrac{e^z}{1+e^z}\\
&\Rightarrow f(-z)=\tfrac{e^{-z}}{1+e^{-z}}=\tfrac{1}{e^z+1}=1-f(z)\\
&\Rightarrow \log\left(1+e^z\right)=\log\left(1+e^{-z}\right)+z\\
&\Rightarrow -z+\log\left(1+e^z\right)=\log\left(1+e^{-z}\right)
\end{aligned}$$

Hence, for the ordinary logistic regression, the following expression can be written:

$$\ell\left(\{\tilde{x}_i, y_i\}_1^{n_0}; \tilde{\beta}\right) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log\left(1 + \exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)\right) + \lambda \|\beta\|_1$$
$$\hat{\beta} \in \arg\min_{\tilde{\beta}} \ell\left(\{\tilde{x}_i, y_i\}_1^{n_0}; \tilde{\beta}\right)$$

where $\tilde{x}_i = \left(1, x_i^T\right)^T$ and $\tilde{\beta} = \left(\beta_0, \beta^T\right)^T$. For the limit of $\hat{\beta}$, i.e. the β is the population-level minimizer, that is to say,

$$\beta \in \arg\min_{\tilde{\beta}} E\left[\log\left(1 + \exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)\right)\right]$$

Also, the appropriate population-level score function and Hessian matrix are set as:

$$S(\beta) = E\left[-\frac{\exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)}{1+\exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)} y_i \tilde{x}_i\right]$$
$$H(\beta) = E\left[\frac{\exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)}{\left(1+\exp\left(-y_i \tilde{x}_i^T \tilde{\beta}\right)\right)^2} \tilde{x}_i \tilde{x}_i^T\right]$$

Now, suppose that except for the primary data set $\left\{\tilde{x}_i^{(0)}, y_i^{(0)}\right\}_1^{n_0}$ there exists another auxiliary data set $\left\{\tilde{x}_i^{(1)}, y_i^{(1)}\right\}_1^{n_1}$ such that $1 \in \mathcal{A}(h)$ for a given small h, the Trans-logistic regression suggests the following optimization procedure:

- Input: Primary $\left\{\tilde{x}_i^{(0)}, y_i^{(0)}\right\}_1^{n_0}$ and auxiliary $\left\{\tilde{x}_i^{(1)}, y_i^{(1)}\right\}_1^{n_1}$.
- Result: $\hat{\beta}_{oracle}$.
- Step 1: With $\lambda_w = c_1(n_1 + n_0)^{-1/2}$, compute

$$\hat{w}^A \in \arg\min_{\tilde{w} \in R^{p+1}} \frac{1}{(n_1 + n_0)} \sum_{k \in \{0,1\}} \sum_{i \in n_k} \log\left(1 + \exp\left(-y_i^{(k)} \tilde{x}_i^{(k)T} \tilde{w}\right)\right) + \lambda_w \|w\|_1.$$

- Step 2: With $\lambda_\delta = c_2(n_0)^{-1/2}$, compute

$$\hat{\delta}^A \in \arg\min_{\tilde{\delta} \in R^{p+1}} \frac{1}{n_0} \sum_{i \in n_0} \log\left(1 + \exp\left(-y_i^{(0)} \tilde{x}_i^{(0)T} \left(\hat{w}^A + \tilde{\delta}\right)\right)\right) + \lambda_\delta \|\delta\|_1.$$

- Output: $\hat{\beta}_{oracle} = \hat{w}^A + \hat{\delta}^A$.

3.5.2 The Statistical Property of Trans-logistic Regression

Condition 1: $H(\beta) = H\left(w^{(1)}\right)$.

Condition 2:

For all data sets, there exists a unique nonzero minimizer $\tilde{\beta}^*$ such that $S\left(\tilde{\beta}^*\right) = 0$ and $c \leq \lambda_{\min}\left(H\left(\tilde{\beta}^*\right)\right) \leq \lambda_{\max}\left(H\left(\tilde{\beta}^*\right)\right) \leq c^{-1}$ for some constants $c > 0$.

Condition 3:

Unique minimizer $\tilde{\beta}^*$ for all data sets satisfies $\left\|\tilde{\beta}^*\right\|_2 \leq C$ for some constants $C > 0$.

The following theorem can be implemented, the proof of which is in the Appendix.B.

Theorem (Convergence rate for Trans-logistic regression)

Let s be the number of support of β. Assume that Condition 1,2 and 3 hold true. If the following condition for h holds true: $s\log p/\left(n_1+n_0\right)+h(\log p/n_0)^{1/2}=o\left((\log p/n_0)^{1/4}\right)$, then

$$\begin{aligned}&\sup_{\beta}\max\left(\frac{1}{n_0}\left(\hat{\beta}_{oracle}-\beta\right)^T H\left(\beta\right)\left(\hat{\beta}_{oracle}-\beta\right),\left\|\hat{\beta}_{oracle}-\beta\right\|_2^2\right)\\&=O_p\left(\frac{s\log p}{n_1+n_0}+\min\left(\frac{s\log p}{n_0},h\sqrt{\frac{\log p}{n_0}},h^2\right)\right)\end{aligned}$$

The proof of this theorem is in the Appendix.B.

Section 4. Method Implementation and Model Analysis

4.1 Logistic Regression with Downsampling Method

After down-sampling the data set, a new data set with the same number of two different levels are generated. The down-sampled data set contains 93,484 observations, of which the two types of response variables each account for 46,742. Then L1 norm regularization with 5-fold cross-validation and grid search is carried out to find the optimized penalty parameter lambda is 0.001003872. The data used for regression is split for training and testing according to the ratio of 8:2. Logistic regression with lasso regularization is performed on the training set.

The coefficients of the model are in column 'down_coef' of Table 5. By keeping all other predictors at the same level, when age increases by 1, the odds of suicide death occurrence increases by $exp(-0.074819) = 0.9279114$ times. 'd_lower' and 'd_upper' represent the lower bond and upper bond of the coefficient at the significance of 5%. If the 95% confidence interval includes 0, it can be concluded that this (level of) variable is not significant at 5% level. The column 'd_Sig' indicated the significance, 'TRUE' and 'FALSE' represent this (level of) variable is or is not significant at 5% level respectively. It can be concluded that 'edu2', variable month, 'race3' and 'race4' are not significant at 5% level. Area under ROC curve (AUC) is a good way to validate the model performance. After using the obtained coefficients to make predictions on the test set, an AUC of 0.870063 is obtained which is shown in Figure 2.

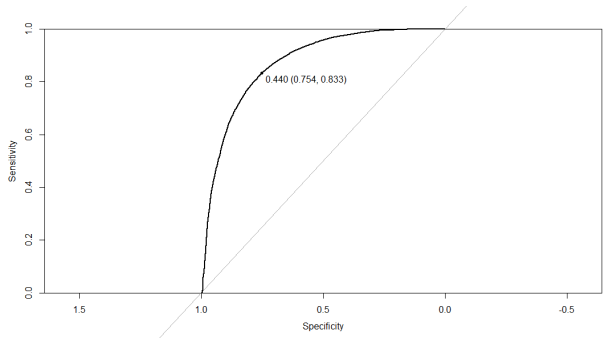

Figure 2: all down sample

Since most variables are of categorical type, in order to have a better understanding of the effect of variables on the results, a step-wise feature adding model is established. The variables are added in the following order: month, race, gender, education, marital status, age, each time the AUC of the model is calculated, Figure 3 is generated. It can be clearly seen from Figure 3 that sex, marital status, and age have a great influence on the response.

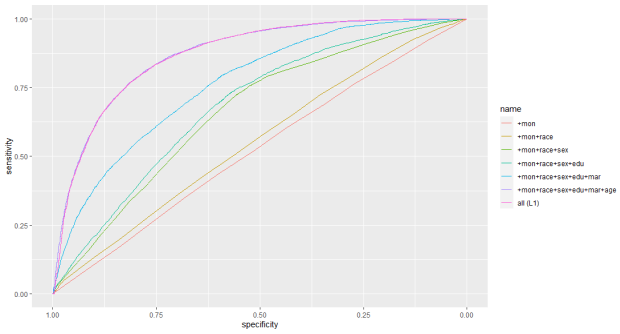

Figure 3: stepwise down sample

4.2 Logistic Regression with Weighted Loss Function

Formula 4 is to calculate the new weights for the minority class and the majority class of 'Y' and 'N'. With default weights, the classifier here will assume that both kinds of label errors have the same cost. But for this unbalanced data set, the wrong prediction of the minority is worse than the wrong prediction of the majority class. Use the entire data set to construct a logistic regression model with regularization, along with weighted loss function, a grid search is performed, and it is found that the optimized penalized parameter lambda is 0.001014139 with L1 norm regularization.

The coefficients of the model are in column 'weighted_coef' of Table 5. By keeping all other predictors at the same level, when age increases by 1, the odds of suicide death occurrence increases by $exp(-0.076467) = 0.9263835$ times. 'w_lower' and 'w_upper' represent the lower bond and upper bond of the coefficient at the significance of 5%. The column 'w_Sig' indicated the significance, 'TRUE' and 'FALSE' represent this (level of) variable is or is not significant at 5% level respectively. After using the obtained coefficients to make predictions on the test set, an AUC of 0.8721828 is obtained which is shown in Figure 4. The step-wise feature adding figure in Figure 5 shows the important variables in the model are basically the same as the down-sampled logistic regression model.

$$weights = \begin{cases} \frac{\text{number of minority class}}{\text{number of all class}} & \text{for majority class} \\ \frac{\text{number of majority class}}{\text{number of all class}} & \text{for minority class} \end{cases} \tag{4}$$

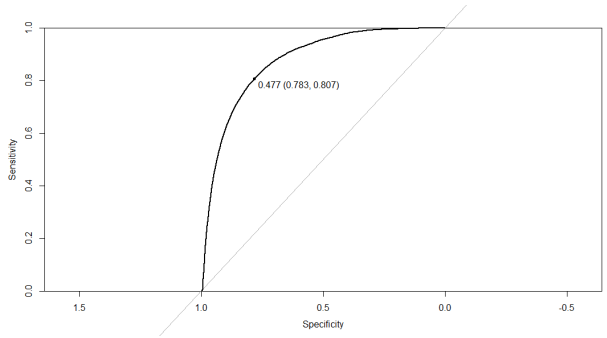

Figure 4: all weighted sample

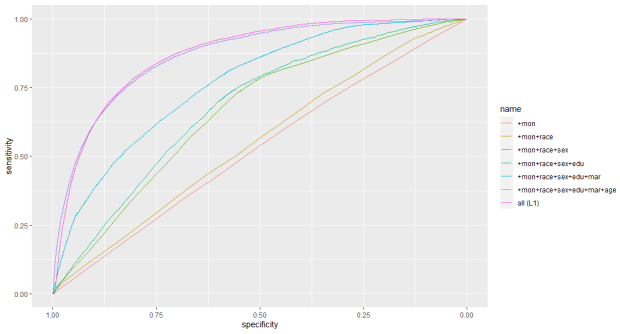

Figure 5: stepwise weighted loss

Table 5: Coefficient and significance

variable	down_coef	d_lower	d_upper	d_Sig	weighted_coef	w_lower	w_upper	w_Sig
(Intercept)	3.927078	3.825892	4.028263	TRUE	4.031376	4.011557	4.051194	TRUE
edu2	0.000587	-0.007895	0.009069	FALSE	-0.000496	-0.002073	0.001081	FALSE
edu3	0.247776	0.191285	0.304267	TRUE	0.224757	0.213814	0.235700	TRUE
edu4	0.442315	0.373794	0.510835	TRUE	0.419355	0.406121	0.432590	TRUE
edu5	0.404306	0.320740	0.487871	TRUE	0.400103	0.384036	0.416170	TRUE
edu6	0.659496	0.587972	0.731020	TRUE	0.620881	0.607070	0.634692	TRUE
edu7	0.501431	0.404929	0.597933	TRUE	0.526277	0.507590	0.544964	TRUE
edu8	0.680135	0.547224	0.813045	TRUE	0.628686	0.603138	0.654233	TRUE
mon02	-0.000871	-0.010305	0.008563	FALSE	-0.000681	-0.002357	0.000995	FALSE
mon03	-0.000371	-0.008242	0.007500	FALSE	-0.000311	-0.001794	0.001172	FALSE
mon04	0.000642	-0.007972	0.009256	FALSE	0.000457	-0.001094	0.002008	FALSE
mon05	0.000232	-0.007329	0.007794	FALSE	0.000202	-0.001236	0.001640	FALSE
mon06	0.003535	-0.064766	0.071836	FALSE	0.000676	-0.000997	0.002348	FALSE
mon07	0.002346	-0.065423	0.070114	FALSE	0.005965	-0.007216	0.019146	FALSE
mon08	0.000571	-0.007828	0.008970	FALSE	0.019911	0.006710	0.033113	TRUE
mon09	0.000910	-0.008684	0.010503	FALSE	0.003562	-0.009810	0.016934	FALSE
mon10	0.000067	-0.007169	0.007303	FALSE	-0.000063	-0.001448	0.001323	FALSE
mon11	-0.029854	-0.099382	0.039674	FALSE	-0.050632	-0.064186	-0.037077	TRUE
mon12	-0.055446	-0.123537	0.012645	FALSE	-0.034609	-0.047834	-0.021385	TRUE
sexM	0.980870	0.939964	1.021777	TRUE	0.987699	0.979787	0.995611	TRUE
age	-0.074819	-0.076114	-0.073523	TRUE	-0.076467	-0.076720	-0.076214	TRUE
marM	-0.207475	-0.255941	-0.159010	TRUE	-0.185242	-0.194643	-0.175840	TRUE
marS	-0.402659	-0.462324	-0.342993	TRUE	-0.384464	-0.396057	-0.372872	TRUE
marW	-0.258116	-0.324637	-0.191596	TRUE	-0.199692	-0.212601	-0.186784	TRUE
race2	-1.356980	-1.423657	-1.290302	TRUE	-1.331619	-1.344593	-1.318646	TRUE
race3	-0.000367	-0.008277	0.007543	FALSE	-0.000933	-0.002808	0.000942	FALSE
race4	0.000411	-0.007596	0.008418	FALSE	0.000386	-0.001137	0.001909	FALSE

4.3 Validation for Transfer Learning

In this session, I will verify the performance of transfer learning, i.e. proving that it indeed can take advantage of the external data to improve the estimator, based on the down-sample data. Denote D as the whole down-sample data. The $\hat{\beta}_D$ is the estimate of the parameter obtained from the previous session according to this whole sample. Denote D_1, D_2 as two half-sample generated by the uniform random sampling without replacement, then $D = \{D_1, D_2\}$. The $\hat{\beta}_{D_1}$ is the estimation of the parameter according to one of the half sample D_1.

To simulate a real-world scenario, in the situation that D_1 is known but D_2 is unknown, the $\hat{\beta}_{D_1}$ seems to be the best choice. However, with the involvement of D_2 as the external data, there are several ways to improve D_1. One most direct way is to combine two half samples into D, then get $\hat{\beta}_D$, however, the transfer learning procedure just proposed suggests another possible way. Here only verify that $\left\|\hat{\beta}_{TL} - \hat{\beta}_D\right\|_2^2 \ll \left\|\hat{\beta}_{D_1} - \hat{\beta}_D\right\|_2^2$. As $\hat{\beta}_D$ is certainly closer than the true β, hence we can here verify that $\hat{\beta}_{TL}$ given by the transfer learning procedure leads a correct way to the true β. This has verified to be true from our code, as

$$\left\|\hat{\beta}_{TL} - \hat{\beta}_D\right\|_2^2 = 0.002268502 \ll 0.03938093 = \left\|\hat{\beta}_{D_1} - \hat{\beta}_D\right\|_2^2$$

Section 5. Model Evaulation and Concluding Remarks

The logistic regression with regularization along with two different methods (down-sampling and weighted loss function) is used for processing the unbalanced data set. To evaluate the model properly, AUC is adopted. The implicit goal of AUC is to deal with the highly skewed distribution of the data set, and not to overfit a single class.

Weighted Logistic regression has a higher AUC of 0.8721828, compared to the down-sampling method with 0.870063. However, the performance of the two logistic regression models is similar, and the model calculation time of the weighted loss function is more than three times that of the down-sampling method. Considering the time consumption of constructing the model, it is concluded that down-sampling is an optimal method for large sample size data, even sacrifice some prediction accuracy.

In future studies, interaction terms and quadratic terms or higher-order terms may be considered to add to the model. In addition, CART model may be a better choice for large sample size data.

References

1. Allison C. Nugent, Elizabeth D. Ballard, Lawrence T. Park & Carlos A. Zarate Jr. Research on the pathophysiology, treatment, and prevention of suicide: practical and ethical issues. BMC Psychiatry volume 19, Article number: 332 (2019)
2. Chao-Ying Joanne Peng, Kuk Lida Lee Gary M. Ingersoll (2002). An Introduction to Logistic Regression Analysis and Reporting Article. The Journal of Educational Research. September 2002 (3-14)
3. Li, Sai, T. Tony Cai, and Hongzhe Li. "Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality." arXiv preprint arXiv:2006.10593 (2020).
4. Ly, Alexander, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. "A tutorial on Fisher information." Journal of Mathematical Psychology 80 (2017): 40-55.
5. National Institute of Mental Health-Suicide Definition https://www.nimh.nih.gov/health/statistics/suicide
6. Rigollet, Philippe, and Alexandre Tsybakov. "Exponential screening and optimal rates of sparse estimation." The Annals of Statistics 39, no. 2 (2011): 731-771.
7. Roger Koenker, Jungmo Yoon. Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy. http://www.econ.uiuc.edu/~roger/research/links/links.pdf
8. Zhou, Shuheng. "Restricted eigenvalue conditions on subgaussian random matrices." arXiv preprint arXiv:0912.4045 (2009).

Appendix

A.1 Data Source and Value of Variable Definitions

Data file please refer to https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/DVS/mortality/mort2019us.zip.

Table 6 to table 13 are definitions of value of variables 'edu','sex','age_dt','age_d','mar','place', and 'race'. Variable 'mon' means Month of Death, its value is from 1 to 12 represents January to December.

Variable 'week' means Day of Week Death, its value is from 1 to 7 and 9, 1 to 7 represents Sunday to Saturday, 9 means unknown.

Table 6: Definition of edu

Variable Name	Full Name
edu	Education (2003 revision)
Value	Definition
1	8th grade or less
2	9 - 12th grade, no diploma
3	high school graduate or GED completed
4	some college credit, but no degree
5	Associate degree
6	Bachelor's degree
7	Master's degree
8	Doctorate or professional degree
9	Unknown

Table 7: Definition of sex

Variable Name	Full Name
sex	Sex
Value	Definition
M	Male
F	Female

Table 8: Definition of age_dt

Variable Name	Full Name
age_dt	Age Detailed Type
Value	Definition
1	Years
2	Months
4	Days
5	Hours
6	Minutes
9	Age not stated

Table 9: Definition of age_d

Variable Name	Full Name
age_d	Age Detailed Number
Value	Definition
001-135	Number
999	Age not stated

Table 10: Definition of mar

Variable Name	Full Name
mar	Marital Status
Value	Definition
S	Never married, single
M	Married
W	Widowed
D	Divorced
U	Marital Status unknown

Table 11: Definition of manner

Variable Name	Full Name
manner	Manner of Death
Value	Definition
1	Accident
2	Suicide
3	Homicide
4	Pending investigation
5	Could not determine
6	Self-Inflicted
7	Natural
Blank	Not specified

Note: Manner of Death equals to 2 means Suicide case, which is denoted by Y. Manner of Death not equals to 2 means not Suicide case, which is denoted by N.

Table 12: Definition of place

Variable Name	Full Name
manner	Manner of Death
Value	Definition

Variable Name	Full Name
0	Home
1	Residential institution
2	School, other institution and public administrative area
3	Sports and athletics area
4	Street and highway
5	Trade and service area
6	Industrial and construction area
7	Farm
8	Other Specified Places
9	Unspecified place
blank	Causes other than W00-Y34, except Y06and Y07

Table 13: Definition of race

Variable Name	Full Name
manner	Manner of Death
Value	Definition
0	Other (Puerto Rico only)
1	White
2	Black
3	American Indian
4	Asian or Pacific Islander

A.2 Data Description Summplement

Table 14 shows the first six observations of import directly from Mortality Multiple Cause Files(2019). Values of variables definition are in Appendix A.1.

Table 14: First Six Observations

edu	mon	sex	age_dt	age_d	mar	week	manner	place	ICD	race
4	1	M	1	36	M	6	7	NA	E141	1
4	1	F	1	63	M	5	7	NA	C55	1
3	1	F	1	97	W	5	7	NA	I698	1
3	1	M	1	76	M	5	7	NA	C80	1
4	1	M	1	64	M	7	7	NA	I633	1
8	1	M	1	74	M	5	7	NA	C250	1

Table 15 to Table 19 are frequency tables of variables Place of Death, Month, Education, Day of Week, and Race. All the unstated values are blank are not counted.

Table 15: Frequency Table of Place of Death

Place of Death	Value	0	1	2	3	4	5	6	7	8
Suicide	Number	31924	905	347	51	1877	1958	510	249	5799
Overall	Number	117895	11712	2226	341	9181	8162	1132	673	18536

Table 16: Frequency Table of Month

Month	Value	01	02	03	04	05	06	07	08	09	10	11	12
Suicide	Num	3480	3265	3764	3658	3712	3740	3814	3860	3802	3765	3337	3423
Overall	Num	225730	203252	223112	206252	207902	198056	202398	202244	198225	211107	212380	226735

Table 17: Frequency Table of Education

Education(2003 version)	Value	1	2	3	4	5	6	7	8
Suicide	Number	1628	5064	17864	7537	3306	5571	1801	849
Overall	Number	230783	263586	1108660	316830	166396	271313	114825	45000

Table 18: Frequency Table of Day of Week

Day of Week	Value	1	2	3	4	5	6	7
Suicide	Number	6160	6796	6664	6178	5990	6143	5689
Overall	Number	356432	359437	363380	357025	357770	361606	361616

Table 19: Frequency Table of Race

Race	Value	1	2	3	4
Suicide	Number	38610	2999	602	1409
Overall	Number	2131072	316930	18658	50733

B. The proof of the Convergence Rate of Trans-logistic Regression

Let the limit of $\hat{w}^A$ is w^A. That is to say,

$$w^A \in \arg\min_{\tilde{w}} \sum_{k\in\{0,1\}} E\left[\log\left(1+\exp\left(-y_i^{(k)}\tilde{x}_i^{(k)T}\tilde{w}\right)\right)\right].$$

Hence, if let $\delta^A = \beta - w^A$, then δ^A can be expressed by $\delta^{(1)}$, i.e. $\beta - w^{(1)}$:

$$\delta^A = \frac{n_1}{n_1+n_0}\delta^{(1)}.$$

Moreover, define the estimation of Hessian matrix:

$$\begin{aligned}
\hat{H} &= \tfrac{n_0}{n_0+n_1}\hat{H}(\beta) + \tfrac{n_1}{n_0+n_1}\hat{H}\left(w^{(1)}\right)\\
\hat{H}(\beta) &= \tfrac{1}{n_0}\sum_{i=1}^{n_0}\frac{\exp\left(-y_i^{(0)}\tilde{x}_i^{(0)T}\tilde{\beta}\right)}{\left(1+\exp\left(-y_i^{(0)}\tilde{x}_i^{(0)T}\tilde{\beta}\right)\right)^2}\tilde{x}_i^{(0)}\tilde{x}_i^{(0)T}\\
\hat{H}\left(w^{(1)}\right) &= \tfrac{1}{n_0}\sum_{i=1}^{n_0}\frac{\exp\left(-y_i^{(0)}\tilde{x}_i^{(0)T}w^{(1)}\right)}{\left(1+\exp\left(-y_i^{(0)}\tilde{x}_i^{(0)T}w^{(1)}\right)\right)^2}\tilde{x}_i^{(0)}\tilde{x}_i^{(0)T}
\end{aligned}$$

Lemma 1 (Restricted eigenvalue condition)

Under the restricted eigenvalue condition, with positive r_1 and r_2 such that $r_1 \left(\log p/ \left(n_1 + n_0\right)\right) = o\left(1\right)$ and $r_2 \left(\log p/n_0\right) = o\left(1\right)$, we have

$$\min \left\{ \inf_{0 \neq u \in B_1(r_1)} \frac{u^T \hat{H} u}{\|u\|_2^2} \geq \phi_0, \inf_{0 \neq u \in B_1(r_2)} \frac{u^T \hat{H}\left(\beta\right) u}{\|u\|_2^2} \right\} \geq \phi_0.$$

Lemma 2 (w part)

Under the conditions 1,2 and 3, for $\hat{u} = \hat{w}^A - w^A$, we have

$$\begin{aligned} &\max \left\{ \left(\hat{u}\right)^T \hat{H} \hat{u}, \|\hat{u}\|_2^2 \right\} = O_P\left(s\lambda_w^2 + \lambda_w h\right) \\ &\|\hat{u}\|_1 = O_P\left(s\lambda_w + h\right), \end{aligned}$$

Lemma 3 (δ part)

Under the conditions 1,2 and 3, for $\hat{v} = \hat{\delta}^A - \delta^A$, we have

$$\begin{aligned} &\max \left\{ \left(\hat{v}\right)^T \hat{H} \hat{v}, \|\hat{v}\|_2^2 \right\} = O_P\left(s\lambda_w^2 + \lambda_w h + \lambda_\delta h\right) \\ &\|\hat{v}\|_1 = O_P\left(s\lambda_\delta + h\right) \end{aligned}$$

Proof of theorem:

Combining the following inequalities

$$\begin{aligned} &\left\|\hat{\beta}_{oracle} - \beta\right\|_2^2 = \left\|\hat{w}^A + \hat{\delta}^A - \left(w^A + \delta^A\right)\right\|_2^2 = \left\|\left(\hat{w}^A - w^A\right) + \left(\hat{\delta}^A - \delta^A\right)\right\|_2^2 \\ &\leq \left\|\hat{w}^A - w^A\right\|_2^2 + \left\|\hat{\delta}^A - \delta^A\right\|_2^2 = \|\hat{u}\|_2^2 + \|\hat{v}\|_2^2 \end{aligned}$$

and

$$\begin{aligned} &\left(\hat{\beta}_{oracle} - \beta\right)^T H\left(\beta\right) \left(\hat{\beta}_{oracle} - \beta\right) \\ &\leq \left(\hat{w}^A - w^A\right)^T H\left(\beta\right) \left(\hat{w}^A - w^A\right) + \left(\hat{\delta}^A - \delta^A\right)^T H\left(\beta\right) \left(\hat{\delta}^A - \delta^A\right), \end{aligned}$$

lemma 2 and lemma 3,We can draw the final conclusion directly by noting $\left\|H\left(\beta\right) - \hat{H}\right\| \to 0$ under the condition 1 as $n \to \infty$.