Two-stage polytmous logistic regression tutorial

Haoyu Zhang *Department of Biostatistics, Johns Hopkins University, Baltimore, MD, U.S.A.*
Ni Zhao *Department of Biostatistics, Johns Hopkins University, Baltimore, MD, U.S.A.*
Thomas U. Ahearn *National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, MD, U.S.A.*
Montserrat García-Closas *National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, MD, U.S.A.*
William Wheeler *Information Management Services, Inc., Rockville, MD, U.S.A*
Nilanjan Chatterjee *Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA*

Cancers are routinely classified into subtypes according to various features, including histo pathological characteristics and molecular markers. Genomic investigations have reported heterogeneous association between loci and cancer subtypes. However, it is not evident what is the optimal modeling strategy for handling correlated tumor features, missing data, and increased degrees-of-freedom in the underlying tests of associations. In this tutorial, we proposed a two-stage polytomous regression framework to handle cancer data with multivariate tumor characteristics. In the first stage, a standard polytomous model is used to specify for all subtypes defined by the cross-classification of different markers. In the second stage, the subtype-specific case-control odds ratios are specified using a more parsimonious model based on the case-control odds ratio for a baseline subtype, and the case-case parameters associated with tumor markers. Further, to reduce the degrees-of-freedom, we allow to specify case-case parameters for additional markers using a random-effect model. We use the EM algorithm to account for missing data on tumor markers. The score-test distribution theory is developed by borrowing analogous techniques from group-based association tests. Through simulations across a range of realistic scenarios, we show the proposed methods outperforms alternative methods for identifying heterogenous associations between risk loci and tumor subtypes.

Keywords: Two-stage polytomous model; Susceptibility variants; Cancer subtypes; EM algorithm; Score tests; Etiologic heterogeneity.

Overview

This vegnette provides an introduction to the 'TOP' package. To load the package, users need to install package from CRAN and TOP from github. The package can be loaded with the following command:

```
# install.packages("devtools")
library(devtools)
#install_github("andrewhaoyu/TOP")
library(TOP)
```

Two-stage polytomous model

In this vegnette, we will decomstrate the methods with a breast cancer example. There are around 5,112 cases and 4,888 controls in the dataset. Four different tumor characteristics were included:

ER (positive vs negative), PR (positive vs negative), HER2 (positive vs negative), grade (ordinal 0, 1, 2).

For simplicity, we will first demonstrate the two-stage model with three binary tumor characteristics (ER, PR, and HER2). These three tumor characteristics could define 8 different breast cancer subtypes (8=2x2x2). We included two covariates, one is a SNP that we are interested. The second one is the first principal component (PC1). Let D_i denote the disease status, taking values in $\{0, 1, 2, \cdots, 8\}$, of the ith ($i \in 1, \cdots, 10,000$) subject in the study. $D_i = 0$ represents a control, and $D_i = m$ represent a subject with disease of subtype m. Let G_i be the genotype for ith subject and X_i be the PC1 for the ith subject. In the first-stage model, we use the standard "saturated" polytomous logistic regression model

$$Pr(D_i = m|G_i, X_i) = \frac{\exp(\beta_m G_i + \eta_m X_i)}{1 + \sum_{m=1}^{8} \exp(\beta_m G_i + \eta_m X_i)}$$

where β_m and η_m is the regression coefficients for the SNP and PC1 for association with the mth subtype.

Because each cancer subtype m is defined through a unique combination of the 4 characteristics, we can always alternatively index the parameters β_m as $\beta_{s_1 s_2 s_3}$, where $s_1, s_2, s_3 \in \{0, 1\}$ for the three binary tumor characteristics. Originally β_1 could be the coefficient of cancer subtype ER-PR-HER2-. With the new index, β_1 could be written as β_{000}, which means the ER, PR, HER2 are all negative. With this new index, we can represent the log odds ratio as

$$\beta_{s_1 s_2 s_3} = \theta^{(0)} + \theta_1^{(1)} s_1 + \theta_2^{(1)} s_2 + \theta_3^{(1)} s_3 + \theta_{12}^{(2)} s_1 s_2 + \theta_{13}^{(2)} s_1 s_3 + \theta_{23}^{(2)} s_2 s_3 + \theta_{123}^{(3)} (s_1 s_2 s_3).$$

Here $\theta^{(0)}$ represents the standard case-control log odds ratio for a reference disease subtype compared to the control and $\theta_k^{(1)}$ represents a case-case log odds ratio associated with the levels of kth tumor characteristics after adjusting for other tumor characteristics, and $\theta_{k_1 k_2}^{(2)}$ represent case-case log odds ratios associated with pairwise interactions among the tumor characteristics and so on. This decomposition is equivalent with the first stage model. Since both the first stage and second stage have 8 parameters. We called this saturated model. The users could construct different two-stage model by assuming different second stage parameters to be 0. For example, the baseline only model:

$$\beta_{s_1 s_2 s_3} = \theta^{(0)}.$$

This model assumes all of the subtypes have the same log odds ratio. So it is equivalent to the standard logistic regression. We can also construct the additive two-stage model by assuming all of the second stage interactions parameters are 0, then the second stage decomposition becomes,

$$\beta_{s_1 s_2 s_3} = \theta^{(0)} + \theta_1^{(1)} s_1 + \theta_2^{(1)} s_2 + \theta_3^{(1)} s_3$$

Furthermore, we could construct the pairwise interaction two-stage model by assuming all of the second stage higher order interactions parameters are 0, then the second stage decomposition becomes,

$$\beta_{s_1 s_2 s_3} = \theta^{(0)} + \theta_1^{(1)} s_1 + \theta_2^{(1)} s_2 + \theta_3^{(1)} s_3 + \theta_{12}^{(2)} s_1 s_2 + \theta_{13}^{(2)} s_1 s_3 + \theta_{23}^{(2)} s_2 s_3$$

```
library(TOP)
#load in the breast cancer example
data(data, package="TOP")
```

```r
#this is a simulated breast cancer example
#there are around 5000 breast cancer cases and 5000 controls disease
data[1:5,]
#four different tumor characteristics were included,
#ER (positive vs negative),
#PR (positive vs negative),
#HER2 (positive vs negative)
#the phenotype file
y <- data[,1:4]
#one SNP
#one Principal components (PC1) are the covariates
SNP <- data[,6,drop=F]
PC1 <- data[,7,drop=F]
#fit the additive two-stage model
model.1 <- TwoStageModel(y=y,additive=cbind(SNP,PC1),
                         missingTumorIndicator = 888)
```

```r
#the model result is a list
#model.1[[4]] are the second stage odds ratio (95% CI)
#and p-value, the baseline effect is the case-control
#odds ratio of the reference subtype (ER-,PR-,HER2-,grade0).
#The main effect are the case-case odds ratio of the tumor characteristics.
(model.1[[4]])
```

```
##   Covariate SecondStageEffect OddsRatio OddsRatio(95%CI low)
## 1       SNP   baseline effect      0.97                 0.84
## 2       SNP    ER main effect      0.98                 0.87
## 3       SNP    PR main effect      1.06                 0.93
## 4       SNP  HER2 main effect      1.15                 0.99
## 5       PC1   baseline effect      1.13                 1.03
## 6       PC1    ER main effect      1.01                 0.93
## 7       PC1    PR main effect      0.95                 0.88
## 8       PC1  HER2 main effect      0.94                 0.86
##   OddsRatio(95%CI high)  Pvalue
## 1                  1.13 0.69200
## 2                  1.11 0.76700
## 3                  1.20 0.37000
## 4                  1.33 0.06380
## 5                  1.24 0.00755
## 6                  1.08 0.87100
## 7                  1.03 0.21500
## 8                  1.03 0.17700
```

```r
#model.1[[5]] are the global association test and
#global heterogeneity test result of the covariates.
model.1[[5]]
```

```
##   Covariate Global test for association Global test for heterogeneity
## 1       SNP                     0.10100                         0.248
## 2       PC1                     0.00921                         0.466
```

```
#model.1[[7]] are the case-control odds ratios
#for all of the subtypes.
head(model.1[[7]])
```

```
##   Covariate    Subtypes OddsRatio OddsRatio(95%CI low)
## 1       SNP ER0PR0HER20      0.97                 0.84
## 2       PC1 ER0PR0HER20      1.13                 1.03
## 3       SNP ER1PR0HER20      0.95                 0.81
## 4       PC1 ER1PR0HER20      1.14                 1.03
## 5       SNP ER0PR1HER20      1.03                 0.94
## 6       PC1 ER0PR1HER20      1.08                 1.02
##   OddsRatio(95%CI high)  Pvalue
## 1                  1.13 0.69200
## 2                  1.24 0.00755
## 3                  1.12 0.55900
## 4                  1.26 0.01010
## 5                  1.12 0.52700
## 6                  1.14 0.00623
```

```
#instead of additive model, you can also try
#different combinations. For example, for the PC1,
#we use the additive model, but for SNP,
#we use the baseline only model.
model.2 <- TwoStageModel(y=y,baselineonly = SNP,
                         additive=PC1,
                         missingTumorIndicator = 888)
```

The results of the function contain three elements: 1. p.dir is the p-value of likelihood ratio test based on emprical distrition. 2. p.aud is the p-value by approximating the null distribution as a mixture of a point mass at zero with probability b and weighted chi square distribution with d degrees of freedom with probality of 1-b. 3. LR is the likelihood ratio test statistics.

References

1. N. Zhao, H. Zhang, J. Clark, A. Maity, M. Wu. Composite Kernel Machine Regression based on Likelihood Ratio Test with Application for Combined Genetic and Gene-environment Interaction Effect (Submitted)