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.*
William Wheeler *Information Management Services, Inc., Rockville, MD, U.S.A*
Montserrat García-Closas *National Cancer Institute, Division of Cancer Epidemiology and Genetics, 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:

```r
# 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 are 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 3 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. $\theta_k^{(1)}$ represents a case-case log odds ratio associated with the levels of kth tumor characteristics after adjusting for other tumor characteristics. We also refer $\theta_k^{(1)}$ as the main effect of the kth tumor characteristic. And $\theta_{k_1 k_2}^{(2)}$ represents how the case-case log odds ratio associated k_1th tumor characteristic is modified by the levels of the k_2th tumor characteristic and vice versa. We also refer $\theta_{k_1 k_2}^{(2)}$ as the pairwise interaction between the k_1th and k_2th tumor characteristic. And $\theta_{123}^{(3)}$ represent the third order interactions of the three tumor characteristics. Since both the first stage and second stage have 8 parameters, this decomposition is equivalent with the first stage polytomous logistic regression model. 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,

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

```r
library(TOP)
#load in the breast cancer example
data(data, package="TOP")
#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.
#the p-value is for individual tumor heterogeneity test of each second stage parameter.
(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
```

```
#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 result is a list containing 9 elements. 1. the second stage parameters 2. the covariance matrix for the second stage parameters. 3. the second stage parameters organzied for each covariate. 4. The case-control odds ratio and case-case odds ratios of tumor characteristics. 5. Global association test and global heterogeneity test result (Wald test based) 6. The first stage parameter organized for each covariate 7. First stage odds ratio of all the subtypes. 8. Likelihood 9. AIC

Two-stage polytomous model self design second stage

Instead of using the hierarchical second stage decomposition we discussed in last section, the two-stage model also allows the user to self design the second stage matrix. For example, we could define five intrinsic breast cancer subtypes based on the four tumor characteristics: ER, PR, HER2, grade. The five intrinsic subtypes are: 1. (ER or PR)+, HER2-, grade 1 or 2; 2. (ER or PR)+, HER2+;

3. (ER or PR)+, HER2-, grade 3; 4. (ER & PR)-, HER2-; 5. ER-PR-HER2-. We are interested in estimating the case-control log odds ratios of these intrinsic subtypes.

```r
library(TOP)
#load in the breast cancer example
data(data, package="TOP")
#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)
#grade (oridinal 1,2,3)
#the phenotype file
y <- data[,1:5]
#generate the combinations of all the subtypes
#by default, we remove all the subtypes with less than 10 cases
z.standard <- GenerateZstandard(y)
M <- nrow(z.standard) #M is the total number of first stage subtypes

#initial a z.design matrix with M rows, and 5 columns
#each row represent a first stage subtype
#each column represent an aggregated subtype
z.design <- matrix(0,M,5)
#define names for the five intrinsic subtypes
colnames(z.design) <- c("HR+_HER2-_lowgrade",
                        "HR+_HER2+",
                        "HR+_HER2-_highgrade",
                        "HR-_HER2+",
                        "HR-_HER2-")
#To construct a self design second stage matrix,
#we need to find the correpsonding first stage subtypes
#belonging to specific aggregated subtypes
#for first subtype HR+_HER2-_lowgrade
idx.1 <- which((z.standard[,1]==1|z.standard[,2]==1)
               &z.standard[,3]==0
               &(z.standard[,4]==0|z.standard[,4]==1))
z.design[idx.1,1] <- 1
#for second subtype HR+_HER2+
idx.2 <- which((z.standard[,1]==1|z.standard[,2]==1)
               &z.standard[,3]==1)
z.design[idx.2,2] <- 1
#for third subtype HR+_HER2-_highgrade
idx.3 <- which((z.standard[,1]==1|z.standard[,2]==1)
               &z.standard[,3]==0
               &z.standard[,4]==2)
z.design[idx.3,3] <- 1
```

```
#for third subtype HR-_HER2+
idx.4 <- which(z.standard[,1]==0&z.standard[,2]==0
               &z.standard[,3]==1)
z.design[idx.4,4] <- 1
#for third subtype HR-_HER2-
idx.5 <- which(z.standard[,1]==0&z.standard[,2]==0
               &z.standard[,3]==0)
z.design[idx.5,5] <- 1
#one SNP and one Principal components (PC1) are the covariates
SNP <- data[,6,drop=F]
PC1 <- data[,7,drop=F]

model.3 <- EMmvpolySelfDesign(y,
          x.self.design = SNP,
    z.design = z.design,
    additive=PC1,
  missingTumorIndicator = 888)
```

```
#model.3[[4]] are the second stage odds ratio (95% CI)
#and p-value of the intrinsic subtypes
#the second stage odds ratios under this model are the case-control odds ratios for the intrinsi
(model.3[[4]])
```

```
##   Covariate   SecondStageEffect OddsRatio OddsRatio(95%CI low)
## 1       SNP  HR+_HER2-_lowgrade      1.02                 0.93
## 2       SNP           HR+_HER2+      1.12                 1.02
## 3       SNP HR+_HER2-_highgrade      1.06                 0.93
## 4       SNP           HR-_HER2+      1.03                 0.73
## 5       SNP           HR-_HER2-      0.94                 0.66
##   OddsRatio(95%CI high) Pvalue
## 1                  1.12 0.7100
## 2                  1.22 0.0161
## 3                  1.22 0.3890
## 4                  1.44 0.8750
## 5                  1.36 0.7570
```

```
#model.1[[5]] are the global association test and
#global heterogeneity test result of the covariates.
#Note global heterogeneity under self designed
#second stage matrix don't have interpretation
model.3[[5]]
```

```
##   Covariate Global test for association Global test for heterogeneity
## 1       SNP                       0.251                          0.16
```

```
#model.1[[7]] are the case-control odds ratios
#for all of the subtypes.
head(model.3[[7]])
```

```
##   Covariate          Subtypes OddsRatio OddsRatio(95%CI low)
## 1       SNP ER1PR0HER20grade0      1.02                 0.93
## 2       SNP ER0PR1HER20grade0      1.09                 0.97
## 3       SNP ER1PR1HER20grade0      1.02                 0.93
## 4       SNP ER0PR0HER21grade0      1.04                 0.97
## 5       SNP ER1PR0HER21grade0      1.02                 0.93
## 6       SNP ER0PR1HER21grade0      1.03                 0.94
##   OddsRatio(95%CI high) Pvalue
## 1                  1.12  0.710
## 2                  1.23  0.150
## 3                  1.12  0.710
## 4                  1.12  0.255
## 5                  1.12  0.710
## 6                  1.14  0.506
```

Fixed effect two-stage model score test (FTOP)

To construct the score test for a fixed effect two stage model, we need two steps. First, we need to fit the model under the null hypothesis that the second stage parameters of SNP is 0. In other words, null of the subytpes is associated with the SNP. Second we need should compute the score and information matrix for each SNP. Based on the score and information matrix, we could construct the score test statistics for global association test.

```
#fit the two-stage model under the null hypothesis
#that the second stage parameters of SNP is 0
#the model only has one covariate PC1
score.support.fixed <- ScoreTestSupportMixedModel(y=y,
                                                  additive=PC1,
                                                  missingTumorIndicator=888)
```

```
## [1] "EM Algorithm Converged"
```

```
#Generate the additive second stage design matrix
z.additive <- cbind(1,z.standard)
#compute the score and information matrix for SNP
score.test.fixed <- ScoreTestMixedModel(y=y,
                                        x=SNP,
                                        z.design = z.additive,
                                        score.test.support=score.support.fixed,
                                        missingTumorIndicator=888)
#the first element is the score
#the second element is the information matrix
score.fixed <- score.test.fixed[[1]]
```

```
infor.fixed <- score.test.fixed[[2]]
#compute the global association test p value
p.value.ftop <- DisplayFixedScoreTestResult(score.fixed,infor.fixed)
print(p.value.ftop)
```

```
## [1] 0.0539
```

In general, I don't recommend to do the global heterogeneity test under the fixed-effect two-stage model. Since the global heterogeneity test requires to estimate the baseline effect of every single SNP, it loses the advantage of score test that it only needs to estimate the parameters of the other covariates under the null hypothesis for one time. Instead, I recommend using Wald test to do the global heterogeneity test. It will automatically give you the results of global heterogeneity test (model.1[[4]]) and individual heterogeneity test (model.1[[5]]). And the Wald test and score test asymptotically have similar power.

The general analysis pipeline is like this: First, we performed the global association test for the whole genome. This step is fast since we only need to fit the function ScoreTestSupportMixedModel one time under the null hypothesis, every single SNP only needs to run the function ScoreTestMixedModel. Then we have the potential regions with SNPs genome-wide significant associated with the disease. For every SNP in these regions, we can fit the function TwoStageModel. This function will give you the global heterogeneity test and individual heterogeneity test results. Since there are limited number of SNPs in these regions, this process won't cost too much time.

Mixed effect two-stage model score test (MTOP)

To construct the score test for a mixed effect two stage model, we need four steps. First, we need to fit the model under the null hypothesis that the second stage parameters of SNP is 0. In other words, null of the subytpes is associated with the SNP. Second, we need should compute the score and information matrix of fixed effect for each SNP. Then, we need to fit the model under the null that the variance of random effect is 0. Finally, we need to compute the score and information matrix of the random effect terms. With the score and information of fixed effect and random effect, we could construct the global association test for the mixed effect two-stage model.

```
#we are going to build a two-stage model with
#baseline parameter and ER case-case parameter as fixed
#We assume the PR, HER2, grade case-case parameter
#to be random
#fit the two-stage model under the null hypothesis
#that the second stage parameters of SNP is 0
#the model only has one covariate PC1
#Generate the z design matrix for fixed effect
z.design.fixed <- cbind(1,z.standard[,1])
#compute the score and information matrix for fixed effect
score.test.fixed <- ScoreTestMixedModel(y=y,
                    x=SNP,
                    z.design=z.design.fixed,
                    score.test.support=score.support.fixed,
```

```r
                    missingTumorIndicator=888)
#the first element is the score
#the second element is the information matrix
score.fixed <- score.test.fixed[[1]]
infor.fixed <- score.test.fixed[[2]]

#fit the two-stage model under the null hypothesis
#that only the random effect is 0
#the model will have two covariates,
#PC1 and the fixed effect of SNP
score.support.random <- ScoreTestSupportMixedModelSelfDesign(y=y,
                        x.self.design  = SNP,
                        z.design = z.design.fixed,
                        additive = PC1,
                        missingTumorIndicator = 888)
```

```
## [1] "EM Algorithm Converged"
```

```r
#Generate the z design matrix for random effect
#PR, HER2 and grade is random effect
z.design.random <- z.standard[,2:4]

#compute the score and information matrix for random effect
score.test.random <- ScoreTestMixedModel(y=y,
                                         x=SNP,
                                         z.design=z.design.random,
                                         score.test.support=score.support.random,
                                         missingTumorIndicator=888)
#the first element is the score
#the second element is the information matrix
score.random <- score.test.random[[1]]
infor.random <- score.test.random[[2]]

#after we get the the fixed effect score, infor
#and random effect score, infor,
#we can combine them through the following function.
#two p value will be generated.
#the first p value for null hypothesis that
#both the fixed effect and the variance of the random effects are 0
#under this case, the first p-value is for the global association test.
#the second p value is for the null hypothesis
#that the variance of random effect is 0
#Under this situation, the second p value is NOT
#the global heterogeneity test p value since ER is fixed
p.value.mtop <- DisplayMixedScoreTestResult(
  score.fixed,
  infor.fixed,
  score.random,
  infor.random
)
print(p.value.mtop)
```

```
## [1] 0.09988269 0.14110520
```

The second p.value of p.value.mtop is for null hypothesis that the random effect is 0. Under this particular situation, the second p.value is not for the p-value of global heterogeneity test, since ER is fixed. To construct the global heterogeneity test for this setting, we need four steps. First, we need to fit the model under the null hypothesis that the fixed effect (ER) and variance of the random effects (PR, HER2, and grade) is 0. In other words, only the baseline effect can be non-zero. Second, we need should compute the score and information matrix of fixed effect (ER) for each SNP. Then, we need to fit the model under the null that the variance of random effect is 0. Finally, we need to compute the score and information matrix of the random effect terms. With the score and information of fixed effect and random effect, we could construct the global association test for the mixed effect two-stage model.

```
#we are going to build a two-stage model with
#baseline parameter and ER case-case parameter as fixed
#We assume the PR, HER2, grade case-case parameter
#to be random.
#fit the two-stage model under the null hypothesis
#that the fixed effect (ER) and
#variance of the random effects (PR, HER2, and grade) is 0.
#the model will have two covariates,
#PC1 and the fixed effect of SNP.
#Generate the z design matrix for fixed effect
z.design.fixed.baseline <- matrix(1,nrow(z.standard),1)
#compute the score and information matrix for fixed effect
score.support.fixed <- ScoreTestSupportMixedModelSelfDesign(y=y,
                        x.self.design  = SNP,
                        z.design = z.design.fixed.baseline,
                        additive = PC1,
                        missingTumorIndicator = 888)
```

```
## [1] "EM Algorithm Converged"
```

```
z.design.ER <- z.standard[,1,drop=F]
score.test.fixed <- ScoreTestMixedModel(y=y,
                    x=SNP,
                    z.design=z.design.ER,
                    score.test.support=score.support.fixed,
                    missingTumorIndicator=888)
#the first element is the score
#the second element is the information matrix
score.fixed <- score.test.fixed[[1]]
```

```r
infor.fixed <- score.test.fixed[[2]]

#fit the two-stage model under the null hypothesis
#that only the random effect is 0
#the model will have two covariates,
#PC1 and the fixed effect of SNP
z.design.fixed <- cbind(z.design.fixed.baseline,z.design.ER)
score.support.random <- ScoreTestSupportMixedModelSelfDesign(y=y,
                        x.self.design  = SNP,
                        z.design = z.design.fixed,
                        additive = PC1,
                        missingTumorIndicator = 888)
```

```
## [1] "EM Algorithm Converged"
```

```r
#Generate the z design matrix for random effect
#PR, HER2 and grade is random effect
z.design.random <- z.standard[,2:4]

#compute the score and information matrix for random effect
score.test.random <- ScoreTestMixedModel(y=y,
                                         x=SNP,
                                         z.design=z.design.random,
                                         score.test.support=score.support.random,
                                         missingTumorIndicator=888)
#the first element is the score
#the second element is the information matrix
score.random <- score.test.random[[1]]
infor.random <- score.test.random[[2]]

#after we get the the fixed effect score, infor
#and random effect score, infor,
#we can combine them through the following function.
#two p value will be generated.
#the first p value for null hypothesis that
#both the fixed effect and the variance of the random effects are 0
#under this case, the first p-value is for the global heterogeneity test.
#the second p.value don't have a good interpretation under this case
p.value.mtop <- DisplayMixedScoreTestResult(
  score.fixed,
  infor.fixed,
  score.random,
  infor.random
)
print(p.value.mtop)
```

```
## [1] 0.2700321 0.1411052
```

References

1. Zhang, H., Zhao, N., Ahearn, T.U, Wheeler W., García-Closas, M., Chatterjee, N., A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics (Submitted)