485 Paper 3 Reproducibility Appendix

Siddarth Marthi, Jiarui Wang, Haoran Cheng

2023-04-05

Overview of appendix: This code is designed to test 6 different models to see which one does the best job of predicting who will win the NCAA tournament. Those models differ in whether to incorporate home advantage variable and date interaction term, as well as whether there is penalty term added to the model formula. and then from there we will use cross-validation to compute negative log likelihood loss, hoping to see which model to prefer and then ranking the teams based on the model. Lastly, we want to calculate Michigan's odds of winning against the teams that they played in the first two rounds of tournament. Multiplication correction with Bonferroni method is performed to control family-wise type I error. Extra part of multicolinearity check on our chosen model is added for paper writing use. 1
We fitted a B-T model without home advantage variable.
2
We fitted penalized version of the B-T model in part 1 with bayesglm().
3
For part a, we included an additional home advantage variable-the intercept, and apply both the non-penalized and penalized model fitting method to it.
For part b, we excluded the intercept, but added another date interaction variable. Similarly, we fitted both non-penalized and penalized version.
4
We calculated negative log likelihood loss to our different models with cv.glmnet(). We pick the non-penalized model without intercept but with date interaction to be the best since it has the smallest loss value. 5
We rank the coefficients of our picked model and get the top 10/bottom 5 team variables correspondingly. 6
We computed the odds of Michigan winning three teams respectively: Delaware.State, St..Francis..PA. and Western.Michigan According to formula:
$logit(odd) = log\frac{P_{i>j}}{P_{i<j}} = log\frac{e^{\beta}_{i}}{e^{\beta}_{j}} = \beta_i - \beta_j.$
7
We applied Bonferroni method to do mulitplicity correction to control type I error of simultaneous hypothesis testing.
Extra part
We used vif() function to check whether there is multicollinearity issue within our model.

##Data paraperation Changing data to fit date to the date variable type

```
ncaa_womens <-
  read.csv("http://stat.lsa.umich.edu/~bbh/s485/data/cbb-womens-2023-03-12.csv")


ncaa_womens$date <- as.Date(ncaa_womens$date)
```

To identify opponents of Michigan in the first two rounds of tournament for later use in part 6:

```
opponent <- ncaa_womens[ncaa_womens$Michigan != 0,]
```

```
row_values <- opponent[1, ]
indices <- which(row_values != 0)
elements <- row_values[indices]
name <- colnames(opponent)[indices]
name
```

```
## [1] "home_win"       "date"           "Delaware.State" "Michigan"
```

```
row_values <- opponent[2, ]
indices <- which(row_values != 0)
elements <- row_values[indices]
name <- colnames(opponent)[indices]
name
```

```
## [1] "home_win"         "date"             "Michigan"         "St..Francis..PA."
```

```
row_values <- opponent[3, ]
indices <- which(row_values != 0)
elements <- row_values[indices]
name <- colnames(opponent)[indices]
name
```

```
## [1] "home_win"         "date"             "Michigan"         "Western.Michigan"
```

The three opponents of Michigan in the first two rounds are: Delaware.State, St..Francis..PA. and Western.Michigan.

#1. Fit a plain-vanilla B-T model To fit a B-T model without date, intercept and Michigan for reference:

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

The reason why we excluded intercept is that it can represent home advantage, the greater of which indicates the greater of home advanrage.

#2. Fitting a version of the model using a penalized form of logistic regression To fit a penalized version of B-T model without date, home advantaage and Michigan for reference, we use arm::bayesglm():

```
btmod2 <- arm::bayesglm(home_win ~ .-date - Michigan  -1, data = ncaa_womens,
                        family = binomial)
```

#3a. Incorporate a model with home team advantage parameter Add intercept to address home advantage:

```
btmod3 <- glm(home_win ~  .-date - Michigan, data = ncaa_womens,
              family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
btmod4 <- arm::bayesglm(home_win ~  .-date - Michigan, data = ncaa_womens,
                        family = binomial)
```

#3b. Incorporating model interaction of the team's variables as a function of game date
Add another term of date interaction (.-date - Michigan-1)*date to take strength change over the season into consideration:

```
btmod5 <- glm(home_win ~ .-Michigan - date -1 + (.-date - Michigan-1)*date,
              data = ncaa_womens, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
btmod6 <- arm::bayesglm(home_win ~ .-Michigan - date -1 +
                          (.-date - Michigan-1)*date, data = ncaa_womens,
                        family = binomial)
```

#4. deciding which model to prefer #change to log-likelihood.

```
matrix1 <- model.matrix(btmod1)
matrix2 <- model.matrix(btmod2)
matrix3 <- model.matrix(btmod3)
matrix4 <- model.matrix(btmod4)
matrix5 <- model.matrix(btmod5)
matrix6 <- model.matrix(btmod6)


mod1 <- cv.glmnet(matrix1, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl1,intercept = FALSE, type.measure = "deviance")
mod2 <- cv.glmnet(matrix2, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl2,intercept = FALSE, type.measure = "deviance")
mod3 <- cv.glmnet(matrix3, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl3,intercept = FALSE, type.measure = "deviance")
mod4 <- cv.glmnet(matrix4, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl4,intercept = FALSE, type.measure = "deviance")
mod5 <- cv.glmnet(matrix5, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl5,intercept = FALSE, type.measure = "deviance")
mod6 <- cv.glmnet(matrix6, ncaa_womens$home_win, alpha=0, nfolds= 15,
                  object = btmodl6,intercept = FALSE, type.measure = "deviance")

mod1
```

```
##
## Call:  cv.glmnet(x = matrix1, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##      Lambda Index Measure       SE Nonzero
## min 0.04048    77  0.4891 0.006887     427
## 1se 0.16341    62  0.4952 0.006198     427
```

```
mod2
```

```
##
## Call:  cv.glmnet(x = matrix2, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##      Lambda Index Measure       SE Nonzero
## min 0.04442    76  0.4897 0.006364     427
## 1se 0.16341    62  0.4954 0.005990     427
```

```
mod3
```

```
##
## Call:  cv.glmnet(x = matrix3, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##      Lambda Index Measure       SE Nonzero
## min 0.03688    78  0.4859 0.007582     427
## 1se 0.16341    62  0.4924 0.007186     427
```

```
mod4
```

```
##
## Call:  cv.glmnet(x = matrix4, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##      Lambda Index Measure       SE Nonzero
## min 0.03688    78  0.4858 0.008069     427
## 1se 0.17934    61  0.4939 0.006777     427
```

```
mod5
```

```
##
## Call:  cv.glmnet(x = matrix5, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##     Lambda Index Measure       SE Nonzero
## min  32.70   100  0.2310 0.001766     855
## 1se  47.44    96  0.2327 0.001778     855
```

```
mod6
```

```
##
## Call:  cv.glmnet(x = matrix6, y = ncaa_womens$home_win, type.measure = "deviance",      nfolds = 15,
##
## Measure: Mean-squared Error
##
##     Lambda Index Measure       SE Nonzero
## min  32.70   100  0.2310 0.001536     855
## 1se  43.23    97  0.2323 0.001542     855
```

```
#log likelihood (4)
```

Due to the fact that the model that incorporates the home team advantage parameter and relative strength of the team at the end of the season without penalty has the smallest log likelihood, we would prefer this model to the others.

5. ranking the teams and presenting the top 10 and bottom 5

```
l <- sort(coef(btmod4), decreasing = TRUE)
m <- sort(coef(btmod4), decreasing = FALSE)
l[1]
```

```
## South.Carolina
##       11.16079
```

```
l[2]
```

```
## Stanford
## 6.447113
```

```
l[3]
```

```
##    UConn
## 6.416081
```

```
l[4]
```

```
##      LSU
## 6.373434
```

```
l[5]
```

```
##     Utah
## 6.265124
```

```
l[6]
```

```
## Virginia.Tech
##       5.95271
```

```
l[7]
```

```
##  Indiana
## 5.794937
```

```
l[8]
```

```
##     Duke
## 5.581368
```

```
l[9]
```

```
## Notre.Dame
##   5.542085
```

```
l[10]
```

```
##     UCLA
## 5.492398
```

```
m[1]
```

```
##  Hartford
## -9.352013
```

```
m[2]
```

```
## Saint.Peter.s
##     -7.930054
```

```
m[3]
```

```
## St..Xavier
##  -4.622275
```

```
m[4]
```

```
##      Navy
## -4.397604
```

```
m[5]
```

```
## Mississippi.Valley.State
##                -4.197128
```

The top 10 strongest team: South.Carolina, Stanford, UConn, LSU, Utah, Virginia.Tech, Indiana, Duke, Notre.Dame, UCLA.
The bottom 5 weakest team: Hartford, Saint.Peter.s, St..Xavier, Navy, Mississippi.Valley.State.

#6. #ask about how to make the odd: Michigan vs Delaware.State, #Michigan vs St..Francis..PA, Michigan vs Western.Michigan #(opponents of Michigan in the first two rounds of tournament)

```
#michigan_odds <- summary(btmod4)$coefficients["Michigan", 1]
D_odds <- summary(btmod4)$coefficients["Delaware.State", 1]
S_odds <- summary(btmod4)$coefficients["St..Francis..PA.", 1]
W_odds <- summary(btmod4)$coefficients["Western.Michigan", 1]
exp(0 -D_odds)
```

```
## [1] 8.04438
```

```
exp(0 - S_odds)
```

```
## [1] 23.85012
```

```
exp(0 - W_odds)
```

```
## [1] 0.764609
```

The odd of Michigan winning Delaware.State: 8.04438 The odd of Michigan winning St..Francis..PA.: 23.85012 The odd of Michigan winning Western.Michigan: 0.764609

#7 part a:

```
D_se <- summary(btmod4)$coefficients["Delaware.State", 2]
S_se <- summary(btmod4)$coefficients["St..Francis..PA.", 2]
W_se <- summary(btmod4)$coefficients["Western.Michigan", 2]
D_se
```

```
## [1] 0.6724356
```

```
S_se
```

```
## [1] 0.7188341
```

```
W_se
```

```
## [1] 0.6063259
```

Standard error of log_odd of Michigan winning Delaware.State is: 0.6724356 Standard error of log_odd of Michigan winning St..Francis..PA. is: 0.7188341 Standard error of log_odd of Michigan winning Western.Michigan is: 0.6063259

```
z_score <- D_odds / D_se
p_value1 <-  (1 - pnorm(abs(z_score)))
p_value1
```

```
## [1] 0.0009655483
```

```
z_score <- S_odds / S_se
p_value2 <-  (1 - pnorm(abs(z_score)))
p_value2
```

```
## [1] 5.111362e-06
```

```
z_score <- W_odds / W_se
p_value3 <-  (1 - pnorm(abs(z_score)))
p_value3
```

```
## [1] 0.3290092
```

```
p_values <- c(p_value1, p_value2, p_value3)
adjusted_p <- p.adjust(p_values, method = "bonferroni")
adjusted_p
```

```
## [1] 2.896645e-03 1.533409e-05 9.870276e-01
```

self-defined part for paper: Check for multicolinearity in btmod4

To check multicolinearity in bemod4 to test one of assumptions in B-T model (essentially a logistic regression model)

```
vif(btmod4)
```

```
##                    Adams.State           Agnes.Scott.College
##                       1.026927                      1.028778
##                     Air.Force                         Akron
##                       1.535084                      1.625156
##                       Alabama                   Alabama.A.M
##                       1.559726                      1.789681
##                  Alabama.State              Alaska.Anchorage
##                       1.764706                      1.032574
##                         Albany                  Alcorn.State
##                       1.667724                      1.705359
##            American.University            Appalachian.State
##                       1.672594                      1.589301
##                        Arizona                 Arizona.State
##                       1.753412                      1.465626
##                    Ark.Baptist                      Arkansas
##                       1.048416                      1.606900
##                  Arkansas.State          Arkansas.Pine.Bluff
##                       1.578482                      1.790169
##               Arlington.Baptist                          Army
##                       1.042206                      1.815887
##                         Auburn                   Austin.Peay
##                       1.488246                      1.532078
##                            BYU                    Ball.State
##                       1.761589                      1.566339
```

```
##                     Baylor                  Bellarmine
##                   1.836206                    1.515023
##         Bellevue.University                     Belmont
##                   1.039453                    1.660474
##            Bethune.Cookman                  Binghamton
##                   1.738662                    1.768277
##       Bloomsburg.University              Bluefield.State
##                   1.047577                    1.052534
##                Boise.State               Boston.College
##                   1.652761                    1.607769
##          Boston.University               Bowling.Green
##                   1.586810                    1.528807
##                    Bradley                     Brescia
##                   1.301567                    1.046168
##                      Brown                      Bryant
##                   1.404409                    1.720550
##                   Bucknell                     Buffalo
##                   1.786988                    1.594939
##                     Butler                    Cal.Poly
##                   1.555437                    1.774041
##       Cal.State.Bakersfield         Cal.State.Fullerton
##                   1.813165                    1.915405
##        Cal.State.Northridge                  California
##                   1.750403                    1.493054
##          California.Baptist                    Campbell
##                   1.695112                    1.863602
##                   Canisius              Central.Arkansas
##                   1.740483                    1.446222
##         Central.Connecticut              Central.Michigan
##                   1.690605                    1.491850
##  Champion.Christian.College                   Charleston
##                   1.046308                    1.595877
##         Charleston.Southern                   Charlotte
##                   1.608026                    1.754013
##                Chattanooga               Chicago.State
##                   1.676349                    1.449451
##          Christian.Brothers                   Cincinnati
##                   1.022170                    1.522119
##                    Clemson               Cleveland.State
##                   1.680903                    1.391597
##            Coastal.Carolina                     Colgate
##                   1.566050                    1.862363
##                   Colorado               Colorado.State
##                   1.819553                    1.596184
##                   Columbia              Converse.College
##                   1.408327                    1.050449
##                Coppin.State                     Cornell
##                   1.571055                    1.422466
##                  Creighton                    Dartmouth
##                   1.631592                    1.241491
##                   Davidson                      Dayton
##                   1.567279                    1.425687
##                     DePaul                    Delaware
##                   1.651516                    1.711569
```

```
##               Delaware.State                      Denver
##                     1.614322                    1.755750
##               Detroit.Mercy                       Drake
##                     1.475037                    1.650070
##                    Drexel                         Duke
##                     1.670897                    1.628771
##                    Duquesne                ELIZABETH.CITY
##                     1.588420                    1.031737
##               East.Carolina          East.Tennessee.State
##                     1.654905                    1.595258
##            Eastern.Illinois             Eastern.Kentucky
##                     1.649971                    1.616679
##            Eastern.Michigan            Eastern.Washington
##                     1.659886                    1.722699
##                        Elon                Emory...Henry
##                     1.661869                    1.043802
##                     Erskine                   Evansville
##                     1.039432                    1.609025
##            Evergreen.State                     Fairfield
##                     1.023693                    1.798887
##          Fairleigh.Dickinson                         Fisk
##                     1.423681                    1.031829
##                     Florida                  Florida.A.M
##                     1.529103                    1.520307
##            Florida.Atlantic            Florida.Gulf.Coast
##                     1.685544                    1.205435
##         Florida.International            Florida.National
##                     1.733342                    1.027681
##               Florida.State                      Fordham
##                     1.631057                    1.535863
##               Fresno.State                       Furman
##                     1.600962                    1.526456
##               Gardner.Webb                  George.Mason
##                     1.215442                    1.557871
##           George.Washington                   Georgetown
##                     1.543332                    1.605014
##                     Georgia              Georgia.Southern
##                     1.561055                    1.542645
##               Georgia.State                  Georgia.Tech
##                     1.587760                    1.676830
##                     Gonzaga                    Grambling
##                     1.296082                    1.694112
##               Grand.Canyon                     Green.Bay
##                     1.613849                    1.475286
##                     Hampton                     Hartford
##                     1.668624                    1.030984
##                     Harvard                      Hawai.i
##                     1.501068                    1.851375
##            Hendrix.College                   High.Point
##                     1.040898                    1.694528
##                     Hofstra                   Holy.Cross
##                     1.671818                    1.656230
##                     Houston             Houston.Christian
##                     1.712441                    1.756527
```

```
##                      Howard                 Howard.Payne
##                    1.615820                     1.007522
##                       IUPUI                        Idaho
##                    1.713776                     1.701902
##                Idaho.State                     Illinois
##                    1.652946                     1.365361
##             Illinois.State                Incarnate.Word
##                    1.564887                     1.759537
##                     Indiana                Indiana.State
##                    1.406796                     1.592971
##                        Iona                         Iowa
##                    1.566899                     1.582826
##                 Iowa.State                Jackson.State
##                    1.875072                     1.276591
##               Jacksonville           Jacksonville.State
##                    1.514809                     1.592689
##             James.Madison              Jarvis.Christian
##                    1.544094                     1.035958
##       Johnson...Wales..NC.               Johnson.C..Smith
##                    1.052310                     1.035839
##     Johnson.University..FL.                       Kansas
##                    1.039703                     1.778809
##                Kansas.City                  Kansas.State
##                    1.696484                     1.837494
##             Kennesaw.State                    Kent.State
##                    1.612229                     1.534424
##                    Kentucky                          LSU
##                    1.459437                     1.161038
##             LSU.Alexandria                LSU.Shreveport
##                    1.020867                     1.026851
##                    La.Salle           La.Sierra.University
##                    1.603114                     1.006154
##                    La.Verne              LaGrange.College
##                    1.022062                     1.027703
##                   Lafayette                        Lamar
##                    1.776602                     1.757582
##                      Lehigh           Lenoir.Rhyne.College
##                    1.793629                     1.058881
##                     Liberty                Life.University
##                    1.337839                     1.021311
##                  Lindenwood                     Lipscomb
##                    1.141240                     1.535709
##                 Little.Rock                Long.Beach.State
##                    1.563184                     1.681287
##      Long.Island.University                     Longwood
##                    1.786353                     1.801694
##                   Louisiana              Louisiana.College
##                    1.566855                     1.051294
##             Louisiana.Tech                     Louisville
##                    1.750488                     1.703761
##             Loyola.Chicago                Loyola.Maryland
##                    1.355090                     1.772965
##           Loyola.Marymount                        Maine
##                    1.536402                     1.615633
```

```
##                   Manhattan                     Marist
##                    1.834903                   1.802911
##                   Marquette                   Marshall
##                    1.684923                   1.687050
##                    Maryland      Maryland.Eastern.Shore
##                    1.484294                   1.558360
##                     McNeese                    Memphis
##                    1.768013                   1.567320
##                      Mercer                   Merrimack
##                    1.556542                   1.905394
##                        Mesa                      Miami
##                    1.015668                   1.663048
##                  Miami..OH.              Michigan.State
##                    1.627638                   1.360990
##           Middle.Tennessee                      Miles
##                    1.383075                   1.024218
##                   Milwaukee                  Minnesota
##                    1.760815                   1.432774
##           Mississippi.State    Mississippi.Valley.State
##                    1.520706                   1.422719
##                   Missouri              Missouri.State
##                    1.637501                   1.628975
##          Missouri.St..Louis             Mitchell.College
##                    1.014260                   1.071820
##                      Mobile                   Monmouth
##                    1.072624                   1.803201
##                    Montana                Montana.State
##                    1.748098                   1.673611
##                Montana.Tech             Montreat.College
##                    1.027390                   1.056355
##             Morehead.State                 Morgan.State
##                    1.658158                   1.456012
##             Mount.St..Mary.s                Murray.State
##                    1.785245                   1.616579
##                   NC.State                  NC.Wesleyan
##                    1.701927                   1.047146
##                       NJIT                 NM.Highlands
##                    1.829274                   1.013591
##                       Navy                     Nebraska
##                    1.157685                   1.399370
##                     Nevada                New.Hampshire
##                    1.606423                   1.722577
##              New.Jersey.City                  New.Mexico
##                    1.056602                   1.589622
##             New.Mexico.State                 New.Orleans
##                    1.713692                   1.686149
##                    Niagara                    Nicholls
##                    1.847602                   1.535950
##                Norfolk.State                North.Alabama
##                    1.402913                   1.560595
##                North.Carolina           North.Carolina.A.T
##                    1.667919                   1.645540
##        North.Carolina.Central                 North.Dakota
##                    1.673305                   1.673897
```

```
##             North.Dakota.State                   North.Florida
##                       1.649641                        1.464986
##                    North.Texas                    Northeastern
##                       1.770220                        1.718552
##                Northern.Arizona                Northern.Colorado
##                       1.782359                        1.724280
##                Northern.Illinois                    Northern.Iowa
##                       1.609794                        1.570344
##                Northern.Kentucky              Northern.New.Mexico
##                       1.767954                        1.021802
##                     Northwestern              Northwestern.State
##                       1.386818                        1.732323
##                      Notre.Dame                         Oakland
##                       1.519851                        1.763458
##                           Ohio                      Ohio.State
##                       1.466210                        1.514041
##                       Oklahoma                  Oklahoma.State
##                       1.696758                        1.896039
##                   Old.Dominion                       Ole.Miss
##                       1.659977                        1.443525
##                          Omaha                    Oral.Roberts
##                       1.764662                        1.693102
##                         Oregon                    Oregon.State
##                       1.854579                        1.658876
##                        Pacific Palm.Beach.Atlantic.University
##                       1.850819                        1.020953
##          Park.University.Gilbert                      Penn.State
##                       1.052886                        1.448306
##                   Pennsylvania                      Pepperdine
##                       1.427859                        1.745994
##                       Pfeiffer                      Pittsburgh
##                       1.051748                        1.458267
##             Pittsburgh.Johnstown                      Point.Park
##                       1.051539                        1.015106
##                       Portland                  Portland.State
##                       1.491613                        1.774941
##                Prairie.View.A.M                   Presbyterian
##                       1.675240                        1.854460
##                      Princeton                     Providence
##                       1.370634                        1.520214
##                   ProvidenceMT                        Purdue
##                       1.015356                        1.399809
##                Purdue.Fort.Wayne                Queens.University
##                       1.766961                        1.488831
##                      Quinnipiac                       Radford
##                       1.606584                        1.886960
##                Regis.University                    Rhode.Island
##                       1.009332                        1.431764
##                           Rice                        Richmond
##                       1.646517                        1.472637
##                          Rider                   Robert.Morris
##                       1.747670                        1.681764
##                        Rutgers                   SE.Louisiana
##                       1.369375                        1.589309
```

```
##              SIU.Edwardsville                         SMU
##                      1.664488                    1.625826
##              Sacramento.State                 Sacred.Heart
##                      1.684403                    1.737152
##                Saint.Joseph.s                  Saint.Louis
##                      1.570233                    1.619209
##                   Saint.Mary.s                Saint.Peter.s
##                      1.809331                    1.007129
##                   Sam.Houston                     Samford
##                      1.516689                    1.609975
##                     San.Diego              San.Diego.State
##                      1.717644                    1.566063
##                  San.Francisco        San.Jos.U.00E9..State
##                      1.728576                    1.562188
##                   Santa.Clara                   Seattle.U
##                      1.788559                    1.536275
##                    Seton.Hall                       Siena
##                      1.646680                    1.884964
##                  South.Alabama               South.Carolina
##                      1.442990                    1.009425
##           South.Carolina.State       South.Carolina.Upstate
##                      1.379727                    1.906348
##                  South.Dakota           South.Dakota.State
##                      1.767967                    1.239843
##                  South.Florida     Southeast.Missouri.State
##                      1.378709                    1.676841
##                      Southern            Southern.Illinois
##                      1.704193                    1.570253
##              Southern.Indiana                Southern.Miss
##                      1.643546                    1.577381
##                 Southern.Utah            Southern.Wesleyan
##                      1.467850                    1.055357
##                   Spring.Hill  St.Thomas.University.Houston
##                      1.049478                    1.042095
##                   St..Andrews              St..Bonaventure
##                      1.055517                    1.453397
##              St..Francis..PA.         St..Francis.Brooklyn
##                      1.784863                    1.916189
##                    St..John.s       St..Thomas...Minnesota
##                      1.613666                    1.709050
##                    St..Xavier                    Stanford
##                      1.391803                    1.642073
##              Stanislaus.State            Stephen.F..Austin
##                      1.029145                    1.453634
##                     Stetson                    Stonehill
##                      1.612331                    1.843229
##                   Stony.Brook                    Syracuse
##                      1.691666                    1.583087
##                           TCU                    Tarleton
##                      1.362638                    1.485701
##                     Temple                    Tennessee
##                      1.646786                    1.556212
##               Tennessee.State               Tennessee.Tech
##                      1.716321                    1.625948
```

```
##             Tennessee.Wesleyan                           Texas
##                       1.014266                        1.885135
##                      Texas.A.M              Texas.A.M.Commerce
##                       1.503094                        1.823540
##       Texas.A.M.Corpus.Christi             Texas.A.M.Texarkana
##                       1.689535                        1.025647
##                  Texas.Lutheran                  Texas.Southern
##                       1.037863                        1.381451
##                     Texas.State                     Texas.Tech
##                       1.599983                        1.749188
##                         Toledo                        Tougaloo
##                       1.436624                        1.047154
##                         Towson                 Trevecca.Nazarene
##                       1.693810                        1.033123
##                            Troy                          Tulane
##                       1.575286                        1.613451
##                           Tulsa                             UAB
##                       1.611989                        1.732104
##                        UC.Davis                       UC.Irvine
##                       1.861429                        1.567302
##                    UC.Riverside                    UC.San.Diego
##                       1.583012                        1.822117
##                  UC.Santa.Barbara                             UCF
##                       1.915302                        1.546583
##                            UCLA                           UConn
##                       1.858074                        1.505195
##                             UIC                       UL.Monroe
##                       1.630542                        1.446146
##                             UMBC                           UMass
##                       1.766779                        1.435268
##                     UMass.Lowell                  UNC.Asheville
##                       1.576014                        1.898916
##                   UNC.Greensboro                  UNC.Wilmington
##                       1.637981                        1.387401
##                             UNLV                      UNT.Dallas
##                       1.229951                        1.039461
##                             USC                     UT.Arlington
##                       1.754194                        1.573161
##                        UT.Martin               UT.Rio.Grande.Valley
##                       1.725743                        1.618577
##                             UTEP                            UTSA
##                       1.733000                        1.819460
##    University.of.the.Southwest                              Utah
##                       1.040460                        1.612990
##                       Utah.State                      Utah.Tech
##                       1.354991                        1.575867
##                      Utah.Valley                             VCU
##                       1.474474                        1.455639
##                  VU.of.Lynchburg                      Valparaiso
##                       1.035411                        1.519917
##                      Vanderbilt                         Vermont
##                       1.466863                        1.511708
##                      Villanova                        Virginia
##                       1.611215                        1.525990
```

```
##          Virginia.Tech                    Wagner
##               1.515740                  1.902760
##             Wake.Forest          Warner.University
##               1.706963                  1.052428
##               Washington          Washington.State
##               1.688854                  1.953878
##             Weber.State             West.Virginia
##               1.397815                  1.827610
##        Western.Carolina          Western.Colorado
##               1.559152                  1.019015
##        Western.Illinois          Western.Kentucky
##               1.650892                  1.781067
##        Western.Michigan             Wichita.State
##               1.526547                  1.725637
##   Wilberforce.University            William...Mary
##               1.031387                  1.730547
##          Wilmington..DE.                  Winthrop
##               1.058626                  1.830534
##                Wisconsin                  Wofford
##               1.419600                  1.559778
##             Wright.State                   Wyoming
##               1.678941                  1.561603
##                 Xavier                      Yale
##               1.381818                  1.499915
##        Youngstown.State
##               1.699058
```

Team strength estimation based on Bradley-Terry model with data from NCAA women's basketball

Haoran Cheng

April 2023

1 Introduction

NCAA Basketball tournament is appealing to millions of sports fans all over the world, making it valuable to explore winning an loss probability of different teams. In this paper, relative strength score of different teams will be predicted with Bradley-Terry model of logistic regression, and the hypothesis that Michigan is the stronger team than its opponents in the first 2 rounds of tournament will be tested. To explain data source and their reliability, it is win-loss result collected for Division 1 contests during the regular season from NCAA women's basketball teams for the 2022-23 season. It contains 430 columns, with the first "home-win" noting whether the home team won in the game, the second "date" noting the specific date of gaming happening in form "year-month-day", and the remaining 428 columns recording win/loss result of those basketball teams. Home-win of 1 represents the win of home team and 0 otherwise. Specific team data entry equal to 1 represents the win of that team in that specific data and 0 otherwise. It's helpful to clarify some default setting for those data studied here. First, there is no tie in each games. Second, team strength is assumed to change over the course of the season, which inspires to incorporate the date variable interaction in the model. Other special terms worth explanation for readers will be addressed in latter paragraphs once they are used.

To generally introduce the method used, a total of three models are fitted, with all of them fitted in two versions: with penalty and without penalty. One of these three models excludes both home-advantage variable and date interaction. One of them excludes date interaction but includes home-advantage variable. The other one excludes home-advantage variable but includes date interaction. The best of those models are picked according to cross validation with negative log likelihood loss plus Ridge penalty, and the strength of one specific basketball team - Michigan is studies against its opponents in the first two tournament rounds with multiplicity correction.

Lastly, to briefly summarize the result generated, mod4a - the model incorporating date interaction without penalty achieves the lowest value of objective function and is therefore chosen as the best model to do further strength evaluation. South.Carolina ranks the highest in mod4a coefficient, as well as its estimated strength score comparing with the others. Hartford ranks the lowest in mod4a coefficient and its estimated strength. Based on the fact that all of the three odd computations involving Michigan are over 0, Michigan is predicted to have higher strength than two of its three opponents in the first two rounds of tournament-Delaware.State and St..Francis..PA.. Also, multiplicity corrected hypothesis rejects two of the null hypothesis that Michigan is not stronger than the rival, but accepts the one with rival to be Western.Michigan.

2 Methodology

It's useful to explain some professional statistical terms used next. First, Bradley-Terry model defines a probability model based on logistic regression, which is able to predict the outcome of paired comparison. Given a pair of individuals i and j drawn from some population, the probability of i winning over j is: $P(i > j) = \frac{P_i}{P_i + P_j}$. Notably, if β_i defines the coefficient of linear predictor within logistic regression, exponential score function is used to parameterize:

$P_i = e^{\beta_i}$. Second, a penalized form of logistic regression, specifically with L2 penalty equal to square of the magnitude of coefficients in this paper, defines the adjusted version of logistic regression compromising the goal of loss minimization with overfitting avoid. It works by adding one more penalized term measuring the magnitude of coefficients to the objective function besides loss calculation.

Third, negative log likelihood loss, which is more efficient in model accuracy evaluation than mean-square loss in case of involving categorical variables, defines the model loss computation as follows:

$L = -\frac{1}{m}\Sigma_{i=1}^{m} y_i log(\hat{y_i})$

Forth, Bonferroni correction defines the method to adjust P values when several dependent or independent hypothesis tests are performed simultaneously on the same data set, which is one specific method of multiplicity correction.

Next, assumptions of Bradley-Terry model, which is essentially a logistic regression model, will be specified. First, it's assumed the model is parameterized exponentially, and therefore $P_i = e^{\beta_i}$ if P_i denotes the team strength score of team i. Second, it's assumed that the probability of pairwise comparison $i > j$ is $P(i > j) = \frac{P_i}{P_i + P_j}$. Third, it's assumed that the dependent variable is binary, which corresponds to the binary value of home-win variable as the response variable. Forth, it's assumed that observations of different games are independent. Possible flaws of this assumption would be discussed further. Forth, it's assumed no multicollinearity among the independent variables, which means there is assumed no linear relationship between team variables.

Now details of methods used will be explained. A total of six models-with three different models fitted in both with and without penalty version, are fitted using the same maximum likelihood method. They are determined through different model disciplines: with or without home advantage variable, with or without date interaction term. Their negative log likelihood loss plus Ridge penalty are computed and compared to rank their performance. Notably, Michigan has be excluded from all the response variables below since it works as reference in the

model. The first model mod1 defined below is fitted without home advantage variable, without date interaction and without penalties:

$homewin = \frac{1}{1+e^{\beta^T(team1,team2,\ldots)+\epsilon}}$

$glm(homewin\ .-date-Michigan-1, family = binomial)$ mod2 is defined below with the same variable as mod1, without home advantage variable, without date interaction but with penalties:

$homewin = \frac{1}{1+e^{\beta^T(team1,team2,\ldots)+\epsilon}}$

$arm :: bayesglm(homewin\ .-date-Michigan-1, family = binomial)$

Model mod3 is fitted with home advantage variable and without date interaction. Both its version with and without penalties are defined:

mod3: $homewin = \frac{1}{1+e^{\beta^T(team1,team2,\ldots)+\beta_0+\epsilon}}$

mod3a without penalty:

$glm(homewin\ .-date-Michigan, family = binomial)$

mod3b with penalty:

$arm :: bayesglm(homewin\ .-date-Michigan, family = binomial)$

Model mod4 defines below is fitted without home advantage variable and with date interaction, Both its version with and without penalties are defined:

mod4: $homewin = \frac{1}{1+e^{\beta^T((date,team1,team2,\ldots),team1,team2,\ldots)+\epsilon}}$

mod4a without penalty:

$glm(homewin\ .-Michigan-date-1+(.-date-Michigan-1)*date, family = binomial)$

mod4b with penalty:

$arm :: bayesglm(homewin\ .-Michigan-date-1+(.-date-Michigan-1)*date, family = binomial)$

Maximum likelihood estimation is performed in all of the no-penalty models to determine coefficients. Intuitively, MLE defines the method to determine regression coefficients with assumed models by maximizing likelihood function with observed data. To apply to the case in this paper, the likelihood function is : $L(x1, x2, x3, \ldots|\beta_1, \beta_2, \ldots) = f(x1, x2, x3, \ldots|\beta_1, \beta_2, \ldots) = \prod_{i=1}^{n} f(x_i|\beta_1, \beta_2, \ldots) = \prod_{i=1}^{n} \frac{1}{1+e^{w^T x_i+b}}$ in the case of Bradley-Terry model with logistic regression

essence. The vector of coefficient values are determined to be the value making derivative of the log of $L(x1, x2, x3, \ldots | \beta_1, \beta_2, \ldots)$ equal to zero.

How Bayesian prior is incorporated to perform penalty also deserves specification here. The default prior proportional to the reciprocal of the variance is used besides maximum likelihood estimation to determine regression coefficients here, and the prior is equivalent to adding a Ridge penalty term to the objective function. One of advantages of using penalized model is to prevent overfitting by taking magnitude of coefficients into consideration, making balance between loss minimization and the complexity of model coefficients therefore achieved. Objective function of Negative Log likelihood loss plus Ridge penalty is applied to rank those models. Specifically, objective $= -\sum_1^n y_i log\hat{y}_i + \lambda \sum_{i=1}^p \beta_i^2$, which constrains the magnitude of coefficients as well as forcing the model to minimize NLL loss. The benefit is to prevent overfitting, in which case extra coefficients are determined when the training process tries to minimize loss of training data, while making the model less capable to deal with new test data because extra coefficients are fitted specifically suitable for the training data.

After picking the best model, relative strength of those teams are determined according to the value of their regression coefficients in the model, with odds of Michigan winning three example teams in the first two rounds are computed to specifically test relative strength of those four teams. Odd of example team i winning example team j is computed through those steps: First, log odd of team i winning team j is determined by $log-odd_(i > j) = log\frac{P(i>j)}{P(j>i} = log\frac{e^{\beta_i}}{e^{\beta_j}} = \beta_i - \beta_j$, given the strength of team i is assumed to be exponential score with formula $P_i = e^{\beta_i}$. The exponent of the difference $e^{\beta_i - \beta_j}$ is further calculated, which directs to the final value of the odd of team i winning team j: $odd = e^{\beta_i - \beta_j}$. It's value exceeding 1 would indicate that Michigan winning probability is more than loss.

Lastly, three hypothesis tests based on three pairs of Michigan-rival comparison are performed to test the null hypothesis that Michigan was no better than the rival. multiplicity correction that controls the family-wise type 1 error rate is

conducted with Bonferroni method. Specifically, denote H1, H2, H3 to be three hypothesis test of Michigan against its three rivals:

H1: null hypothesis: Michigan was no better than Delaware.State. Alternative: Michigan was stronger.

H2: null hypothesis: Michigan was no better than St..Francis...PA.. Alternative: Michigan was stronger.

H3: null hypothesis: Michigan was no better than Western.Michigan. Alternative: Michigan was stronger.

Assume their respective solo p-value are computed to be: p1, p2 and p3. Assume the significant level $\alpha = 0.05$. The Bonferroni correction will reject the null hypothesis for each hypothesis if: $pi \leq \frac{\alpha}{3}$. The join family-wise type I error is therefore constrained at significant level of 0.05 as a result. Without which correction, the joint family-wise type I error would be underestimated.

3 Results

	Home advantage	Date interaction	Objective function vale
Mod1	No	No	0.4896
Mod2	No	No	0.4866
Mod3a	Yes	No	0.4875
Mod3b	Yes	No	0.4894
Mod4a	No	Yes	0.2309
Mod4b	No	Yes	0.2310

table1: objective function results of six models and their incorporation

First, according to table 1, mod4a - the version with no penalty ranks the first in objective minimization and has the value to be 0.2309. Mod1 performs the worse and achieves the objective function value to be 0.4896. Regarding the others, mod2 achieves the value of 0.4866, mod3a achieves 0.4875, mod3b achieves 0.4894, and mod4b achieves 0.2310. Accordingly, mod4a is the best one picked to further evaluate strengths of teams.

Second, teams are ranked according to their regression coefficients in mod4a to compare their relative strengths. Accordingly, the top 10 strongest teams with

their coefficients are:

South.Carolina, 11.16; Standford, 6.45; Uconn, 16.42; LSU, 6.37; Utah, 6.27; virginia.tech, 5.95; Indiana, 5.79; Duke, 5.58; Notre.Dame, 5.54; UCLA, 5.49.

The bottom 5 weakest teams with their coefficients are: Hartford, -9.35; Saint.Peter.s, 7.93; St.Xavier, -4.62; Navy, -4.40; Mississippi.Valley.State, -4.20.

The estimated strongest team is South.Carolina with strength score $e^{11.16}$, and the weakest team is predicted to be Hartford with strength score $e^{-9.35}$. Notably, the coefficient of Michigan is 0 since it's reference in the model, and its strength score is $e^0 = 1$.

	Odds
Michigan VS Delaware.State	8.04
Michigan VS St..Francis..PA.	23.85
Michigan VS Western.Michigan	0.76

table3: odds of Michigan vs three opponents

Respectively, Odds of Michigan winning Delaware.State is 8.04. Odds of Michigan winning St..Francis..PA. is 23.85. Odds of Michigan winning Western.Michigan is 0.76. Since odd being over one would indicate the higher probability of Michigan winning, Michigan are predicted to be able to win Delaware.State and St..Francis..PA.. However, Michinga is estimated to lose against Western.Michigan.

	P-value	Adjusted-p-value
H1	0.00097	0.00290
H2	5.111362E-06	1.533409E-05
H3	0.32901	0.98703

table4:Multiplicity corrected hypothesis test with original/adjusted p-value

With H1 and H2 having adjusted p-values smaller than 0.05, it's significant enough to reject the null that Michigan was no better than the rival. However, with H3 having adjusted p-value greater than 0.05, it's accepted that Michinga

was no better than Western.Michigan.

4 Discussion

To summarize the finding in this paper, the model with date interaction but without penalty has the best performance in minimizing objective function, indicating that it performs well not only in minimizing NLL loss but also in balancing the magnitude of coefficients. This model estimates South.Carolina to be the strongest team and Hartford to be the weakest. Michigan is estimated to have larger probability to win against Delaware.State and St..Francis... PA. - two of its rivals in the first two rounds of tournament, given that the odds of Michigan competing those them are greater than 1.

However, two assumptions made before fitting those models can sometimes be unrealistic and make our results unreliable. First, independence between observations of game data can be questionable. The result of Game happening earlier can impact teams playing later, especially when two teams involved are ranking close in score board. For instance, if team Michigan and team Standford ranks first and second in the score board and they are competing for the champion near the end of season, the result of game involving Michigan happening in the same round, but earlier can influence morale of Standford playing later. Second, error-in-variables issue in the original data may distort final estimation. Specifically, two teams playing in the same game may play in neutral third-party site instead of campus stadiums of themselves. In other words, regarding the intercept as home advantage variable is unreasonable because winning of any team involved would not contribute to the home-win response since none of them are the home team. One possible solution could be creating a new indicator variable named non-third-playing: making it return true if third-party site is not involved and false if third-party site is involved. Then replacing the intercept term with this new indicator variable can make the home-advantage analysis more trustworthy.