Introducing Common Techniques to Address Issues within Multiple Logistic Regression Analysis: A National Football League Case Study

Stephen Donald Parziale, M.Sc.
The University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Abbreviated abstract: As the use of predictive modelling spreads throughout sports, researchers are facing many issues in their analyses. In this project, three common issues in binary logistic multiple regression are addressed: longitudinal responses, separation and multicollinearity. Four unique models will be introduced, a baseline Generalized Linear Model that deals with none of these issues, a Hierarchical Generalized Linear Model that accounts for longitudinal responses, a binary logistic regression model with Firth's method to deal with separation issues, and finally a Ridge Regression technique that accounts for multicollinearity. A common dataset with NFL metrics will be used so we can analyze how the results from each model differ while using the same data. With improved understanding of how to address these issues, predictive sports analytics will become more accurate and useful in practice.

Related publications:

1. Guffey, K. (2020). *DEALING WITH SEPARATION ISSUES IN LONGITUDINAL DATA*, The University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Objectives

- Introduce a “baseline” model that estimates a team’s probability of making the playoffs, correcting for no issues. Then introduce 3 concurrent models that account for the issues in our data.
- Using all 4 of our models, compare the estimates
- Understand how each of our predictor variables (3 offensive, 3 defensive) effect the team’s probability of making the playoffs

$$\ln\left[\frac{\pi_{it}}{(1-\pi_{it})}\right] = \beta_0 + \beta_1 PTDFor + \beta_2 RTDFor + \beta_3 TOFor + \beta_4 PTDAgainst + \beta_5 RTDAgainst + \beta_6 TOcreated$$

Baseline logistic regression model

Predictors:

$$\beta 1 = Passing\ Touchdowns$$
$$\beta 2 = Rushing\ Touchdowns$$
$$\beta 3 = Turnovers$$
$$\beta 4 = Passing\ Touchdowns\ Allowed$$
$$\beta 5 = Rushing\ Touchdowns\ Allowed$$
$$\beta 6 = Turnovers\ Created$$

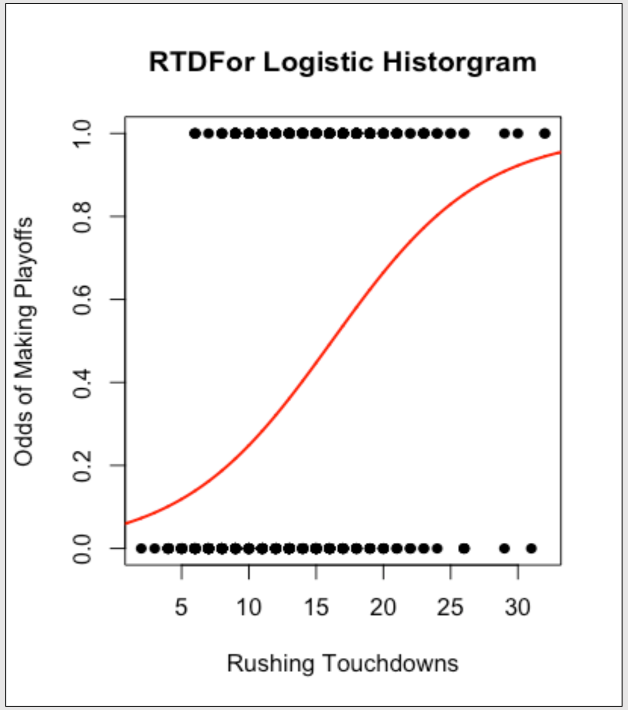
Research Questions

- How can a model be constructed to account for multiple years of data on the same subjects?
- How can a model be constructed to account for the issue of separation in binary multiple logistic regression?
- How can a model be constructed to account for the issue of multicollinearity in binary multiple logistic regression?
- How do the results from each model compare using data from the same dataset?

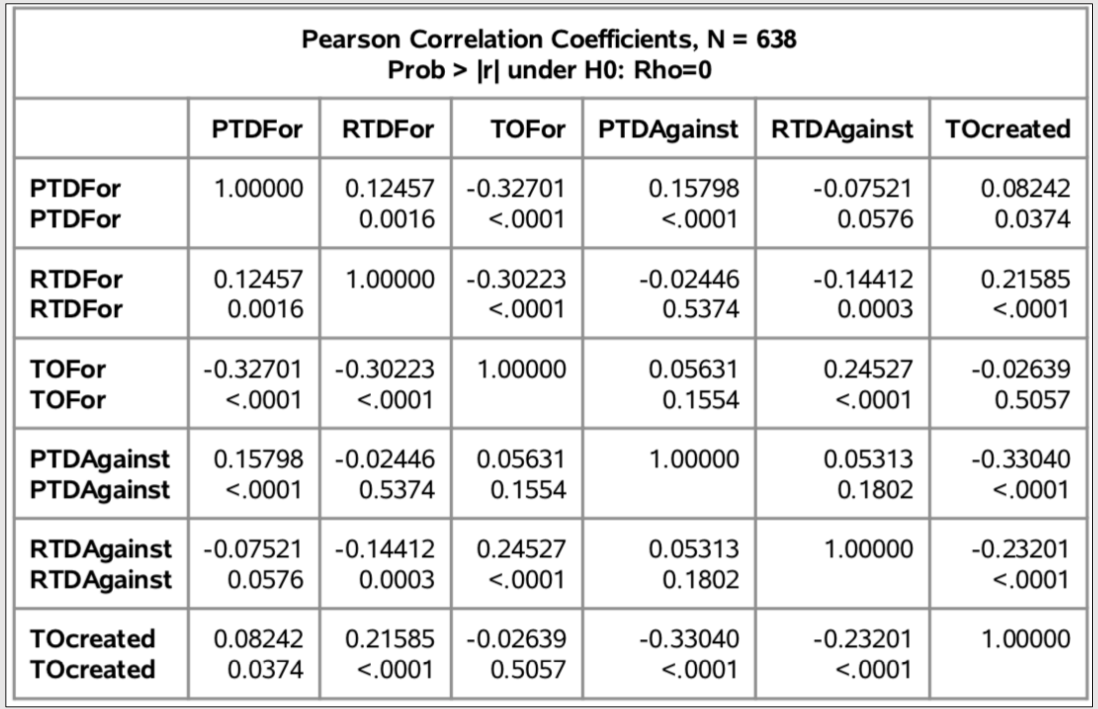
UNIVERSITY OF NORTHERN COLORADO

Methods

- The following R packages were used:
- glm – model that accounts for no issues
- hglm – model that accounts for repeated measures
- logistf – model that accounts for separation issues
- glmnet – model that accounts for multicollinearity



Separation issues



Pearson Correlation Coefficients, N = 638
Prob > |r| under H0: Rho=0

	PTDFor	RTDFor	TOFor	PTDAgainst	RTDAgainst	TOcreated
PTDFor PTDFor	1.00000	0.12457 0.0016	-0.32701 <.0001	0.15798 <.0001	-0.07521 0.0576	0.08242 0.0374
RTDFor RTDFor	0.12457 0.0016	1.00000	-0.30223 <.0001	-0.02446 0.5374	-0.14412 0.0003	0.21585 <.0001
TOFor TOFor	-0.32701 <.0001	-0.30223 <.0001	1.00000	0.05631 0.1554	0.24527 <.0001	-0.02639 0.5057
PTDAgainst PTDAgainst	0.15798 <.0001	-0.02446 0.5374	0.05631 0.1554	1.00000	0.05313 0.1802	-0.33040 <.0001
RTDAgainst RTDAgainst	-0.07521 0.0576	-0.14412 0.0003	0.24527 <.0001	0.05313 0.1802	1.00000	-0.23201 <.0001
TOcreated TOcreated	0.08242 0.0374	0.21585 <.0001	-0.02639 0.5057	-0.33040 <.0001	-0.23201 <.0001	1.00000

Multicollinearity issues

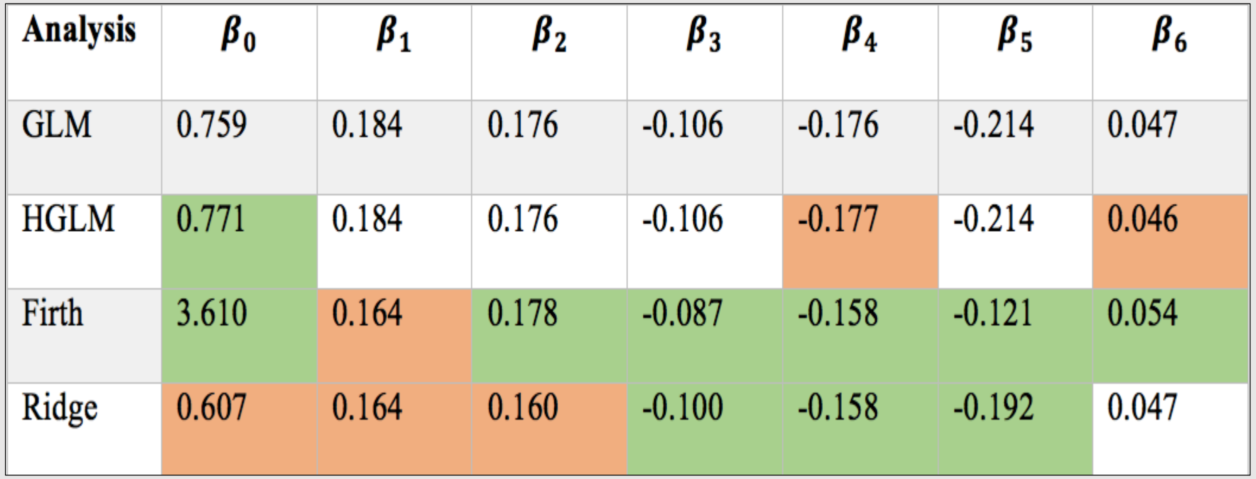
Mixed effects logistic model accounts for repeated measures of individual team (hglm)

Firth’s method is a systematic correction to the score function, which is used to calculate the maximum likelihood estimate (logistf)

Ridge Regression reduces the standard errors by adding a bias penalty to the log-likelihood function (glmnet)

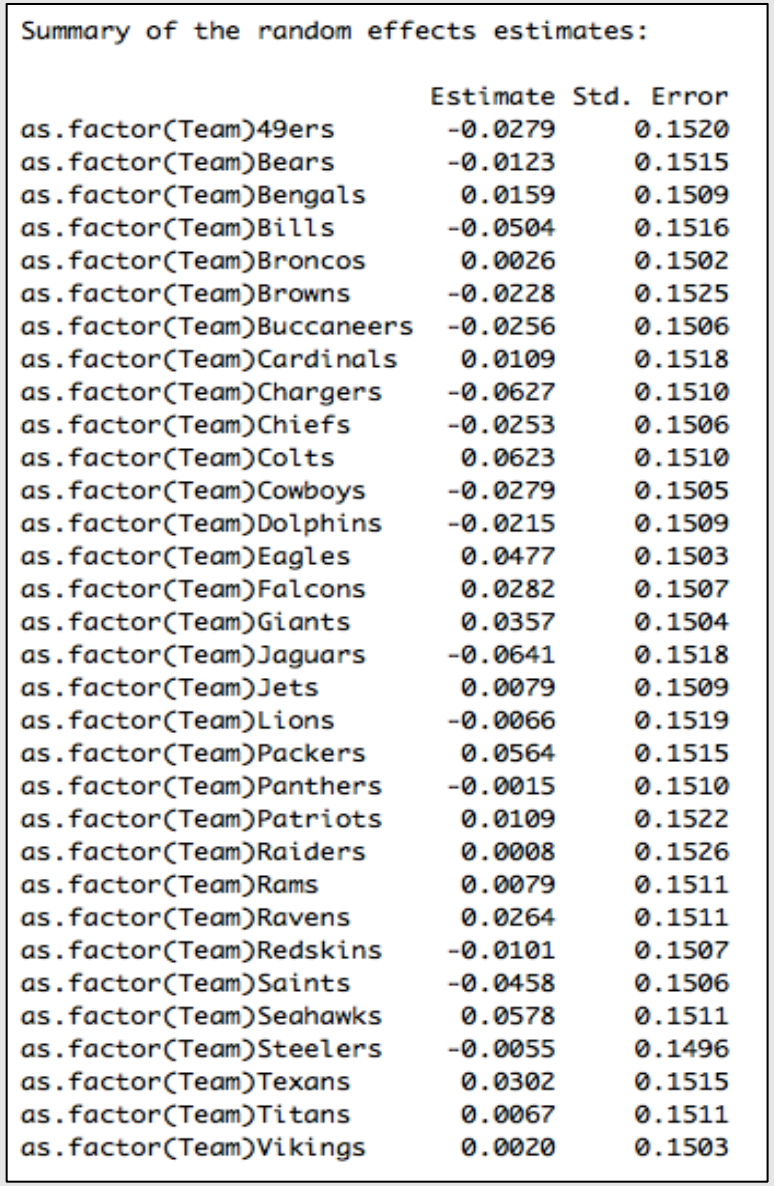
Results

- Minimal differences between our baseline model and our hglm model
- The glmnet model produced the most change, as most estimates shrunk towards 0
- The logistf model significantly changed the intercept (4x increase), but other variables had minimal change
- Every time a team scores an additional rushing or passing touchdown, their odds of making the playoffs increases. When they allow either touchdown, their odds decrease
- Every time a team creates a turnover, their odds increase, but when they commit one their odds decrease



Analysis	β_0	β_1	β_2	β_3	β_4	β_5	β_6
GLM	0.759	0.184	0.176	-0.106	-0.176	-0.214	0.047
HGLM	0.771	0.184	0.176	-0.106	-0.177	-0.214	0.046
Firth	3.610	0.164	0.178	-0.087	-0.158	-0.121	0.054
Ridge	0.607	0.164	0.160	-0.100	-0.158	-0.192	0.047

Model coefficients between analyses



```
Summary of the random effects estimates:

                              Estimate Std. Error
as.factor(Team)49ers           -0.0279     0.1520
as.factor(Team)Bears           -0.0123     0.1515
as.factor(Team)Bengals          0.0159     0.1509
as.factor(Team)Bills           -0.0504     0.1516
as.factor(Team)Broncos          0.0026     0.1502
as.factor(Team)Browns          -0.0228     0.1525
as.factor(Team)Buccaneers      -0.0256     0.1506
as.factor(Team)Cardinals        0.0109     0.1518
as.factor(Team)Chargers        -0.0627     0.1510
as.factor(Team)Chiefs          -0.0253     0.1506
as.factor(Team)Colts            0.0623     0.1510
as.factor(Team)Cowboys         -0.0279     0.1505
as.factor(Team)Dolphins        -0.0215     0.1509
as.factor(Team)Eagles           0.0477     0.1503
as.factor(Team)Falcons          0.0282     0.1507
as.factor(Team)Giants           0.0357     0.1504
as.factor(Team)Jaguars         -0.0641     0.1518
as.factor(Team)Jets             0.0079     0.1509
as.factor(Team)Lions           -0.0066     0.1519
as.factor(Team)Packers          0.0564     0.1515
as.factor(Team)Panthers        -0.0015     0.1510
as.factor(Team)Patriots         0.0109     0.1522
as.factor(Team)Raiders          0.0008     0.1526
as.factor(Team)Rams             0.0079     0.1511
as.factor(Team)Ravens           0.0264     0.1511
as.factor(Team)Redskins        -0.0101     0.1507
as.factor(Team)Saints          -0.0458     0.1506
as.factor(Team)Seahawks         0.0578     0.1511
as.factor(Team)Steelers        -0.0055     0.1496
as.factor(Team)Texans           0.0302     0.1515
as.factor(Team)Titans           0.0067     0.1511
as.factor(Team)Vikings          0.0020     0.1503
```

UNIVERSITY OF NORTHERN COLORADO

UCSAS 2021