Traumatic Brain Injury in Collegiate Sports

Brennan Hall

University of California Santa Barbara

March 22, 2018

Abstract

This paper discusses a dataset with records of the proportion of collegiate student-athletes who had suffered a Traumatic Brain Injury (TBI), segmented according to the student's Gender, Sport that he or she played, and academic Year in which at least one such injury was suffered. The researcher formulates and presents a parametric model of the binary random variable describing whether a student will suffer a TBI based on the data from the sample. The topic of interest is to determine whether the proportion of students who suffer a TBI differs according to a student's Gender, Sport, or the Year in which the student played the Sport. The researcher concludes based on the sample that the Sport is the only statistically significant variable that affects the proportion of students who suffer a TBI. Although critiques and potential further analysis is provided.

1 Motivation

We have been presented a dataset pertaining to cases of traumatic brain injury (TBI) among collegiate athletes. The dataset is a record of the number of students who suffered at least one TBI across different combinations of factors, including *Gender* (males and female), *Sport* (football, hockey, basketball, baseball, and gymnastics), and *Year* (1990, 1991, and 1992). This allows for a total of 30 unique combinations since we treat the academic years as factors rather than a continuous variable. For ease of computations and interpretations, we create an additional column in the dataset that represents the proportion of TBIs (*Proportion*) that occurred per each factor combination. A brief display of the dataset can be found in the Appendix in Table 4.

The goal of our study is to determine if sufficient evidence exists to support the claim that the proportion of students experiencing a TBI differs according to the sport played, gender, or academic year.

2 Exploratory Analysis

Our analysis begins with considering the individual factors against *Proportion*. We do this initially via boxplots as seen in Figure 1. We note the change of proportion per level of *Sport* is clearly the most defined while change per level of *Gender* and *Year* seem mostly negligible, but require further investigation.

Figure 1: Boxplots of *Gender*, *Sport*, and *Year* against *Proportion*

However, continuing this analysis, we consider possible change within joint factors, eg. change of proportion per level of *Sport* among males and females. Figure 2 displays the possible interaction effect as we notice some considerable differences

Figure 2: Boxplot of *Sport* per *Gender*=Female and *Gender*=Male

between the two genders. Conversely, Figure 3 and Figure 4 display a seeming lack of interaction effect between *Gender* and *Year* and *Year* and *Sport*, respectively, as there

Figure 3: Boxplot of *Year* per *Gender*=Female and *Gender*=Male

is relatively little change between the two genders and three years. This indicates to us that the interaction between *Sport* and *Gender* will likely be the only important one to consider for our model.

Figure 4: Boxplot of *Sport* per *Year*=1990, 1991, and 1992

One final note we make about the dataset is that there are zero cases of TBI out of 1627 possible male gymnasts over all three academic years and only 1 case of TBI out of 11,807 possible female gymnasts. This should clearly indicate that it is highly unlikely for gymnasts to suffer a TBI, but we will see some strange results corresponding to the model estimates for this variable due to this nature.

3 Model

In this section we first present to you the final model that we have found best modeled the variable of interest then discuss our model selection methodology later in the section. Our model is constructed under the framework of a General Linear Model (GLM). Specifically, we have decided to treat our variable of interest, the binary variable (TBI) that takes a value of 1 if the student had a TBI and a value of 0 if the student did not have a TBI, as a Binomial random variable. In the GLM framework, we present models by decomposing it into three components: a **random** component, a **systematic** component, and a **link** function that connects the random component with the systematic component by a linear structure. In our case we have chosen to model the variable $Y = \frac{TBI}{m}$, where m denotes the total sample size so that Y is has a Binomial Proportion distribution. Thus our model is presented below.

Random: $Y_i = \frac{TBI_i}{m_i}$, for $i = 1, 2, \ldots, 30$

$Y_i \overset{indep}{\sim} \text{BinProp}(m_i, p_i)$ with known parameters m_i and unknown parameters p_i

$\mathbb{E}[Y_i|\mathbf{X}] = \mathbb{E}\left[\frac{TBI_i}{m_i}\middle|\mathbf{X}\right] = p_i$

Link: $g(p_i) = \eta_i = \text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$

Systematic:

$$\eta_i = \beta_0 + \beta_{\text{Male}} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} + \sum_{j \in \mathcal{J}} \beta_j \mathbb{1}_{[\text{Sport}_i=j]} + \sum_{j \in \mathcal{J}} \beta_{\text{Male}*j} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} \mathbb{1}_{[\text{Sport}_i=j]}$$

$$\mathcal{J} = \{\text{Basketball, Football, Hockey, Gymnatstics}\}$$

where $\mathbb{1}_{[\cdot]}$ denotes the set indicator function and the subscript i denotes the i^{th} observation of the respective variable in our dataset. Each β corresponds to an unknown parameter for which we will attain estimates.

3.1 Methodology

In our search for an effective model that would best accomplish the goal of this study, we considered multiple models that may work but eventually decided on the model presented above. In particular, we want to address three models that we considered where the process we went through in arriving at the third and final model will give insight into our final conclusion. We will denote the following models as Model (1), Model (2), and Model (3), respectively. In each model, our **random** component and **link** function are the same as the final model[1] presented above so here we will only give the **systematic** component to differentiate the models.

$$\eta_i = \beta_0 + \beta_{\text{Male}} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} + \sum_{j \in \mathcal{J}} \beta_j \mathbb{1}_{[\text{Sport}_i=j]} + \sum_{k \in \mathcal{K}} \beta_k \mathbb{1}_{[\text{Year}_i=k]} \tag{1}$$

$$\eta_i = \beta_0 + \beta_{\text{Male}} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} + \sum_{j \in \mathcal{J}} \beta_{\text{Male}*j} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} \mathbb{1}_{[\text{Sport}_i=j]} + \sum_{j \in \mathcal{J}} \beta_j \mathbb{1}_{[\text{Sport}_i=j]} + \sum_{k \in \mathcal{K}} \beta_k \mathbb{1}_{[\text{Year}_i=k]} \tag{2}$$

$$\eta_i = \beta_0 + \beta_{\text{Male}} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} + \sum_{j \in \mathcal{J}} \beta_{\text{Male}*j} \mathbb{1}_{[\text{Gender}_i=\text{Male}]} \mathbb{1}_{[\text{Sport}_i=j]} + \sum_{j \in \mathcal{J}} \beta_j \mathbb{1}_{[\text{Sport}_i=j]} \tag{3}$$

[1]Models with a different random component and/or link function were considered, but each was found to be inferior to the models presented in this paper and so are not addressed in this paper.

where, in each model,

$$\mathcal{J} = \{\text{Basketball, Football, Hockey, Gymnatstics}\}$$
$$\mathcal{K} = \{1991, 1992\}$$

Model (1) seems like a natural place to begin our investigation as it includes every variable of interest but no higher order interaction terms so there is no extra complexity. The summary table for Model (1) gives us a brief look into some of the conclusions we could draw from this model.

The asterisks next to the coefficient estimates in Table 2 indicate that nearly all coefficient estimates are significantly non-zero for Model (1) except for the ones corresponding to *Year*=1991 and *Year*=1992 and *Sport*=Gymnastics. This would mean that all of our variables of interest except *Year* have some non-zero effect on the proportion of students who suffered a TBI.

However, looking at a few diagnostics of this model indicates it is not ideal and can be improved upon. First, the Binomial Proportion model under the GLM framework assumes a *dispersion parameter*, ϕ, equal to 1. The dispersion parameter is part of what describes the variance of a random variable and so is important when determining standard errors of a model's estimates. We can get an estimate for this parameter via the Pearson X^2 statistic,

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)/m_i}$$

Model (1) gives an estimate, $\tilde{\phi} = 1.277$, which is not cause for great concern, but we see later that other models have estimates closer to 1. Noting the interaction elements found in our exploratory data analysis in Section 2, we want to investigate how this impacts our model. Since we are not entirely convinced that *Gender* or *Year* are significant on their own, we do not include an interaction term between these two. Similarly, we are not convinced that any interaction between *Sport* and *Year* is significant so we start with adding the interaction term between *Sport* and *Gender*, bringing us to Model (2). The additional term in Model (2) compared to Model (1) accounts for this interaction. The summary of Model (2) in Table 2 shows us that including this interaction term makes the coefficient estimate for *Gender*=Male insignificant (zero lies in the 95% confidence interval for the true value of the parameter).

We also see some reasons to prefer Model (2) over Model (1) despite the loss of

significance in our parameter estimates. For one, we see that the estimate for the Binomial Proportion distribution's dispersion parameter is 20% closer to 1 than the estimate from Model (1). We also see the AIC value for Model (2) is smaller than Model (1). The AIC value is a measure of quality of a model for a given set of data, relative to other models, where smaller values indicate a higher quality. Finally, an Analysis of Variance (ANOVA) test, presented in Table 1, indicates that Model (2) is significantly different than Model (1) and so, if we reduce our model back to Model (1), we will lose information concerning the proportion of TBI. Because of this result, we conclude that, although the coefficient parameters for *Gender* and the interaction between *Gender* and *Sport* are not significant, their inclusion in the model is important in modeling the proportion of TBI.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Model (1)	22	29.27			
Model (3)	20	22.20	2	7.07	0.0292

Table 1: ANOVA table assuming χ^2 distribution of Deviance

We next consider removing *Year* from our model since it's main effect has been insignificant throughout of our previous two models. Model (3) drops the *Year* term from our model which results in only minor changes to the results of Model (2). It is clear from the summaries in Table 2 that there is very little difference between Model (2) and Model (3). No term has become significant as a result of dropping *Year* from the model so, in terms of the goal of our research, our conclusion will, largely, be the same regardless of which model we choose. We elect to use Model (3) as it has a lower AIC value and model simplicity is preferred.

3.2 Model Estimates and Interpretation

In general because of our choice in the model's link component, each of our model coefficient estimates are interpreted as the additive change in the log-odds ratio due to the corresponding covariate assuming its factor level, holding all other covariates constant. Therefore the exponential of our estimates will correspond to the multiplicative change in the odd's ratio. In Table 3 we present the 95% exponentiated confidence intervals for each of these parameters. These represent the interval that the true value of the respective exponentiated parameter is 95% likely to be inside.

We read this as saying that, for example, 95% of the time being a male student will increase your odds of suffering a TBI somewhere between 0.6171 times and 1.3570

	Model (1)	Model (2)	Model (3)
(Intercept)	-8.029^{***}	-8.129^{***}	-8.044^{***}
	(0.132)	(0.165)	(0.146)
year1991	0.010	0.009	
	(0.110)	(0.110)	
year1992	0.152	0.154	
	(0.102)	(0.102)	
sportBasketB	0.714^{***}	0.876^{***}	0.843^{***}
	(0.136)	(0.189)	(0.188)
sportFootB	1.734^{***}	1.906^{***}	1.882^{***}
	(0.119)	(0.167)	(0.166)
sportGymn	-1.484	-1.277	-1.332
	(1.006)	(1.012)	(1.011)
sportHockey	1.474^{***}	1.226^{***}	1.200^{***}
	(0.152)	(0.245)	(0.245)
sexM	-0.275^{**}	-0.079	-0.090
	(0.085)	(0.201)	(0.200)
sportBasketB:sexM		-0.344	-0.329
		(0.273)	(0.273)
sportFootB:sexM		-0.374	-0.367
		(0.238)	(0.238)
sportGymn:sexM		-15.244	-15.198
		(3328.270)	(3411.516)
sportHockey:sexM		0.389	0.399
		(0.314)	(0.314)
AIC	162.347	160.377	159.278
$\tilde{\phi}$	1.2770	1.0199	1.0117

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Summary of Model (1), Model (2), and Model (3)

	2.5 %	97.5 %
(Intercept)	0.0002	0.0004
genderM	0.6171	1.3570
sportBasketB	1.6152	3.3765
sportFootB	4.7851	9.1946
sportGymn	0.0149	1.2042
sportHockey	2.0292	5.3166
genderM:sportBasketB	0.4201	1.2266
genderM:sportFootB	0.4340	1.1030
genderM:sportGymn	0.0000	0.0000
genderM:sportHockey	0.8104	2.7782

Table 3: 95% confidence interval of exponentiated parameters

times. Therefore, it is likely that being a male student or female student will not contribute to any additional likelihood of suffering a TBI since the multiple of 1 falls in this interval. Considering different sports now, playing football will increase your odds somewhere between 4.7851 times and 9.1946 times. So football, as well as basketball and hockey, have significant chance of increasing your odds of suffering a TBI. On the other hand, being a gymnast seems to possibly lead to a decrease in your odds of suffering from a TBI. The fact that we were able to come to a model that excluded the covariate *Year* implies that it has a very short confidence interval that includes the value of 1. And, in fact, computing the appropriate intervals under Model (2) gives us the intervals $(0.8130, 1.2527)$ and $(0.9544, 1.4267)$ for the parameters $Year = 1991$ and $Year = 1992$, respectively.

4 Concluding Thoughts

Our analysis leads us to the conclusion that *Sport* is the only highly significant variable that influences the proportion of students that suffer a traumatic brain injury. The gender of the student seemed likely to be significant as well, but after we controlled for the sport that the particular gendered student played via the interaction term, we could not conclude that *Gender* had any significant affect on the proportion of TBIs.

However, there are many areas in which we think our analysis could be improved upon given certain conditions. For one, the dataset provided was quite limited in the number of unique combinations of factors. We would like to have considered a study that lasted for many more years, possibly allowing us to consider *Year* as a continuous variable and incorporate time series analysis. If the number of academic years sampled

was increased then the additional data points would give us more stability in our parameter estimates. The Binomial Proportion GLM model's parametric inferences rely on an asymptotically Gaussian distribution, but in our model's current state, the limited size of dataset does not allow us to confidently assume that our parameters' distribution converges. Therefore we must enter a disclaimer on the above confidence intervals that their computation may not be theoretically justified and our parameter estimates may not be unbiased.

We also would like to continue our analysis with a more replicated dataset across multiple universities. The non-replicated dataset we used for this analysis can lead to a biased good fit, meaning our model's results may misleadingly indicate a good fit when, in reality, it poorly estimates the true parameters and mean function. The replicated data would fix this issue, or we could extend our model to consider regional differences in the proportion of students who suffer TBIs.

Additionally, there are certain aspects of the data left unknown to us that we may have found useful to account for, such as whether there were students who suffered multiple TBIs. Knowing not only the number of students who suffered a TBI but also the total number of TBIs suffered per year, sport, and gender would give us an additional control variable and perhaps more insight into whether certain sports, genders, or years were more or less likely to have repeated TBIs.

5 Appendix

	Gender	Sport	Year	Proportion	TotalStudents
1	F	FootB	1990	0.00204	24981
2	F	FootB	1991	0.00205	22934
3	F	FootB	1992	0.00221	27167
4	F	Hockey	1990	0.00137	8762
5	F	Hockey	1991	0.00098	7122
6	F	Hockey	1992	0.00082	8531
⋮	⋮	⋮	⋮	⋮	⋮

Table 4: First 6 lines of the dataset