What Causes Concussions? A Look into the Relationships between Concussion, Gender, Sport, and Year

Chirag Kashyap

STA 138 Project 1

998388067

Introduction

The possibility of suffering a concussion is a risk players take when participating in contact sports. Concussions can be a harrowing experience to go through, because they impact the control of one's body and can even have long-lasting effects on movement, learning, speaking, and temperament. This paper does not try to solve the concussion problem. Armed with a categorical dataset, this paper looks into the independence between concussions, gender, type of sport, and year. By comparing multiple log linear models using Akaike's Information Criterion (AIC), this paper hopes to find the best fitted model for the data along with the conditional odds ratio for that model.

Material and Methods

Data Set: http://www.stat.ufl.edu/~winner/data/concussion.txt
Data Description: http://www.stat.ufl.edu/~winner/data/concussion.txt

The description of the dataset given by the source is as follows: "Counts of concussions among collegiate athletes in 5 sports for 3 years by gender". This dataset was used in a study called "Sex Difference and the Incidence of Concussions Among Collegiate Athletes" published in the Journal of Athletic Training in 2003. This study did not use log linear modeling to test independence between incidence of concussions, gender, type of sport, and year. Instead, they used the dataset and chi-squared tests to conclude that "female athletes sustained a higher percentage of concussions during games than male athletes". This conclusion is irrelevant to this paper since we are testing completely different things using different methods.

There are three types of categorical variables in this dataset: ordinal, nominal, and binary. Nominal variables have categories without a natural ordering, while ordinal variables do have ordered categories. Binary variables only have two categories, which are usually 'success and 'failure'. This data set has two nominal variables, one binary variable, and one ordinal variable. Gender is a nominal variable with two categories: Male and Female. Type of sport is also a nominal variable with five categories: Soccer, Lacrosse, Basketball, Baseball/Softball, and Gymnastics. Incidence of concussions is a binary variable with only two categories: an athlete suffered a concussion or did not. The only ordinal variable is year, which has three categories: 1997, 1998, and 1999.

In order to find the best fitting model for the data, I will form multiple log linear models. Because this data set has four factors, testing all of the possible models would take an enormous amount of time. For that reason, I jumped straight from the mutual independence model ([G][S][Y][I]) to the pairwise association model ([GS][GY][GI][SY][SI][YI]). All of the models in between had high AIC values and P-Values less than 0.0001, meaning that they fit the data terribly. Below is a description of the models I will fit.

Model	Description
[G][S][Y][I]	Only the main effects; the mutual independence model
[GS][GY][GI][SY][SI][YI]	Includes all the pairwise interactions between the four factors

[GSI][Y]	Year is independent of Gender, Sport, and Incidence of Concussion
[GSY][I]	Incidence of Concussion is independent of Gender, Sport, and Year
[GYI][S]	Sport is independent of Gender, Year, and Incidence of Concussion
[SYI][G]	Gender is independent of Sport, Year, and Incidence of Concussion
[GSY][GSI]	Given Gender and Sport, Year is independent of Incidence of Concussions
[GSY][GYI]	Given Gender and Year, Sport is independent of Incidence of Concussions
[GSY][SYI]	Given Sport and Year, Gender is independent of Incidence of Concussions
[GYI][GSI]	Given Gender and Incidence of Concussions, Year is independent of Sport
[GSI][SYI]	Given Sport and Incidence of Concussions, Gender is independent of Year
[GYI][SYI]	Given Year and Incidence of Concussions, Gender is independent of Sport
[GSI][GSY][GYI]	Given Gender and Sport, Year is independent of Incidence of Concussions and Given Gender and Year, Sport is independent of Incidence of Concussions and Given Gender and Incidence of Concussions, Year is independent of Sport
[GSI][GSY][SYI]	Given Sport and Year, Gender is independent of Incidence of Concussions and Given Gender and Sport, Year is independent of Incidence of Concussions and Given Sport and Incidence of Concussions, Gender is independent of Year
[GSY][GYI][SYI]	Given Year and Incidence of Concussions, Gender is independent of Sport and Given Sport and Year, Gender is independent of Incidence of Concussions and Given Gender and Year, Sport is independent of Incidence of Concussions
[GSI][GYI][SYI]	Given Year and Incidence of Concussions, Gender is independent of Sport and Given Sport and Incidence of Concussions, Gender is independent of Year and Given Gender and Incidence of Concussions, Year is independent of Sport
[GSI][GSY][GYI][SYI]	Includes all three-way interactions between the four factors; the model of homogenous association

Where G = Gender, S = Type of Sport, I = Incidence of Concussion, Y = Year

I will test the fit of all of these models using the G^2 goodness-of-fit test, which will output a P-value with which I can distinguish the models. A P-value greater than 0.05 will indicate suitable models. In addition, I will use AIC to select the final model which not only fits the data, but also has a low amount of parameters. AIC compares all of the suitable models and adds twice the number of parameters to a model's G^2 value in order to make sure that the final chosen model is not overly complex.

Results

Model	Degrees of Freedom	P-Value	AIC
[G][S][Y][I]	51	<0.0001	52924.57
[GS][GY][GI][SY][SI][YI]	30	<0.0001	6216.92
[GSI][Y]	26	<0.0001	6215.03
[GSY][I]	22	0.1374	105.28
[GYI][S]	28	<0.0001	6220.5
[SYI][G]	22	<0.0001	6217.56
[GSY][GSI]	18	0.3737	103.3
[GSY][GYI]	20	0.0862	109.08
[GSY][SYI]	14	0.5018	105.32
[GYI][GSI]	24	<0.0001	6218.76
[GSI][SYI]	18	<0.0001	6216.35
[GYI][SYI]	20	<0.0001	6221.06
[GSI][GSY][GYI]	16	0.2584	107.2
[GSI][GSY][SYI]	10	0.9452	104.05
[GSY][GYI][SYI]	12	0.3752	108.91
[GSI][GYI][SYI]	19	<0.0001	6219.71
[GSI][GSY][GYI][SYI]	8	0.9077	107.39

Where G = Gender, S = Type of Sport, I = Incidence of Concussion, Y = Year

Based on the P-values above, there are 8 models that reasonably fit the data. However, the two models with the lowest AIC scores distinguished themselves in various ways. Model [GSI][GSY][SYI] had the second lowest AIC and the highest P-value, while model [GSY][GSI] did not have a great P-value, but is favored by AIC because of its lower complexity. Both models reasonably fit the data and are more favored by AIC compared to any of the other models, but it is still challenging to decide between the two. So I decided to compare the two models using G^2. By taking the difference of their G^2 values and degrees of freedom, it is easy to see if the two models are statistically different, or statistically the same model.

```
LR tests for hierarchical log-linear models

Model 1:
 count ~ Gender + Sport + Year + Id + Gender:Sport + Gender:Year + Sport:Year + Gender:Id + Sport:Id + Year:Id +
Gender:Sport:Year + Gender:Sport:Id
Model 2:
 count ~ Gender + Sport + Year + Id + Gender:Sport + Gender:Year + Sport:Year + Gender:Id + Sport:Id + Year:Id +
Gender:Sport:Year + Gender:Sport:Id + Sport:Year:Id

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   19.298032 18
Model 2    4.046791 10  15.251241         8        0.05444
Saturated  0.000000  0   4.046791        10        0.94521
```

The model comparison test has H_a: The two models are the same and H_o: The larger model is better. The difference in G^2 between model [GSY][GSI] and model [GSI][GSY][SYI] had a P-value of 0.054, just above our cutoff P-value at 0.05. Because of our cutoff, we fail to

reject H_a and can assume that model [GSY][GSI] is statistically the same as model [GSI][GSY][SYI]. With a different P-value cutoff, like 0.10, these two models would have been statistically different.

In normal circumstances, I would still probably chose model [GSI][GSY][SYI], because it was the most consistent. However, Professor Azari told us to choose the simplest model if two models were proven to be statistically the same, so in accordance with that I will choose model [GSY][GSI]. Model [GSY][GSI] says that given gender and type of sport, year is independent of incidence of concussions. Although this conclusion is not groundbreaking, it does shed some light on how to improve the next study. Since year is independent of incidence of concussions, the researcher should lump all incidence of concussions together and get rid of the year variable while keeping the gender and type of sport variables in the next study.

The conditional odds ratio between year and incidence of concussion given type of sport and gender turned out to have a 95% confidence interval of (0.979, 1.203) and an estimated conditional odds ratio of 1.086. The confidence interval suggests of a strong, positive conditional association between year and incidence of concussion given type of sport and gender. This only cements the previous assertion that year is independent of incidence of concussion given type of sport and gender.

Conclusion and Discussion

By fitting many various models for a four-factor dataset, I was able find a suitable model with an understandable meaning that was not overly complex. Using AIC and the model comparison test, I was able to further narrow down the set of suitable models to the model [GSY][GSI]. This model theorizes that given gender and type of sport, year is independent of incidence of concussions. This theory could help further next studies, which could remove the year variable, since it is independent of the incidence of concussions. This would allow for a study that could explain the root of the differences in gender and type of sport among concussions.

Model [GSY][I] especially interested me, because it had a P-value of 0.1374 and an AIC of 105.28. Given the suitability of this model to the data, its meaning is very peculiar. [GSY][I] suggests that incidence of concussions is independent of gender, type of sport, and year, which goes against the very fabric of studies on the relationship between concussions and sports. I thought it was very interesting that this model was suitable for this dataset.

Code Appendix

```
library(MASS)
concussion <- read.table("C:/Users/ckashyap/Downloads/concussion.csv", quote="\"",
comment.char="")
names(concussion) = c("Gender", "Sport", "Year", "Id", "count")

full = loglm(count~Gender*Sport*Year*Id, data=concussion, param = T) #Saturated Model
fit3.4 = update(full, .~. - Gender:Sport:Year:Id) #All 3-way Associations
fit3.3.1 = update(fit3.4, .~. -Gender:Sport:Year) #[GSI][GYI][SYI]
fit3.3.2 = update(fit3.4, .~. -Gender:Sport:Id) #[GSY][GYI][SYI]
```

```
fit3.3.3 = update(fit3.4, .~. -Gender:Year:Id) #[GSI][GSY][SYI]
fit3.3.4 = update(fit3.4, .~. -Sport:Year:Id) #[GSI][GSY][GYI]
fit3.2.1 = update(fit3.4, .~. -Gender:Sport:Year -Gender:Sport:Id) #[GYI][SYI]
fit3.2.2 = update(fit3.4, .~. -Gender:Sport:Year -Gender:Year:Id) #[GSI][SYI]
fit3.2.3 = update(fit3.4, .~. -Gender:Sport:Year -Sport:Year:Id) #[GYI][GSI]
fit3.2.4 = update(fit3.4, .~. -Gender:Sport:Id -Gender:Year:Id) #[GSY][SYI]
fit3.2.5 = update(fit3.4, .~. -Gender:Sport:Id -Sport:Year:Id) #[GSY][GYI]
fit3.2.6 = update(fit3.4, .~. -Gender:Year:Id -Sport:Year:Id) #[GSY][GSI]
fit3.1.1 = update(fit3.4, .~. -Gender:Sport:Year -Gender:Sport:Id -Gender:Year:Id) #[SYI]
fit3.1.2 = update(fit3.4, .~. -Gender:Sport:Year -Gender:Sport:Id -Sport:Year:Id) #[GYI]
fit3.1.3 = update(fit3.4, .~. -Gender:Sport:Id -Gender:Year:Id -Sport:Year:Id) #[GSY]
fit3.1.4 = update(fit3.4, .~. -Gender:Sport:Year -Gender:Year:Id -Sport:Year:Id) #[GSI]
indep = loglm(count~Gender+Sport+Year+Id, data = concussion, param = T) #Mutual
Independence Model
fit2.6 = update(indep, .~.^2) #All Pairwise Associations

extractAIC(indep)
extractAIC(fit2.6)
extractAIC(fit3.4)
extractAIC(fit3.3.1)
extractAIC(fit3.3.2)
extractAIC(fit3.3.3)
extractAIC(fit3.3.4)
extractAIC(fit3.2.1)
extractAIC(fit3.2.2)
extractAIC(fit3.2.3)
extractAIC(fit3.2.4)
extractAIC(fit3.2.5)
extractAIC(fit3.2.6)
extractAIC(fit3.1.1)
extractAIC(fit3.1.2)
extractAIC(fit3.1.3)
extractAIC(fit3.1.4)

anova(fit3.2.6,fit3.3.3,test="Chisq")

full = glm(count~Gender*Sport*Year*Id, data=concussion, family = "poisson") #Saturated
Model
fit3.4 = update(full, .~. - Gender:Sport:Year:Id) #All 3-way Associations
fit3.2.6 = update(fit3.4, .~. -Gender:Year:Id -Sport:Year:Id) #[GSY][GSI]

odds = exp(coef(fit3.2.6)["Year:Id"])

est_YI=coef(summary(fit3.2.6))["Year:Id","Estimate"]  # estimate
se_YI=coef(summary(fit3.2.6))["Year:Id","Std. Error"] # standar error, leave a space between
```

"Std." and "Error"
CI_YI=c(exp(est_YI-1.96*se_YI),exp(est_YI+1.96*se_YI))