Scott Herman
Predict 411 GLM

Moneyball Assignment

BACKGROUND

The purpose of this analysis is to develop a predictive model that will accurately tell us the number of wins each team will have over the course of a 162 game season. In order to develop an accurate model, we will first explore our data set consisting of roughly 2,276 observations, each representing a professional baseball team from 1871 up until 2006. We will then clean and prep our data set, which will enable us to develop a variety of linear regression models. Each model will be developed through a number of variable selection procedures. The results from each developed model will be compared, and the model that generates the most accurate prediction of wins will be selected as our final model.

DATA QUALITY

The first step in performing our exploratory analysis will be conducting a data quality check. Then, we will identify observations that may be missing, extreme or erroneous, and assess their validity in order to decide whether any corrections are necessary. This will then enable us to explore the relationships between our predictor and response variables. In order to accurately predict the number of wins each team will have, we have chosen to construct our model utilizing the fifteen predictor variables listed below in Table 1:

Table 1: Predictor Variables

PREDICTOR VARIABLE	DEFINITION	IMPACT
TEAM_BATTING_H	Base Hits	Positive
TEAM_BATTING_2B	Doubles	Positive
TEAM_BATTING_3B	Triples	Positive
TEAM_BATTING_HR	Home Runs	Positive
TEAM_BASERUN_SB	Stolen Bases	Positive
TEAM_BATTING_BB	Walks Earned	Positive
TEAM_BATTING_HBP	Batters Hit by Pitch	Positive
TEAM_PITCHING_SO	Strikeouts by Pitchers	Positive
TEAM_FIELDING_DP	Double Plays	Positive
TEAM_BATTING_SO	Strikeouts by Batters	Negative
TEAM_BASERUN_CS	Caught Stealing	Negative
TEAM_PITCHING_H	Allowed Hits	Negative
TEAM_PITCHING_HR	Allowed Home Runs	Negative
TEAM_PITCHING_BB	Allowed Walks	Negative
TEAM_FIELDING_E	Errors	Negative

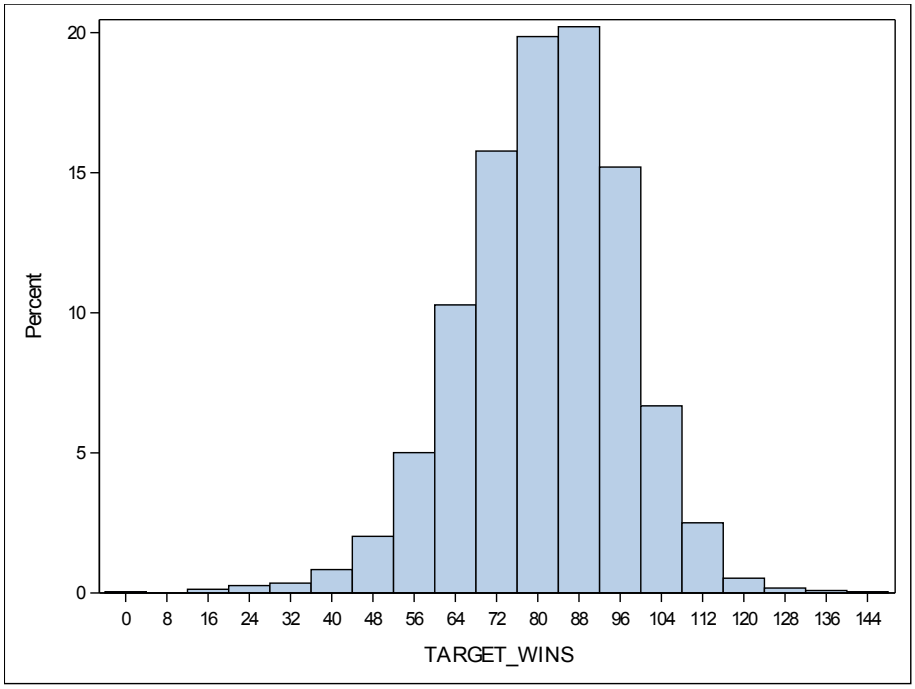
Given the fact that these observations were recorded over such a wide range of time (1876-2006), our first assumption tells us that there is a high likelihood that the data quality could be suspect. Over the course of time, we know that there have been a number of changes within

the game including changes to the fundamental rules, stadium dimensions, playing field surfaces, attendance levels, as well as the number of teams, which also effects the number of games played throughout each season. Since these changes could certainly impact the validity and accuracy of our model, it is important that we take these factors into consideration when ultimately determining how to deal with outliers and potential missing data points.

EXPLORATORY DATA ANLYSIS (EDA)

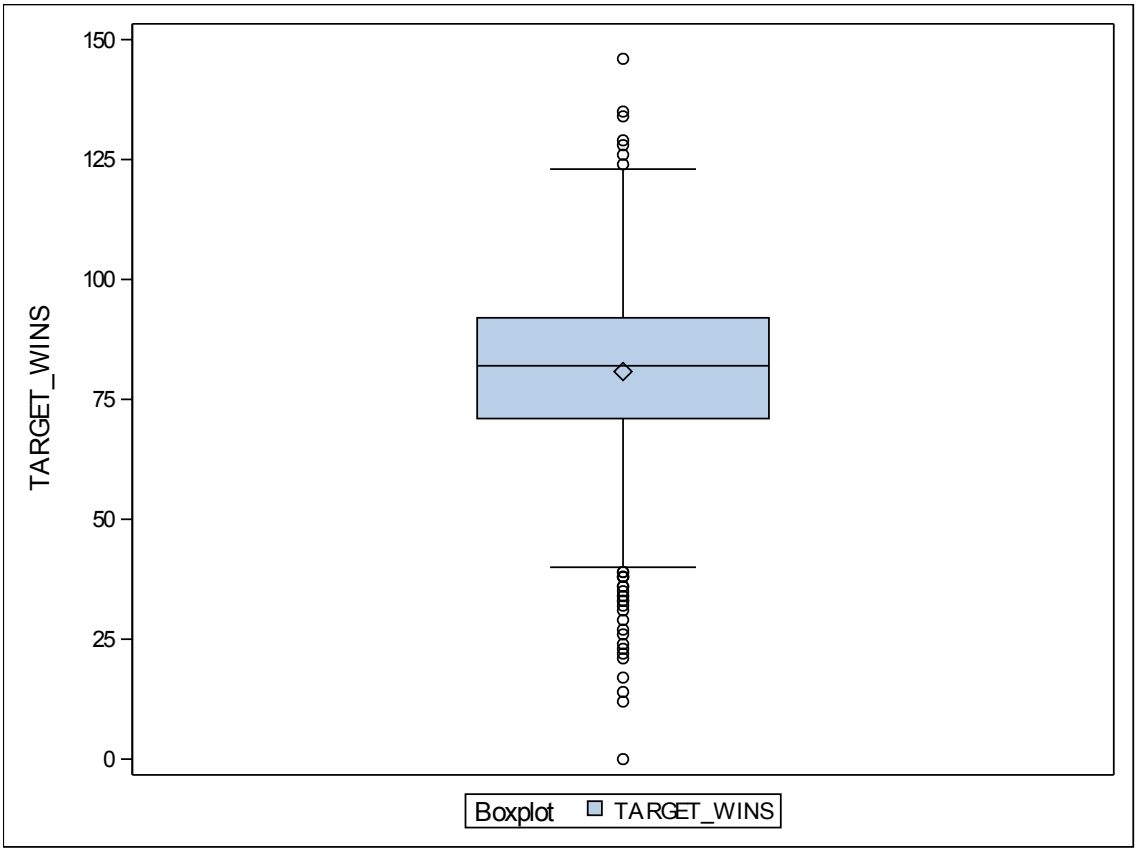
In investigating our response variable, TARGET_WINS per a 162 game season, we will need to test the assumption that these observations are normally distributed. We've plotted these observations in the histogram given in Figure 1 below, which shows us that this variable is in fact, normally distributed.

Figure 1: TARGET_WINS DISTRIBUTION



In order to explore this variable further, the Box Plot given in Figure 2 shows us that these observations range anywhere from 0 wins, all the way up to 142 wins. To test the validity of these points, we've conducted additional research to compare our observations with recorded team wins data collected by **mlb.com.** After examining this data, we confirm that the number of actual wins for any team ranges between 20 and 116 wins. These findings show that these values are outside the actual number of wins recorded. This indicates that the validity of these extreme points are questionable and will need to be further assessed.

Figure 2: TARGET_WINS BOXPLOT



The quantiles for TARGET_WINS are given below, which were produced through SAS using the Univariate Procedure. Table 2 below, indicates that only the lowest and highest 1% of our values are outside the range of possibility.

Table 2: TARGET_WINS QUANTILES

Level	Wins Quantile
100% Max	146
99%	114
95%	104
90%	100
75% Q3	92
50% Median	82
25% Q1	71
10%	61
5%	54
1%	38
0% Min	0

The next step is performing a similar exploration of our predictor variables. The resulting output from the SAS Means procedure, represented in Table 3 below, indicates that we have six different predictor variables that possess a substantial number of missing values. These missing observations will need to be addressed during our data preparation step, as they would negatively impact the accuracy of our model.

Table 3: Predictor Variables with High N Missing Values

PREDICTOR VARIABLE	N Missing Values	Pct. Missing
TEAM_BATTING_HBP	2,085	92%
TEAM_BASERUN_CS	772	34%
TEAM_FIELDING_DP	286	13%
TEAM_BASERUN_SB	131	6%
TEAM_BATTING_SO	102	4%
TEAM_PITCHING_SO	102	4%

Additionally, we've identified to additional variables which contain records with extreme values. The boxplots represented in Figure 3 and Figure 4 below, show these outlier observations in our variables for PITCHING_SO and TEAM_PITCHING_H, respectively.

Figure 3: PITCHING_SO

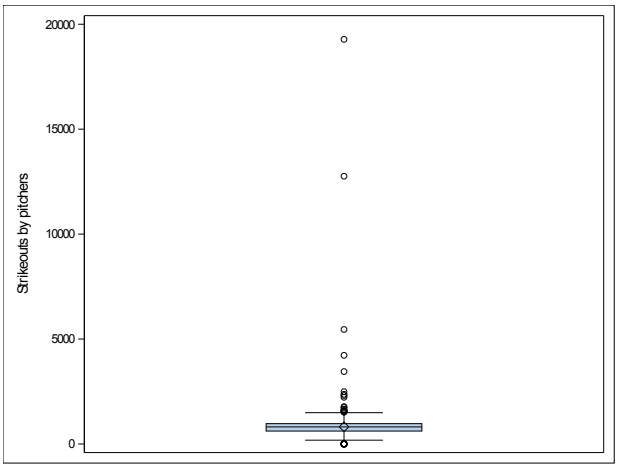
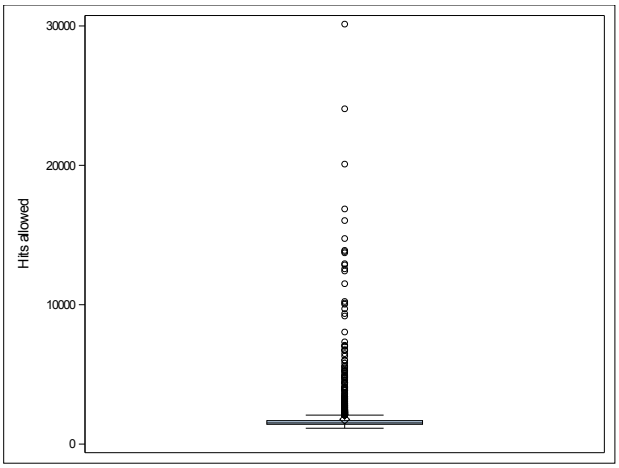


Figure 4: TEAM_PITCHING



The main goal of our exploratory data analysis is to evaluate the correlation each predictor variable has on our response variable, TARGET_WINS. This will be assessed by examining the correlation coefficient and R-Squared results which reveals both the strength and direction of the relationship each of our predictor variables has on the number of wins in a season. Below, Table 4 summarizes the R-Squared value of each of our predictor variables has on the response variable, TARGET_WINS.

Table 4: Predictor Variable Correlation to TARGET_WINS

PREDICTOR VARIABLE	R Value	p Value
TEAM_BATTING_H	0.38877	< .001
TEAM_BATTING_2B	0.28910	< .001
TEAM_BATTING_BB	0.23256	< .001
TEAM_PITCHING_HR	0.18901	< .001
TEAM_FIELDING_E	-0.17648	< .001
TEAM_BATTING_HR	0.17615	< .001
TEAM_BATTING_3B	0.14261	< .001
TEAM_BASERUN_SB	0.13514	< .001
TEAM_PITCHING_BB	0.12417	< .001
TEAM_PITCHING_H	-0.10994	< .001
TEAM_PITCHING_SO	-0.07844	0.0003
TEAM_BATTING_HBP	0.0735	0.3122
TEAM_FIELDING_DP	-0.03485	0.1201
TEAM_BATTING_SO	-0.03175	0.1389
TEAM_BASERUN_CS	0.0224	0.3853

DATA PREPARATION:

Moving on to our data preparation stage, we'll proceed with eliminating the values collected for TARGET_WINS that are below 20 and above 116. Additionally, in our exploratory analysis, we were able to identify that our predictor variable TEAM_BATTING_HBP, is missing over 90% of the values in our data set. Thus, we've removed this variable entirely from our data set.

Additionally, we've identified five other variables that require correction due to missing values. Rather than removing these values from the data set, we've decided to replace these missing points with the mean values for each of the respective observations with each predictor variable set. These variables are TEAM_BASERUN_CS, TEAM_FIELDING_DP, TEAM_BASERUN_SB, TEAM_BATTING_SO, and TEAM_PITCHING_SO. In order to record these changes, we've renamed each of these variables to indicate that they've been imputed by dropping the prefix of "TEAM_" for each variable name. Furthermore, we want to indicate and track the missing values we've removed, and have added five new variables to our data set to measure the possible potential of their predictive ability.

Next, we need to address the outlier records we identified above in order to correct our values for PITCHING_SO and TEAM_PITCHING_H. We've chosen to eliminate the largest extreme values from each of these variables, leaving us now with a total of 2,249 records in our final data set. Table 5 below, illustrates the descriptive statistics for all of the remaining variables included in our data set.

Table 5: Descriptive Statistics for Variables contained in Final Data Set

Variable	Label	# Miss	Mean	Median	Max	Min	Std Dev
TARGET_WINS		0	81	82	116	21	15
TEAM_BATTING_H	Base Hits by batters	0	1467	1454	2496	992	136
TEAM_BATTING_2B	Doubles by batters	0	241	238	458	69	46
TEAM_BATTING_3B	Triples by batters	0	55	47	223	0	28
TEAM_BATTING_HR	Homeruns by batters	0	100	103	264	0	60
TEAM_BATTING_BB	Walks by batters	0	504	513	878	29	119
TEAM_PITCHING_H	Hits allowed	0	1729	1517	20088	1137	1101
TEAM_PITCHING_HR	Homeruns allowed	0	106	108	343	0	61
TEAM_PITCHING_BB	Walks allowed	0	553	537	3645	119	164
TEAM_FIELDING_E	Errors	0	241	158	1898	65	215
BASERUN_CS	Caught Stealing (fixed)	0	53	53	201	11	19
MISSING_BASERUN_CS		0	0	0	1	0	0
FIELDING_DP	Double Plays (fixed)	0	146	146	228	52	25
MISSING_FIELDING_DP		0	0	0	1	0	0
BASERUN_SB	Stolen Bases (fixed)	0	125	106	697	18	85
MISSING_BASERUN_SB		0	0	0	1	0	0
BATTING_SO	Strikeouts by Batters(fixed)	0	738	736	1273	0	237
MISSING_BATTING_SO		0	0	0	1	0	0
PITCHING_SO	Strikeouts by Pitchers(fixed)	0	742	779	1273	0	238
MISSING_PITCHING_SO		0	0	0	1	0	0

Now that we've made some updates to our set of predictor variables, we want to examine the correlation each of these will have on our target variable, which is given by Table 6, below.

Table 6: New Predictor Variable Set and Correlation to TARGET_WINS

PREDICTOR VARIABLE	R Value	p Value
TEAM_BATTING_H	0.35165	< .001
TEAM_BATTING_2B	0.26744	< .001
TEAM_BATTING_BB	0.25481	< .001
TEAM_PITCHING_HR	0.19692	< .001
TEAM_BATTING_HR	0.19551	< .001
TEAM_FIELDING_E	-0.18706	< .001
TEAM_PITCHING_BB	0.13012	< .001
TEAM_BATTING_3B	0.11092	< .001
BASERUN_SB	0.11039	< .001
TEAM_PITCHING_H	-0.07108	.0007
PITCHING_SO	-0.02341	0.2671
BATTING_SO	-0.02341	0.2672
FIELDING_DP	-0.01736	0.4105
BASERUN_CS	0.00038	0.9855

In reviewing the results from the table above, it appears that it might make sense to create an additional variable to remove multicollinearity. A new variable was created called TEAM_EXTRA_BASE, which combines TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR into one variable. An additional PROC REG procedure was ran in order to understand the correlation to TARGET_WINS for our final predictor set of variables.

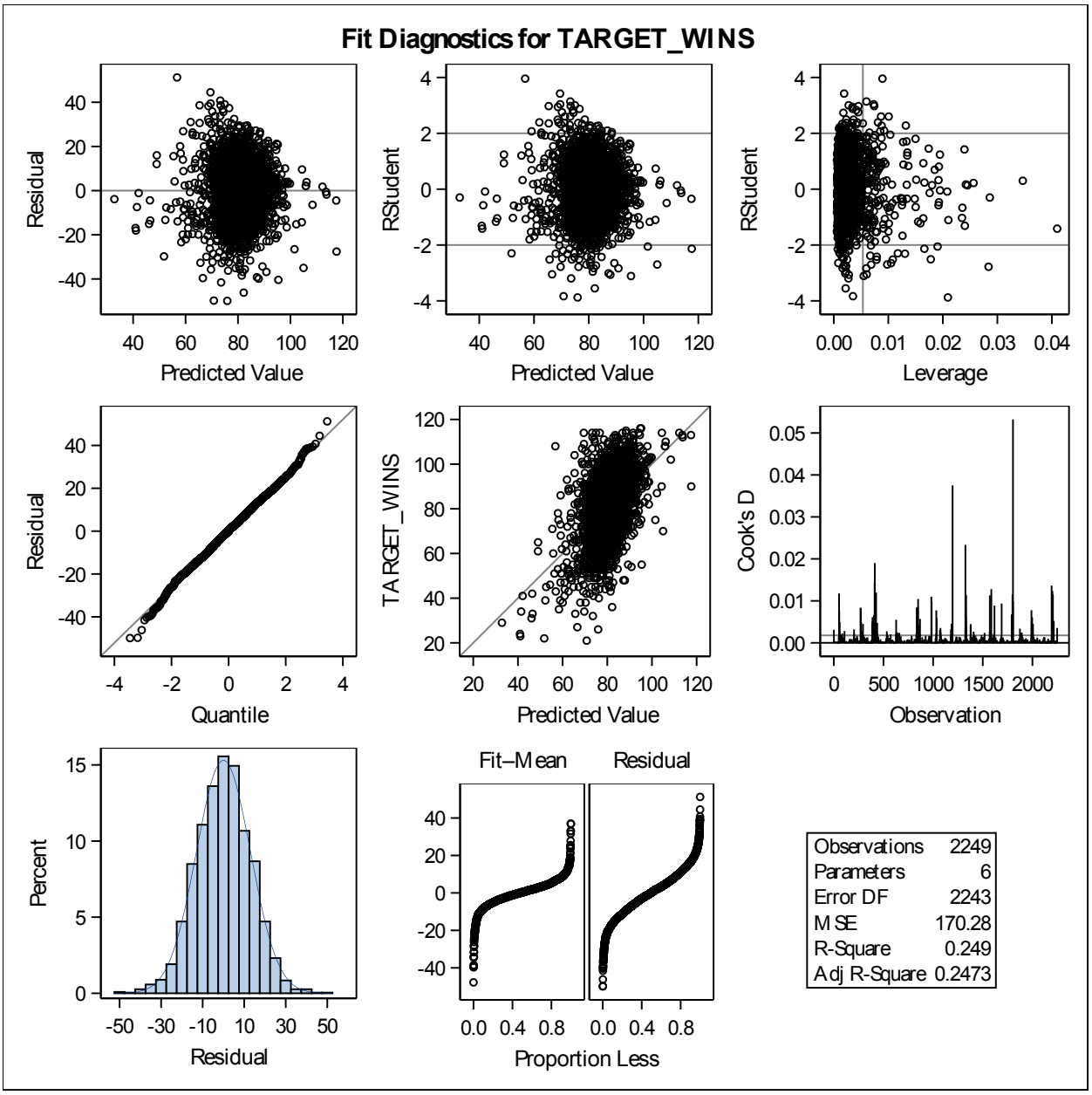
Table 7: New Predictor Variable Set and Correlation to TARGET_WINS

PREDICTOR VARIABLE	R Value	p Value
TEAM_BATTING_H	0.35165	< .001
TEAM_EXTRA_BASE	0.33571	< .001
TEAM_BATTING_BB	0.25481	< .001
TEAM_PITCHING_HR	0.19692	< .001
TEAM_FIELDING_E	-0.18706	< .001
TEAM_PITCHING_BB	0.13012	< .001
BASERUN_SB	0.11039	< .001
TEAM_PITCHING_H	-0.07108	.0007
PITCHING_SO	-0.02341	0.2671
BATTING_SO	-0.02341	0.2672
FIELDING_DP	-0.01736	0.4105
BASERUN_CS	0.00038	0.9855

MODEL RESULTS

In developing an accurate predictive model, we've utilized three different selection procedures; Forward Selection, Stepwise Selection and Manual Selection. Each of these models yielded similar results, which led us to select the model that was the easiest to interpret and consisted of the fewest number of variables. This final model utilized manual selection, and these results are summarized below.

Figure 5: Goodness of Fit for our Final Model



A summary table of the Forward Selection results for our final model is listed below.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	11.39410	3.34248	3.41	0.0007
TEAM_BATTING_H	Base Hits by batters	1	0.04079	0.00274	14.87	<.0001
TEAM_EXTRA_BASE		1	0.01305	0.00507	2.57	0.0101
BASERUN_SB	Stolen Bases (fixed)	1	0.03744	0.00373	10.04	<.0001
TEAM_BATTING_BB	Walks by batters	1	0.00884	0.00321	2.75	0.0060
TEAM_FIELDING_E	Errors	1	-0.02035	0.00203	-10.00	<.0001

The formula for our final model is as follows:

TARGET_WINS = 11.349410 +

0.004079*TEAM_BATTING_H +

0.01305*TEAM_EXTRA_BASE +

0.03744*BASERUN_SB +

0.00884*TEAM_BATTING_BB –

0.02035*TEAM_FIELDING_E

CONCLUSION

Although we were able to develop a model with a fairly strong R-Squared value, we feel the accuracy of our model could be improved. First, we cited a number of data quality issues from our substantially large data set which spanned across over 100 years of baseball seasons. We feel strongly that the accuracy of our final model would be improved if we limited the sample size of the number of seasons from which we analyze data for. Given the number of changes the league has seen as it has evolved over the years, we believe it would make sense to limit our data set to cover only the last 10 seasons. This would not only provide a statistically significant sample size of data to observe, but this would also bring a more consistent basis from which to compare from. For example, in the late 1990's and early 2000's we know that there have been a large number of cases citing players for using illegal performance enhancing supplements that would have had a direct impact on both individual and team performance.

To conclude, we feel that additional research is required in order to develop a better, more accurate model.