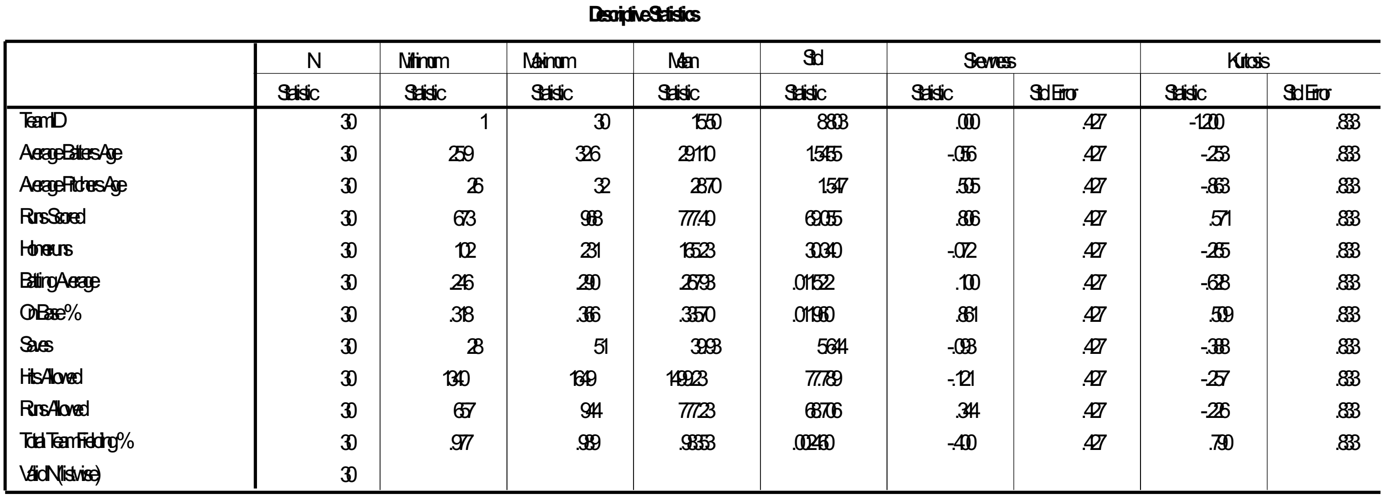
Introduction

We are running a discriminant analysis to try and predict whether or not a Major League Baseball team will make the playoffs. We are running the analysis for the 2005-2007 MLB seasons. Also we are trying to see based on offense statistics and defense statistics how well the discriminant analysis function can predict the teams that will make the playoffs. The offense statistics that will be our independent variables are: Runs Scored, Batting Average, On Base Percentage, Average Batters Age, and Homeruns. And for defense our independent variables will include; Hits allowed, Runs allowed, Total Team Fielding Percentage, Saves and Average Pitchers Age. Our dependent variable, what we are trying to predict, is making the playoff or not, Playoffs. Which indicates whether or not a team made the playoffs or didn't make the playoffs for that year. The discriminant analysis will try and predict which teams should have made the playoffs based on the statistics we indicate, and compare them to the actual results to see how accurate the model is at predicting the teams that made the playoffs. Also the anlaysis did not always choose 8 teams to make the playoff (4 from the American League and 4 from the National League) but due to the data provided it is impossible to make the model consistently provide 8 teams being predicted, so the while in those cases the accuracy may be a little off but the data still provides interesting and important results to our analysis.

2007

Descriptive Statistics



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Team ID	30	1	30	15.50	8.803	.000	.427	-1.200	.833
Average Batters Age	30	25.9	32.6	29.110	1.5455	-.056	.427	-.253	.833
Average Pitchers Age	30	26	32	28.70	1.547	.505	.427	-.863	.833
Runs Scored	30	673	968	777.40	69.055	.806	.427	.571	.833
Homeruns	30	102	231	165.23	30.340	-.072	.427	-.265	.833
Batting Average	30	.246	.290	.2673	.01152	.100	.427	-.628	.833
On Base %	30	.318	.366	.3357	.01196	.861	.427	.509	.833
Saves	30	28	51	39.93	5.644	-.093	.427	-.388	.833
Hits Allowed	30	1340	1649	1499.23	77.789	-.121	.427	-.257	.833
Runs Allowed	30	657	944	777.23	68.706	.344	.427	-.226	.833
Total Team Fielding %	30	.977	.989	.9835	.00246	-.400	.427	.790	.833
Valid N (listwise)	30								

For Offense and Defense independent variables there exists no problems with skewness and Kurtosis. So the data for 2007 has no normality problems and the data is sufficient to use.

Offense

Level of measurement and sample size issues

The variables being used in a discriminant analysis should be non-metric for the dependenant variable and metric for the independent variables, which in this analysis and the following analysis's is true so the measurement level requirement is satisfied. The minimum ratio of valid cases to independent variables for discriminant analysis is 5 to 1, with a preferred ratio of 20 to 1. In this analysis, there are 30 valid cases and 5 independent variables. The ratio here is 6 to 1 so the ratio exceeds the minimum. So the sample size requirement for discriminant analysis is satisfied.

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		30	100.0
Exclud	Missing or out-of-	0	.0

ed	range group codes		
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		30	100.0

In addition to the requirement for the ratio of cases to independent variables, discriminant analysis requires that there be a minimum number of cases in the smallest group defined by the dependent variable. The number of cases in the smallest group must be larger than the number of independent variables, and preferably contain 20 or more cases. In this analysis, the number of cases in the smallest group does not contain more than 20 but it does contain more than 5, which is the number of independent variables. This requirement is also met. This will be the same for all preceding years and whether it is the offense or defense data, so for the preceding analysis it will not be included because the data will be redundant if provided.

Prior Probabilities for Groups

	Prior	Cases Used in Analysis	
Made the Playoffs	Unweighted	Weighted	Unweighted
Did Not Make Playoffs	.733	22	22.000
Made Playoffs	.267	8	8.000
Total	1.000	30	30.000

<u>Assumption of homogeneity of variance</u>

If we fail to reject the null hypothesis and conclude that the variances are equal, we use the SPSS default of using a pooled covariance matrix in classification. And in this case the significance of .149 > .05 so we fail to reject and the homogeneity is satisfied in this case.

Test Results

Box's M		29.504
F	Approx.	1.384
	df1	15
	df2	703.353
	Sig.	.149

Overall Relationship

The Wilks' lambda statistic for the test of the function (Wilks' lambda=.496) had a probability of p=0.003 which was less than or equal to the level of significance of 0.05. Which indicates that there is an overall relationship.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.496	17.891	5	.003

Multicollinearity

Like multiple regression, multicollinearity in discriminant analysis is identified by examining tolerance values. While tolerance is routinely included in the output for the stepwise method for including variables, it is not included for simultaneous entry of variables. If a tolerance problem occurs in a simultaneous entry problem, SPSS will include a table titled "Variables Failing Tolerance Test." So since SPSS did not include this table it indicates that multicollinearity is not a problem in the analysis. No problem with multicollinearity exists in this year's data or any other years so it will not be included in the analysis from here on out.

Role of independent variables in predicting group membership

Functions at Group Centroids

Made the Playoffs	Function
	1
Did Not Make Playoffs	.588
Made Playoffs	-1.616

Unstandardized canonical discriminant functions evaluated at group means

In this analysis this discriminant function assigns positive and negative values to separate the subgroups, making the playoffs and not making the playoffs. This is used to differentiate between the two groups.

Structure Matrix

	Function
	1
On Base %	-.729
Runs Scored	-.597
Batting Average	-.370
Average Batters Age	.153
Homeruns	-.132

Looking at the structure matrix all of the independent variables correlate with making the playoffs since they have negative values. As you can see On-Base % (-.729) is the most important variable contributing to making the playoffs. Followed by Runs Scored (-.597). The Average Batters age actually correlates with making the playoffs but since the discriminant function thinks a higher avg is better it says that it will have the opposite effect, but in reality the teams that made the playoffs have batters with a lower average age which indicates younger players result in a better chance of a team making the playoffs but it is not as strong a factor as 3 of the independent variables(.153). And finally when looking at the effect each independent variable has on the overall function homeruns was the least important in determining whether or not a team makes the playoffs (-.132).

<u>Classification using the discriminant model</u>

Classification Results[a]

			Predicted Group Membership		
		Made the Playoffs	Did Not Make Playoffs	Made Playoffs	Total
Original	Count	Did Not Make Playoffs	22	0	22
		Made Playoffs	2	6	8
	%	Did Not Make Playoffs	100.0	.0	100.0
		Made Playoffs	25.0	75.0	100.0

[a] 93.3% of original grouped cases correctly classified.

After looking at the classification results the model correctly classified 93.3% of the original group cases. In this analysis only ARZ and CHC were classified as not making the playoffs when they actually made the playoffs. Since it is not possible to have the model predict 8 teams to make the playoffs, 4 rom each division, it only predicted 6 teams to make the playoffs when 8 teams are required for the playoffs. Even with the error in the model it still provides significant results in predicting a teams participation in the playoffs. On-Base %, Runs Scored, Batting Average, Average Batters Age, and Homeruns all are important offense statistics to determine whether or not a team will make the playoffs. While some are more important than others they all provide information to help predict the results.

Defense

Assumption of homogeneity of variance

Test Results

Box's M		7.640
F	Approx.	.358
	df1	15
	df2	703.353
	Sig.	.988

Tests null hypothesis of equal population covariance matrices.

The assumption of homogeneity of variance is satisfied in this analysis. Since the significant value .988 > .05 we fail to reject the null hypothesis that tests the null hypotheses that the group variance-covariance matrices are equal. Since we fail to reject we will use the group variance-covariance matrices and can conclude that homogeneity is satisfied.

Overall Relationship

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.719	8.417	5	.135

While if you used the normal way to determine the significance of the overall model .135 > .05 we would say there is no overall relationship. But when looking at the independent variables separately the results are significant.

Role of independent variables in predicting group membership

Functions at Group Centroids

	Function
Made the Playoffs	1
Did Not Make Playoffs	-.364
Made Playoffs	1.002

Unstandardized canonical discriminant functions evaluated at group means

In the discriminant function it seperates between the two supgroups, making the playoffs and not making the playoffs, here the variables with negative values will relate to teams who did not make the playoffs and positive values will correlate with teams who did make the playoffs.

Structure Matrix

	Function
	1
Total Team Fielding %	.691
Runs Allowed	-.674
Saves	.514
Hits Allowed	-.511
Average Pitchers Age	.444

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

When reviewing the structure matrix 2 variables load on the not making the playoffs, Runs Allowed(-.674) and Hits Allowed(-.511), and 3 variables relate to making the playoffs Team Fielding %(.691), Saves(.514), and Average Pitchers Age(.444). So teams that allowed more runs and hits were predicted to not make the playoffs. And teams that had a higher fielding percentage, more saves and a higher pitcher's average age have a better chance of making the playoffs.

Classification using the discriminant model

Classification Results[a]

			Predicted Group Membership		
		Made the Playoffs	Did Not Make Playoffs	Made Playoffs	Total
Original	Count	Did Not Make Playoffs	21	1	22
		Made Playoffs	4	4	8
	%	Did Not Make Playoffs	95.5	4.5	100.0
		Made Playoffs	50.0	50.0	100.0

[a] 83.3% of original grouped cases correctly classified.

The overall model correctly classified 83.3% of the original cases. This is pretty significant and even though it showed that there wasn't an overall relationship when looking at variables independently and the overall classification rate it is clear a relationship exists. This time like the offense analysis for 2007 they said ARZ and CHC should not have made the playoffs, also they conclude based on the defense statistics that LAA and NYY should not have made the playoffs. And they predicted SD make the playoffs when they did not make the playoffs. This result gave me more hope in the actual model because in 2007 SD had to play COL in a extra game to see which would advance to the playoffs so it would make sense they model predicted them to make it.

2006

Descriptive Statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Average Batters Age	30	25.6	33.5	29.150	1.4180	.447	.427	2.646	.833
Average Pitchers Age	30	25.9	32.5	28.770	1.5358	.729	.427	.498	.833
Runs Scored	30	689	930	786.63	57.188	.439	.427	-.088	.833
Homeruns	30	124	236	179.53	26.186	.160	.427	-.299	.833
Batting Average	30	.255	.287	.26927	.008948	.240	.427	-.731	.833
On Base%	30	.314	.363	.33640	.010338	.242	.427	.525	.833
Saves	30	24	54	40.03	6.547	-.177	.427	.266	.833
Hits Allowed	30	1385	1648	1502.43	69.375	-.027	.427	-.891	.833
Runs Allowed	30	675	971	786.63	64.107	.594	.427	1.344	.833
Total Team Fielding%	30	.978	.989	.98317	.002534	-.173	.427	.087	.833
Valid N (listwise)	30								

Here you can see there is not a problem with skewness, and there exists a slight problem in kurtosis for Average Batters Age and Runs Allowed. So we will proceed with caution in the analysis when looking at these to variables.

Offense

Assumption of homogeneity of variance

Test Results

Box's M		13.488
F	Approx.	.633
	df1	15
	df2	7033.553
	Sig.	.849

Tests null hypothesis of equal population covariance matrices.

The assumption of homogeneity of variance is satisfied in this analysis. Since the significant value .849 > .05 we fail to reject the null hypothesis that tests the null hypotheses that the group variance-covariance matrices are equal. Since we fail to reject we will use the group variance-covariance matrices and can conclude that homogeneity is satisfied.

Overall Relationship

The Wilks' lambda statistic for the test of the function (Wilks' lambda=.833) had a probability of p=0.459 which was greater than or equal to the level of significance of 0.05. Which indicates that there is not an overall relationship.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.833	4.662	5	.459

Role of independent variables in predicting group membership

Functions at Group Centroids

Playoffs	Function 1
Did Not Make the Playoffs	-.261
Made the Playoffs	.718

Unstandardized canonical discriminant functions evaluated at group means

In the discriminant function it seperates between the two supgroups, making the playoffs and not making the playoffs, here the variables with negative values will relate to teams who did not make the playoffs and positive values will correlate with teams who did make the playoffs.

Structure Matrix

	Function

	1
On Base %	.671
Average Batters Age	.640
Runs Scored	.611
Batting Average	.466
Homeruns	-.048

In the discriminant model 4 of the 5 statistics relate to teams making the playoffs. The independent variable that most significantly influences whether a team makes the playoffs or not is On-Base % (.671), followed by Average Batters Age (.640), Runs Scored(.611), and lastly Batting Average(.466). And for 2006 Homeruns correlates with teams not making the playoffs and that is why it has a negative value (-.048) but since it is less than .30 it really doesn't have much of an effect on the model.

Classification using the discriminant model

Classification Results[a]

		Playoffs	Predicted Group Membership		Total
			Did Not Make the Playoffs	Made the Playoffs	
Original	Count	Did Not Make the Playoffs	21	1	22
		Made the Playoffs	6	2	8
	%	Did Not Make the Playoffs	95.5	4.5	100.0
		Made the Playoffs	75.0	25.0	100.0

[a] 76.7% of original grouped cases correctly classified.

Though the model successfully predicted 76.7% of the original groups, it only correctly classified 25% of the teams that made playoffs. In this case they only predicted 3 teams to make the playoffs which doesn't really cooperate with the MLB, because 8 teams make the playoffs, but there is no way to account for this error. So since there is no overall relationship this analysis isn't really useful. But when looking at previous and following year models it is important to still consider these results in trying to determine if this model can be used to predict whether or not a team makes the playoffs. So in this case they only predicted LAD, NYY, and SF to make the playoff, of which only MIL and NYY actually made the playoffs. And they predicted NYM, STL, SD, DET,MIN and OAK. This is interesting because the World Series in 2006 was between STL and DET which would mean both teams should not have made the playoffs based on these statistics. It is also important to mention that many other variables that can predict a team making the playoffs that is far beyond the scope of this model.

Defense

Assumption of homogeneity of variance

Test Results

Box's M		27.511
F	Approx.	1.290
	df1	15
	df2	703.353
	Sig.	.202

Tests null hypothesis of equal population covariance matrices.

The assumption of homogeneity of variance is satisfied in this analysis. Since the significant value .202 > .05 we fail to reject the null hypothesis that tests the null hypotheses that the group variance-covariance matrices are equal. Since we fail to reject we will use the group variance-covariance matrices and can conclude that homogeneity is satisfied.

Overall Relationship

The Wilks' lambda statistic for the test of the function (Wilks' lambda=.548) had a probability of p=0.009 which was not less than or equal to the level of significance of 0.05 but close enough for this anlysis. Which indicates that there is an overall relationship.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.548	15.316	5	.009

Role of independent variables in predicting group membership

Functions at Group Centroids

	Function
Playoffs	1
Did Not Make the Playoffs	.529
Made the Playoffs	-1.454

Unstandardized canonical discriminant functions evaluated at group means

In the discriminant function it seperates between the two supgroups, making the playoffs and not making the playoffs, here the variables with negative values will relate to teams who made the playoffs and positive values will correlate with teams who did not make the playoffs.

Structure Matrix

	Function
	1
Runs Allowed	.870
Saves	-.474
Hits Allowed	.440
Average Pitchers Age	-.400
Total Team Fielding %	-.227

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

In the discriminant model 3 of the 5 statistics relate to teams making the playoffs. The independent variable that most significantly influences whether a team makes the playoffs or not is Saves (-.474), followed by Average Pitchers Age (-.400), and lastly Total Team Fielding %(-.227) And for 2006 Run Allowed(.870) and Hits Allowed (.440) correlates with teams not making the playoffs.

Classification using the discriminant model

Classification Results[a]

			Predicted Group Membership		
		Playoffs	Did Not Make the Playoffs	Made the Playoffs	Total
Original	Count	Did Not Make the Playoffs	21	1	22
		Made the Playoffs	2	6	8
	%	Did Not Make the Playoffs	95.5	4.5	100.0
		Made the Playoffs	25.0	75.0	100.0

[a] 90.0% of original grouped cases correctly classified.

The discriminant model was able to correctly classify 90% of the original group cases which makes the model extremely significant. In this model they did not predict STL and OAK to make the playoffs when they actually made the playoffs and predicted HOU to make the playoffs when they didn't actually make the playoffs. So you could say based on this model, the independent variables used can be used to predict whether or not a team will make the playoffs.