Paul Brendel
Final Project

INTRODUCTION

One of the most entertaining aspects of Major League Baseball (MLB) is watching a player hit for extra bases (doubles, triples, and home runs). This power was on full display during baseball's steroid era, which is believed to have occurred roughly between the late 1980's to the late 2000's[1]. On November 15, 2005 MLB and the players' association agreed on a plan to significantly strengthen steroid testing and penalties (including a lifetime ban for 3rd offenses)[2].

One statistic that measures the ability to hit for power is isolated power (ISO), which tells you the number of extra bases the player has per at bat (ISO = (2B + (2*3B) + (3*HR) / AB)[3]. Various factors may influence a team's average ISO. American League (AL) teams may have a higher ISO than those of National League (NL) teams because AL teams have a designated hitter bat instead of a pitcher, who is usually a very poor batter. Teams with the highest salaries may have the highest ISOs since they can afford to sign the best players. Teams who play home games in stadiums with more hitter-friendly park factors (park dimensions, weather, air density/quality, etc.) may have a better ISO than those of teams who play in more pitcher-friendly parks. I suspect that each team's average ISO decreased after MLB's strict steroid testing and penalty system was put into place. I also believe that teams in the AL with the highest salaries and with the most hitter-friendly home park factors will have the highest ISO.

METHODS

Team-level data for this study was obtained from fangraphs.com and baseballreference.com. Thirty teams were included and measurements for the outcome, ISO, were taken from 1998, 2001, 2004, 2007, 2010, and 2013, reflecting time before and after the implementation of the stricter steroids policies. There are 14 AL teams and 15 NL teams. The Houston Astros have a missing value for AL/NL since they moved from the AL to the NL in 2013. There are otherwise no missing values (the data is balanced). Salary and park factor (PF) was set to a fixed value by taking the average of the teams' 1998 and 2013 values. Fangraphs calculates park factor as a value where 100 is considered average, >100 is considered hitter-friendly, and <100 is considered pitcher-friendly. Salary and PF will be examined both as binary variables (salary: <median vs. => median, PF: <100 vs. =>100) and continuous variables (using salary/$1,000,000). All variables were approximately normally distributed. Mean ISO values by covariate group are displayed in Table 1. The overall mean ISO was .153 (s.d. = .020), the mean salary was $72,185,692.88 (s.d. = $25,208,705.75), and the mean PF was 100.13 (s.d. = 4.68). There were two salary outliers, belonging to the Yankees and Dodgers, and one PF outlier, belonging to the Rockies, but outliers were not removed. We will be using PROC MIXED in SAS 9.4 to fit mixed effects linear models. These flexible models are used for hierarchical data and allow one to model both the mean and the variance/covariance[4]. This model is an ideal choice for modeling a longitudinal, continuous outcome, such as yearly ISO.

RESULTS

A profile plot of 8 randomly selected teams (Figure 1) did not reveal any obvious trends in ISO over time. Two-sample T tests assuming equal variances were performed comparing the ISOs of AL vs. NL, high salary vs. low salary, and high PF vs. low PF (Table 1). A statistically

significant difference ($p<.05$) was seen when comparing salary and PF groups, but not when comparing leagues. Empirical summary plots (Figure 2) corroborate this conclusion, as it appears that the high salary, high PF teams have a distinctly higher ISO over time. These plots also show a dip in ISO after 2004, which supports the main hypothesis. A correlation matrix of tri-yearly ISO showed a wide range of Pearson correlation coefficients with no obvious pattern by lag time (Table 2).

Three different mean structures were considered for the mixed effects model: unstructured means, linear, and a linear spline with a knot at 2004. Each model was fit using maximum likelihood and models were compared based on AICC. The results are shown in Table 3. The linear spline model was selected due to it having the most negative AICC value. This model is appropriate if there is a different linear slope pre-2004 vs. post-2004, as I suspect. Using the linear spline model, the covariance pattern was selected in the same fashion (Table 4). The autoregressive(1) covariance structure was selected based on its AICC. This structure implies that the correlation between ISO values decreases exponentially over time, which makes sense because each team's roster becomes more dissimilar over time.

Lastly, using the linear spline model with autoregressive(1) covariance, models with different covariates were compared to determine the appropriate covariate set (Table 5). Examining the full model, league is not a significant predictor of ISO ($p=.46$), as we would expect based on the previous T test results. Therefore, the final model includes salary and PF as covariates. These covariates become more significant predictors of ISO when changed from binary to continuous variables. Based on this final model, ISO increased by .0034 tri-yearly from 1998-2004 and then decreased tri-yearly by .0066 (.0034-.010) from 2004-2013. A team with a $100,000,000 salary and a 100 park factor would be expected to have the following ISOs: .163 (1998), .167 (2001), .170 (2004), .164 (2007), .157 (2010), .150 (2013).

CONCLUSION

A limitation of this study was the use of fixed instead of time-varying covariates. Some teams have significantly varying spending based on whether they are competing or rebuilding. In addition, a wider range of years, encompassing a fuller range of the steroid era, would have been ideal. Future analyses may perform similar methods with more yearly data and on the player-level instead of the team-level. It may be interesting to exclusively examine players who attempt to hit for power instead of those who "hit for contact" based on the players' swing launch angles.

The results of the analysis support the hypothesis that teams' ISOs decreased over time after stricter steroid policies were implemented in MLB. The secondary hypothesis was partially supported – teams with higher salaries and more favorable home park factors had higher ISOs. However, a team's league was not a significant predictor of ISO. While the average ISO of a designated hitter is surely higher than that of a pitcher, the gap was not big enough to make an impact when examining team-averaged ISOs.

REFERENCES

1. http://www.espn.com/mlb/topics/_/page/the-steroids-era

2. http://www.baseballssteroidera.com/steroid-era-timeline-text.htm

3. http://www.fangraphs.com/library/offense/iso/

4. Singer JD. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Education and Behavioral Sciences*. 1998;23:323-355.

TABLES/FIGURES

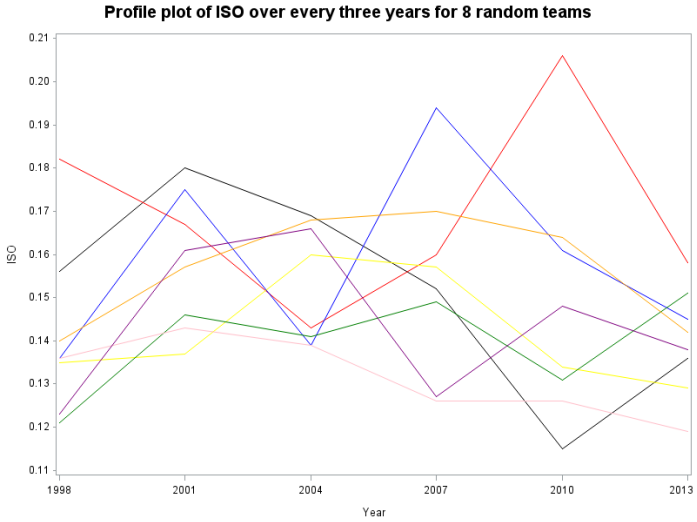

Figure 1: Profile plot of eight random teams.

Group	**Number of teams per group***	**ISO Mean (St. dev.)**	**T statistic (p-value) (degrees of freedom)****
Overall	30	.153 (.020)	n/a
AL	14	.156 (.021)	T=1.22 (p=.22) (DF=172)
NL	15	.152 (.019)	
High salary (=> $69,058,116)	15	.159 (.020)	T=-3.58 (p<.001) (DF=178)
Low salary (<$69,058,116)	15	.148 (.019)	
Neutral/hitter-friendly park (PF => 100)	16	.158 (.020)	T=-2.95 (p=.0037) (DF=178)
Pitcher-friendly park (PF < 100)	14	.149 (.020)	

* With 6 measurements over time for each team
** From 2-sample T Test assuming equal variances
Table 1: ISO comparisons by covariate group.

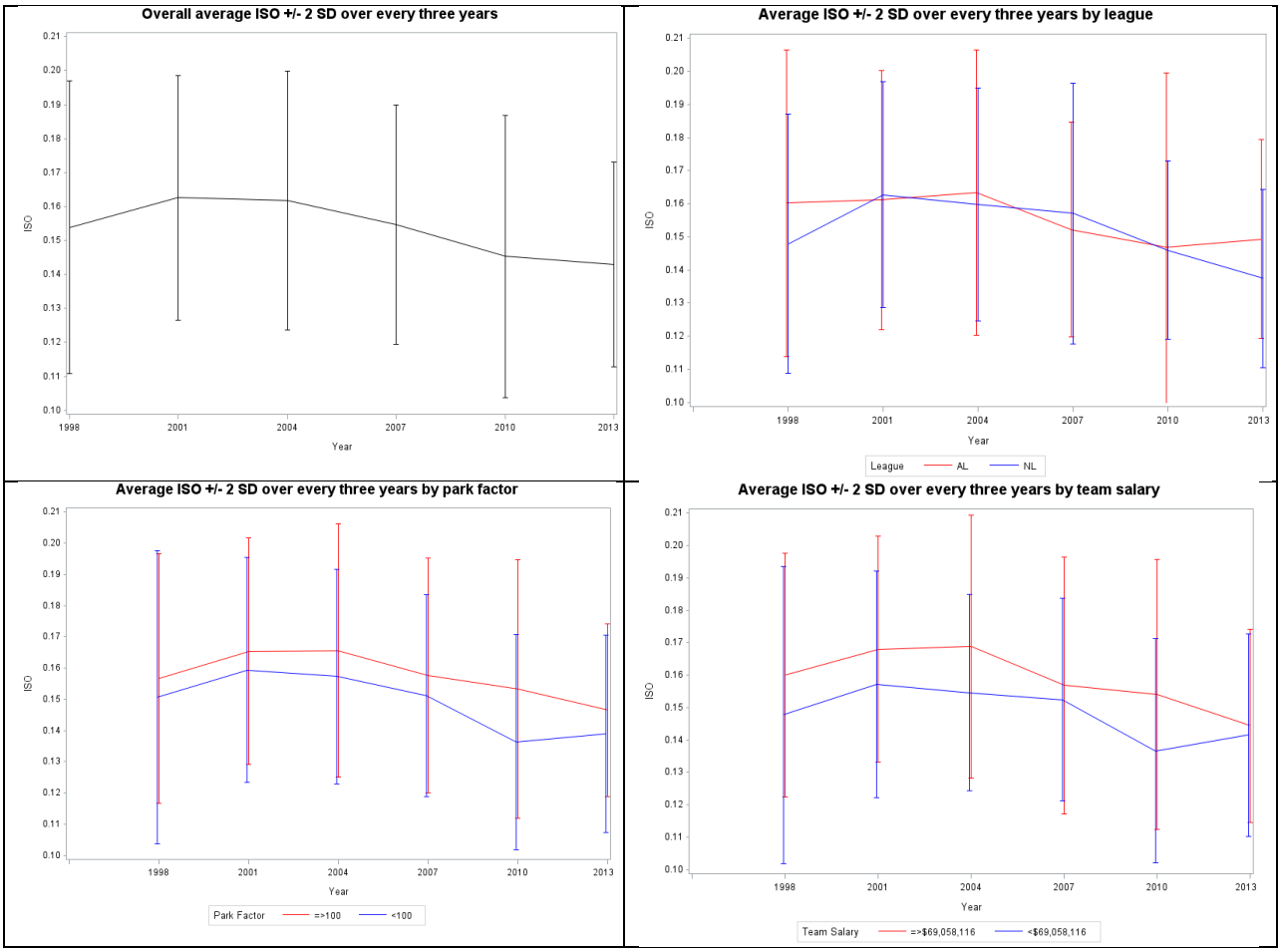

Figure 2: Empirical summary plots of ISO over time.

Year	1998	2001	2004	2007	2010	2013
1998	1.00	0.35	0.35	-0.05	0.097	0.42
2001	.	1.00	0.52	0.11	0.22	-0.084
2004	.	.	1.00	0.15	0.24	0.089
2007	.	.	.	1.00	0.46	-0.00010
2010	.	.	.	.	1.00	0.080
2013	.	.	.	.	.	1.00

Table 2: Correlation matrix of ISO over time.

Model	-2 log likelihood	AICC	BIC
Unstructured means	-939.6	-866.6	-838.5
Linear	-928.5	-866.9	-840.9
Linear spline (knot at 2004)	-936.3	-871.9	-845.3

Table 3: Comparison of mean structures. The linear spline structure was selected.

Covariance structure	Number of parameters	-2 log likelihood	AICC	BIC
Unstructured	21	-936.3	-871.9	-845.3
Variance components	1	-911.8	-897.2	-888.3
Compound symmetry (random intercept)	2	-915.9	-899.0	-888.9
Heterogeneous compound symmetry	7	-918.8	-890.5	-875.0
Autoregressive(1)	2	-918.6	-901.7	-891.6
Heterogeneous AR(1)	7	-921.4	-893.2	-877.7
Toeplitz	6	-924.1	-898.2	-883.7
Heterogeneous TOEP	11	-927.2	-889.3	-870.0

Table 4: Comparison of covariance structures. The autoregressive(1) structure was selected.

Covariates	Estimate	Standard error	T statistic (p-value)
Model 1			
Intercept	0.15	0.0052	29.28 (<.001)
Year	0.0035	0.0022	1.58 (.12)
Year_3	-0.010	0.0033	-3.15 (.002)
Model 2			
Intercept	0.16	0.0058	27.06 (<.001)
Year	0.0035	0.0022	1.58 (.12)
Year_3	-0.010	0.0033	-3.11 (.002)
League=AL	0.0026	0.0034	0.75 (.46)
Park factor=low (binary)	-0.0061	0.0035	-1.74 (.09)
Salary=low (binary)	-0.0085	0.0035	-2.47 (.02)
Model 3			
Intercept	0.16	0.0053	29.71 (<.001)
Year	0.0035	0.0022	1.61 (.11)
Year_3	-0.010	0.0032	-3.26 (.001)
Park factor=low (binary)	-0.0062	0.0034	-1.83 (.08)
Salary=low (binary)	-0.0084	0.0034	-2.49 (.02)
Model 4			
Intercept	0.022	0.035	0.63 (.53)
Year	0.0034	0.0021	1.63 (.11)
Year_3	-0.010	0.0031	-3.32 (.001)
Park factor (cont.)	0.0012	0.00034	3.46 (.002)
Salary/$1,000,000 (cont.)	0.00018	0.000063	2.87 (.008)

Table 5: Examining different covariates using the linear spline model with autoregressive(1) covariance.