EVALUATION AND RANKING OF MINOR-LEAGUE HITTERS USING A STATISTICAL MODEL

by

GARY BRENT JOHNSON

B.A., Doane College, 1999

A REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2006

Approved by:

Major Professor
Thomas M. Loughin

ABSTRACT

Traditionally, major-league scouts have evaluated young "position players," those who are not pitchers, using the "Five Tools": hitting for average, hitting for power, running, throwing, and fielding. However, "sabermetricians," those who study the science of baseball, e.g. Bill James, have been trying to evaluate position players using quantifiable measures of performance. In this study, a factor analysis was used to determine underlying characteristics of minor-league hitters. The underlying factors were determined to be slugging ability, lead-off hitting ability, "patience" at the plate, and pure-hitting ability. Additionally, an ordinal response was created from the number of at-bats and on-base plus slugging percentage in the majors during the 2002-05 seasons. The underlying characteristics along with other variables such as a player's age, position, and level in the minors are used in a cumulative logit logistic regression model to predict a player's probability of notable success in the majors. The model is built upon data from the 2002 minor-league season and data from the 2002, 2003, 2004, and 2005 major-league seasons.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to especially thank Dr. Tom Loughin, my academic advisor, for his help and support on this research. His genuine love for sports and statistics invigorated me. The quality and extensiveness of this work were continually improved with his guidance.

I would also like to thank my committee members, Dr. Paul Nelson and Dr. Dallas Johnson. Their comments and suggestions always invite me to take a new perspective.

Lastly, I would like to thank the rest of the Kansas State University Department of Statistics faculty. Their help through coursework and instruction was invaluable to my work.

Many, many thanks!!

I. Introduction

...By the time Billy [Beane] was fourteen, he was six inches taller than his father and doing things that his father's books failed to describe. As a freshman in high school he was brought up by his coach, over the angry objections of the older players, to pitch the last varsity game of the season. He threw a shutout with ten strikeouts, and went two for four at the plate. As a fifteen-year-old sophomore, he hit over .500 in one of the toughest high school baseball leagues in the country. By his junior year he was six foot four, 180 pounds and still growing, and his high school diamond was infested with major league scouts, who watched him hit over .500 again. In the first big game after Billy had come to the scouts' attention, Billy pitched a two-hitter, stole four bases, and hit three triples...

...He encouraged strong feelings in the older men who were paid to imagine what kind of pro ballplayer a young man might become. The boy had a body you could dream on. Ramrod-straight and lean but not so lean you couldn't imagine him filling out. And that face! Beneath an unruly mop of dark brown hair the boy had the sharp features the scouts loved. Some of the scouts still believed they could tell by the structure of a young man's face not only his character but his future in pro ball. They had a phrase they used: "the Good Face." Billy had the Good Face...

...They all missed the clues. They didn't notice, for instance, that Billy's batting average collapsed from over .500 in his junior year to just over .300 in his senior year. It was hard to say why. Maybe it was the pressure of the scouts. Maybe it was that the other teams found different ways to pitch to him, and Billy failed to adapt. Or maybe it was plain bad luck. The point is: no one even noticed the drop-off. 'I never looked at a single statistic of Billy's,' admits one of the scouts. 'It wouldn't have crossed my mind. Billy was a five-tool guy. He had it all.' Roger Jongewaard, the Mets' head scout, says, 'You have to understand: we don't just look at performance. We were looking at talent.' But in Billy's case, talent was a mask. Things went so well for him so often that no one ever needed to worry about how he behaved when they didn't go well. Blalock [his coach] worried, though. Blalock lived with it. The moment Billy failed, he went looking for something to break. One time after Billy struck out, he whacked his aluminum bat against a wall with such violence that he bent it at a right angle. The next time he came to the plate he was still so furious with himself that he

insisted on hitting with the crooked bat. Another time he threw such a tantrum that Blalock tossed him off the team. 'You have some guys that when they strike out and come back to the bench all the other guys move down to the other end of the bench,' says Blalock. 'That was Billy.'

...Billy could run and Billy could throw and Billy could catch and Billy even had the presence of mind in the field. Billy was quick-witted and charming and perceptive about other people, if not about himself. He had a bravado, increasingly false, that no one in a fifty-mile radius was ever going to see through. He looked more like a superstar than any actual superstar. He was a natural leader of young men. Billy's weakness was simple: he couldn't hit...

...In his last three and a half years of pro ball Billy watched a lot more baseball than he played, and demonstrated an odd knack for being near the center of other people's action. 'The Forrest Gump of baseball,' he later called himself. He was on the bench when the Twins won the 1987 World Series and also when the A's won the 1989 World Series. He was forever finding himself next to people who were about to become stars. He'd played outfield with Lenny Dykstra and Darryl Strawberry. He'd subbed for Mark McGwire and Jose Canseco. He'd lockered beside Rickey Henderson. In his slivers of five years in the big leagues he played for four famous managers: Sparky Andersen, Tom Kelly, Davey Johnson, and Tony La Russa. But by the end of 1989 his career stat line (301 at bats, .219 batting average, .246 on-base percentage, .296 slugging percentage, and 11 walks against 80 strikeouts) told an eloquent tale of suffering. You didn't need to know Billy Beane at all – you only needed to read his stats – to sense that he left every on-deck circle in trouble. That he had developed neither discipline nor composure. That he had never learned to lay off a bad pitch. That he was easily fooled. That, fooled so often, he came to expect that he would be fooled. That he hit with fear. That his fear masqueraded as aggression. That the aggression enabled him to exit the batter's box as quickly as possible. One season in the big leagues he came to the plate seventy-nine times and failed to draw a single walk. Not many players do that...(Lewis, 2003)

The above passage is an excerpt from the book, Moneyball: The Art of Winning an Unfair Game. This book wonderfully illustrates the why and how of a new approach taken by some to evaluate baseball players. Traditionally, major-league scouts have

evaluated young "position players," those who are not pitchers, using the "Five Tools": hitting for average, hitting for power, running, throwing, and fielding. They evaluated the players while watching them in person, during either games or try-outs. However, there are many stories like the one of Billy Beane, and there is a problem with this. Beane was signed by the Mets in 1980 for $125,000; Luis Montanez, who some scouts compared to Shawon Dunston, received $2.75 million from the Chicago Cubs as their number one draft pick in 2000 according to baseballprospectus.com, and he has yet to play a game in a Cubs uniform. Professional baseball is a very expensive industry. Teams cannot afford to make too many mistakes. They need an efficient, trustworthy method for evaluating young players.

Statisticians try to encourage researchers to make data-based decisions. It should be considered that position players can be evaluated using quantifiable measures of performance. Some "sabermetricians," those who study the science of baseball, have been trying to do that or at least something similar for many years. (Sabermetrics is a term originating from the organization SABR, or the Society for American Baseball Research.) Bill James, often known as the father of sabermetrics, has, for example, created a statistic called runs-created. Runs-created is used to predict the number of runs a team will score based on offensive statistics like hits, walks, and total bases. It is also often used to estimate how many runs an individual player can "create" for his team. Other sabermetric measures of offensive production such as on-base percentage plus slugging percentage (OPS) have already become commonly used by players, managers, and media.

I hypothesize that the offensive performance of prospective minor-league position players can be evaluated by a statistical model in such a way as to predict their performance in the major leagues. In order to assess this hypothesis, I will attempt to build a statistical model, using offensive statistics such as hits, homeruns, walks, and strikeouts that both accurately and precisely predicts offensive production in the major-leagues.

An overview of the approach to be taken follows:

1. Gather data on prospective minor-league position players' performance
2. Observe their major-league performance, if any
3. Model some measure(s) of their major-league performance against measures of their minor-league performance
4. Make conclusions about significant predictors and final model
5. Discuss ideals and reality regarding collected data and scope of inference

II. Data Collection

II.1 Brief Discussion of the Minor Leagues

In order from least to most advanced, the minor leagues consist of the rookie, advanced rookie, short-season A, A, A+, AA, and AAA leagues. The teams in these leagues are independently owned and operated, but they are directly affiliated with a major-league organization. There are also independent leagues, but they have no affiliation with the Major League Baseball. Mike Blake discusses each of these leagues in his book, The Minor Leagues (Blake, 1991). Further discussion of the minor leagues can be found on

the Wikipedia website.[1] The purpose of these leagues is to develop young players so that one day they may be ready to be called up by a major-league affiliate.

The rookie, advanced rookie, and short-season A leagues are all short-season leagues taking place from June through September. They consist mostly of players recently drafted out of high school and some out of college. These players are young enough that they may never have lived away from home or made decisions without the influence of their parents. A large portion of these young players are still honing their basic life skills, as well as their baseball skills. Their future success may not yet be predictable with any desired level of accuracy.

The single-A and double-A leagues are the levels at which most serious player evaluation takes place. The single-A league consists of players moving up from the short-season leagues, some high first-round draftees (particularly those with college experience), and possibly very successful players from foreign rookie leagues. For many of these players, this is a second or third promotion in the minors. They are commonly trying to work on control as pitchers and consistency as hitters. The double-A league is often the level from which players are called up to the majors. This is a level where a few small remaining player faults are being corrected, and the level of competition is quite good.

Lastly, there is the triple-A league. This league has a very interesting mix of players. Some are major-league players getting in some practice while rehabilitating an injury (technically, these players may be found at any long-season league). Some are major-league-caliber players getting one last evaluation before being called up. And, some are

[1] Wikipedia's URL is http://en.wikipedia.org/wiki/Minor_league_baseball.

players at or near major-league quality, for whom the major-league team doesn't have an available roster spot. It may be that an all-star or very good player already plays their position, or the minor-league player's defense is not good enough to fill an open position, or a combination of both. (Only 25 players may be a part of the major-league club until September 1st when this is expanded to 40.) It is not uncommon for poor-fielding sluggers to get "stuck" at the triple-A level for the remainder of their careers, thus becoming what some call a "quadruple-A player." This issue will become relevant later when a *minor-league prospect* is defined.

So, a general profile of the players in the minor leagues is

- Short-season league players are very young and their skills very raw, and thus, the leagues are "speculative" and filled with players of unknown potential
- Single-A players are potentially good players working on control as pitchers and consistency as hitters
- Double-A players are very good players often called directly up to the majors
- Triple-A players are very good players at or near major-league quality, for whom often there is not an open roster spot, or the defense required to play their position is lacking; some are major-leaguers rehabilitating an injury

So, many players in the single-A, double-A, and triple-A leagues have at least some potential to become major-leaguers. And therefore, these leagues are the focus of the data collection.

II.2 Data Sources

The following sources were scoured for player-performance statistics from the single-A, double-A, and triple-A leagues during the 1994-98 seasons. (Collecting data from this time period would give the prospects more than ample opportunity to progress and play in the majors, if they were ever going to do so.)

- John Sickels, noted prospect analyst and author of The Baseball Prospect Book
- www.minorleaguebaseball.com
- www.baseballamerica.com
- www.usatoday.com
- sportsillustrated.cnn.com
- www.thebaseballcube.com
- www.baseball1.com
- www.baseball-reference.com
- www.baseball-links.com – This site contained many links, into which I checked.
- individual team websites such as the Durham Bulls, Wichita Wranglers, and Fort Myers Miracle
- groups.yahoo.com/group/baseball-databank

The last source in the above list is a large webgroup (over 700 members) of sabermetricians, both amateur and professional, and simply fans of baseball statistics.

No response was received from John Sickels. Minorleaguebaseball.com, Baseball America, USA Today, Sports Illustrated, and the individual teams had statistics for the current season, 2005. Baseball1 and Baseball-Reference had major-league statistics only. The baseball-databank webgroup was also in search of a source of minor league statistics. The following message was received from someone at The Baseball Cube:

> "Basically, I have statistics during this time period [1994-98] for players who have major league experience and minor leaguers who were active in 2002 or later. I have 6582 records for batting statistics and 6690 records for pitching statistics." - Anonymous

This data would not have been ideal because these data contain only those minor-leaguers from 1994-98 who made it to the majors. They represent a biased sample of minor-leaguers. Those players who washed out between 1994 and 2002 would not have been included. There would have been no data with which to distinguish those who achieved future success from those who didn't.

In the end, a workable-but-not-ideal data set was obtained: complete offensive minor-league statistics were found for the 2002 season from the baseball-databank webgroup, as well as major-league offensive statistics from the 2002, 2003, 2004, and 2005 seasons. This one-year minor-league sample limits the ability to obtain precise predictions in the case of some players. A player may have one really good or really bad season in the minors due to little more than chance or injury. It is not possible to take an average over multiple minor-league seasons in order to reduce these effects. Also, because the sample is from only three years ago, many of the players have not yet had adequate time to reach their full potential, and hence it is unknown whether such players might still become productive major-leaguers or even all-stars. This forces a change in the research question I will pursue. It is now, "What offensive statistics are indicators of major-league success in the next three seasons?" If primary interests are short-term, then this question is relevant. Or if one can assume that indicators of long-term success are no different from those of short-term success, then the original hypothesis can still be tested. There is no way to check this, however.

II.3 Defining a Prospect

One last step must be taken in order to create a working dataset. A minor-league prospect must be defined. Not all minor league players are prospects on the way up. There are the quadruple-A players as discussed in Section II.1, as well as those players performing rehabilitation stints. Researchers have shown that the production of baseball players tends to peak around 28 years of age (Krohn, 1983). Because of this, "prospects" are players who have advanced through the minor leagues at a reasonable rate and may therefore spend the years surrounding their peak level in the major leagues. The following age restrictions on "prospects" have been imposed:

- A prospect in A must be no older than 22.
- A prospect in A+ must be no older than 23.
- A prospect in AA must be no older than 24.
- A prospect in AAA must be no older than 25.

If a prospect advances one level each year, the minimum rate a true prospect should advance, this will place him in the majors at least two years before the average age at which players reach their peak performance. A looser age restriction was considered, but this included several players who were explicitly classified as too old for the competition at that level by a popular minor-league player evaluator (Sickels, 2003).

Lastly, some players meet the age requirements for a "prospect" but already have substantial major-league experience. Sickels (2003) no longer considers these players "prospects." His rule is that if a player has played in at least 50 major-league games by the end of a season, that player is no longer a "minor-league prospect." The rule adopted

for the present study is players who have played in at least 50 major-league games prior to a particular minor-league season are no longer "minor-league prospects."

III. Analysis Methods and Definitions

III.1 Overview of Approach

The following steps are taken to predict major-league performance from minor-league data:

1. Define a measurement of major-league "success" or "performance"
2. Use a factor analysis on numerous minor-league hitting statistics to identify independent, underlying factors common to all offensive categories
3. Build logistic regression model providing an adequate fit to the data, and rank players based on predicted probability of success

III.2 Determine Measurement of Major-League Success

In order to determine key indicators of success in the major leagues, a measurement of success is needed. This measurement should take into account both production and longevity. (The word 'longevity' is used loosely in this case because there is a 3-year limit on the length of time a player can have played by the end of the 2005 season.) On-base percentage plus slugging is a currently accepted standard for measuring productivity because it measures both primary components of hitting: a player's ability to get on base and his ability to drive other players around the bases with extra-base hits. (This hitting

statistic is defined in the Glossary.) At-bats are a standard measure of longevity. Figure 1 shows how these two statistics are used to define our discrete response for success. If a player has not gotten a major-league at-bat, then his response for the variable, Major League Baseball (MLB) success, is a zero. As a player's production increases over an increased number of at-bats, the value of this response also increases until reaching a maximum response of four.

The idea behind the discrete variable, MLB success, came from an unknown author, who wrote an article on the website, www.birdsinthebelfry.com, comparing the rates of success of baseball amateur draftees drafted out of high school vs. college. The author used the following subjective responses to evaluate the players:

0 - Player never reached the major leagues
1 - Player had a "cup of coffee" in the majors
2 - Player was a major league "journeyman"
3 - Player was a starting position player
4 - Player was a star at the major league level

The quantitative responses for MLB success were an attempt to imitate these categorical rankings based on our limited observation period.

Values for the Response Variable "MLB Success"

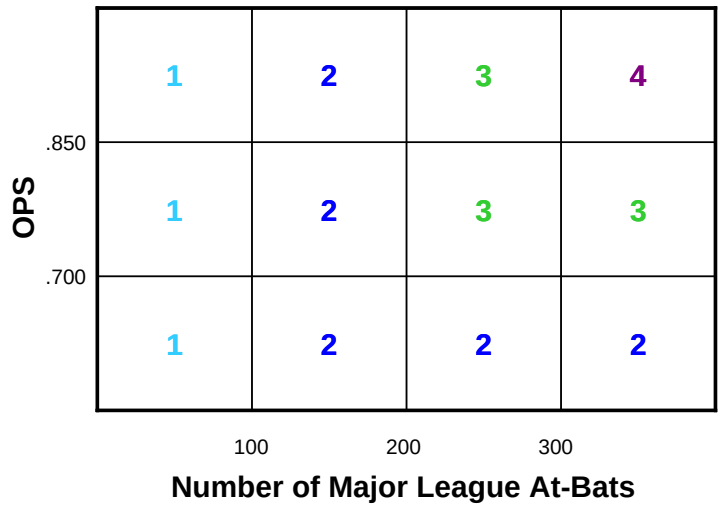

Figure 1

Recall that the prospects being researched have had only three seasons since the 2002 minor-league season, in which to observe major-league performance. The cutoffs defining longevity have been defined accordingly using my best judgment. The cutoffs defining level of production can also be reasoned using the following anecdotal evidence. Consider the Milwaukee Brewers for example. All nine of their starting position players posted an OPS of at least .700 during the 2005 season. One starter, Geoff Jenkins, posted an OPS of .888, which is at an "all-star level." He was not an all-star in 2005, as fellow teammate Carlos Lee was, but Jenkins was an all-star in 2003 when his OPS was .913. Consider the 2005 World Series Champ, the Chicago White Sox, as another example. Again, eight of the nine starting position players posted an OPS of at least .700. (Scott Podsednik's OPS was .699.) Paul Konerko, whose OPS was .911, was an all-star. Ironically, Podsednik was also an all-star because his on-base percentage was good, he stole 44 bases while only being caught 9 times prior to the all-star break (the next best in

the AL was 27), and he plays good defense in left field. I am speculating that the findings will be similar with most major-league teams. Thus, these discrete rankings of major-league success are reasonable.

III.3 Preliminary Analyses of Success Rate of Prospects

It would be beneficial to determine the chance that a prospect gets at least a taste of the majors in the next three seasons, as well as to determine whether or not this probability changes significantly given certain factors. The probability of a prospect getting at least a taste of the majors in the next three seasons can be estimated by the sample proportion of players who have at least one at-bat in the majors out of all prospects in 2002. (Pinch-runners who appear in major-league games but have no at-bats are excluded, but the number of pinch-runners is thought to be small enough so as to affect the estimation only slightly.) Among the factors that may be associated with whether or not a player plays in the majors in the next three seasons are the player's position and the highest minor-league level played in 2002.

Bill James (James, 1985) discusses what he calls the "defensive spectrum." This spectrum places the defensive position most easily played at the left end and the most difficult at the right end, as shown below.

1B ➔ LF ➔ RF ➔ 3B ➔ CF ➔ 2B ➔ SS

The catcher, being so different from all the other positions, is not included in the spectrum. However, in the data set to be analyzed, the center, left, and right field positions are named more generally as outfield. So, the defensive spectrum used in this research is as shown below.

1B ➔ OF ➔ 3B ➔ 2B ➔ SS

In this light, position (excluding catcher) can be perceived as an ordinal variable. The levels of MLB success are also ordinal. Correlation between these two ordinal categorical variables can be examined. As the level of defensive difficulty increases, do responses on the level of MLB success tend to increase, or stay about the same?
The anticipated direction of the relationship is not clear, nor is it clear that it must be monotone. It may be that shortstops have a higher rate of MLB success because their position is so difficult to play. They may get called up to the majors even if their hitting abilities are not strong. It may also be that first-basemen have a higher rate of MLB success because they tend to be good hitters, and their position is not as demanding. In this case, it would be useful to treat position (including catchers) as a nominal variable, and examine whether or not the row-mean scores for success differ between positions.

The other possible factor that will be considered, minor-league level played at in 2002, can also be perceived as an ordinal variable. Thus, the correlation between it and level of MLB success can also be examined. It would be expected that players who play at a high

level in 2002 should have a relatively higher probability of at least making it to the majors.

These analyses are marginal analyses not accounting for the possible confounding effects of performance in the minors. But, they may lead to indications of potential predictors in the final modeling of major-league success.

III.4 Exploration of Trends across Levels of MLB Success

An exploratory step in the analysis is to identify which offensive variables exhibit trends across increasing levels of MLB success. The offensive variables that are considered follow (all statistics are defined in the glossary):

- Singles
- Doubles
- Triples
- Homeruns
- Homeruns per At-bat
- Strikeouts
- Walks
- Strikeout-to-walk ratio
- On-base percentage (OBP)
- Isolated power (Isopower)
- On-base percentage plus slugging percentage (OPS)
- Runs created

An analysis of variance F-test is performed in order to determine for which offensive variables the mean responses differ significantly for at least one level of MLB success. In addition, a series of contrasts tests whether or not significant patterns exist for any

offensive variables across levels of MLB success. The contrasts and their respective contrast coefficients follow:

Contrast	Description	Coefficients
Linear	Test for linearly increasing or decreasing trend	-2 -1 0 1 2
Quadratic	Test for increasing and later decreasing pattern or vice versa	-2 1 2 1 -2
0 vs. 1-4	Test if group 0 mean is different from the mean of groups 1-4	-4 1 1 1 1
Linear across 1-4	Test for linearly increasing or decreasing trend across groups 1-4	0 -3 -1 1 3

III.5 Factor Analysis

III.5.1 The Objectives

In a study with many variables being measured on each player, it is easy to believe that these variables are related to one another in many different ways. A factor analysis (FA) is therefore a reasonable next step in the analysis. An objective of a FA according to Johnson (1998) is to use a set of variables to derive a new set of uncorrelated variables, called *underlying factors*, with the hope that these new variables will be few in number and give a better understanding of the data being analyzed. These new variables can then be used in future analyses of the data. In order for the underlying factors to give a better understanding of the data, reasonable interpretations of them must exist. A solid

understanding of baseball ought to give these interpretations. But, most importantly, the new variables will be independent of one another. Hence, they may be used together in a regression model, for example, and their effects on responses can be interpreted more cleanly.

III.5.2 The Model and Assumptions

Johnson (1998) discusses all that follows. Consider a p-variate response vector $\mathbf{x}$ from a population that has mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. In this research, the response vector consists of the hitting statistics of interest listed in section III.3. The general FA model assumes there are m underlying factors ($m < p$) denoted by f_1, f_2, ..., f_m such that

$$x_j = \mu_j + \lambda_{j1} f_1 + \lambda_{j2} f_2 + \ldots + \lambda_{jm} f_m + \eta_j \quad \text{for } j = 1,2,\ldots,p \tag{3.5.1}$$

In the preceding model, we assume that

1 the f_k's are independently and identically distributed with mean 0 and variance 1 for $k = 1, 2, ..., m$;
2 the η_j's are independently distributed with mean 0 and variance ψ_j for $j = 1, 2, ..., p$; and
3 f_k and η_j have independent distributions for all combinations of k and j, $k = 1, 2, ..., m$ and $j = 1,2, ..., p$.

The variables f_1, f_2, ..., f_m are the newly created underlying factors called *common factors* because they are common to all p original hitting statistics. The unknown parameters η_1, η_2, ..., η_p are called *specific factors* because they describe the residual effect due to the jth hitting statistic. Lastly, ψ_j is called the specific variance of the jth

response variable because it describes the player-to-player variation specific to the *j*th hitting statistic. So, the response to the *j*th hitting statistic can be thought of as a function of its overall mean μ, the *m* underlying factors common to all *p* hitting statistics, and some residual error due to player-to-player differences. The multipliers, the λ_{jk}'s, are called *factor loadings* of the *j*th hitting statistic on the *k*th factor. These factor loadings measure the contribution of the *k*th common factor to the *j*th hitting statistic.

III.5.3 Factor Analysis Equations

It must be determined if $\mathbf{f}$, $\mathbf{\Lambda}$, and $\boldsymbol{\eta}$ exist such that $\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{\eta}$, which is Equation (3.5.1) in matrix form where $\boldsymbol{\mu}$ is equal to zero. (The reason for $\boldsymbol{\mu}$ being set to zero will be made clear later in this section.) First, it can be noted that $\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{\eta}$ implies that

$$\begin{aligned} \mathbf{\Sigma} &= \text{Cov}(\mathbf{x}) \\ &= \text{Cov}(\mathbf{\Lambda f} + \boldsymbol{\eta}) \\ &= \mathbf{\Lambda} \cdot \text{Cov}(\mathbf{f}) \cdot \mathbf{\Lambda}' + \mathbf{\Psi} \\ &= \mathbf{\Lambda I \Lambda}' + \mathbf{\Psi} \\ &= \mathbf{\Lambda \Lambda}' + \mathbf{\Psi} \end{aligned}$$

So, it is easier to instead try to find $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ so that

$$\mathbf{\Sigma} = \mathbf{\Lambda \Lambda}' + \mathbf{\Psi} \qquad (3.5.2)$$

The relationships described in Eq. (3.5.2) are called the *factor analysis equations*. Rather than analyze $\mathbf{\Sigma}$ directly, most factor analysis procedures are applied to standardized

versions of **X**, say **Z**, and its respective correlation matrix **P**. (This is the reason for **μ** being set to zero.) This then means that **Λ** is the matrix of correlations between the z_j's (standardized hitting statistics) and the f_k's.

The factor loading matrix **Λ** is not unique. By multiplying by an orthogonal matrix **T** (called a rotation), $\mathbf{\Lambda}^* = \mathbf{T\Lambda}$ is also a loading matrix. Different rotations may yield solutions that are more easily interpreted than others. For more details, see Johnson (1998). Finally, note that the variance of x_j can be partitioned according to the amount explained by each factor, and the proportion of the variance of x_j that is explained by the common factors is called the *communality* of the *j*th hitting statistic. The communality of the *j*th hitting statistic is $\sum_{k=1}^{m} \lambda_{jk}^2$.

III.5.4 Solving the Factor Analysis Equations

Prior to trying to solve the FA equations, an estimate of the number of underlying factors *m* is needed. A good place to start is with a scree plot. This is a plot of the eigenvalues associated with *principal components*, which are new uncorrelated variables that account for as much of the variability in the data as possible, against their rank in descending order. Eigenvalues that are greater than one represent principal components that explain more variability than any of the original standardized variables. The number of eigenvalues that are greater than one, then, gives a good initial estimate of the number of underlying factors driving the values of the variables being measured. Johnson (1998) provides some rules to determine if more or fewer factors are needed.

1 Do not include trivial factors, i.e. factors that have one and only one of the original variables loading on them.

2 Many computing programs will produce matrices of differences between the observed correlations between variables and those that are reproduced by the FA solution. If these differences are quite small, you might be able to reduce the number of factors. If they are quite large (many greater than .25 and some greater than .40, perhaps), then the number of factors might need to be increased. (SAS's Proc Factor produces such matrices.)

There are a number of different methods available in SAS's Proc Factor that can be used to solve the FA equations. If the data are multivariate normal, we use the maximum likelihood method because it is generally known to be good in this situation. If the data are not multivariate normal, then we use the principal factoring method without iteration. Both methods require prior estimates of the communalities for each response variable. We use the squared multiple correlation of a variable with all the remaining variables as prior communality estimates, which is also available in Proc Factor.

Once a set of factors has been derived, a recommended next step is to rotate the factors by multiplying by an orthogonal matrix. There are many procedures for doing this. Most try to make as many factor loadings as possible near zero and maximize as many of the remaining as possible. We use the Varimax method proposed by Kaiser (1958) and recommended by Johnson (1998).

Lastly, because the newly-derived factors are to be used in subsequent analyses, a score must be assigned for each of the new variables for each hitter in the data set. A method

implemented by SAS's Proc Factor to do this is Thompson's method or the regression method. Thompson noted that for normally distributed data the joint distribution of **z** (standardized **x**) and **f** is

$$\begin{bmatrix} z \\ \mathbf{f} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{P} & \Lambda \\ \Lambda' & \mathbf{I} \end{bmatrix} \right)$$

This implies that the conditional expectation of **f** given that $\mathbf{z} = \mathbf{z}^*$ is

$$\mathrm{E}\left[\mathrm{f} \mid \mathrm{z} = \mathrm{z}^*\right] = \Lambda' \mathrm{P}^{-1} \mathrm{z}^*$$

Therefore, Thompson's method estimates the vector of factor scores for the *r*th individual as $\mathrm{f}_r = \hat{\Lambda}' \mathrm{R}^{-1} \mathrm{z}_r$. Thompson's method is implemented by SAS's Proc Factor.

If the assumption of normality does not hold for the data, Bartlett's method or the weighted least-squares method can be used. However, this method is not implemented by SAS's Proc Factor. Therefore, the method for computing factor scores implemented in the case of non-normality in this research is an ad-hoc mentioned by Johnson (1998). This method computes the factor scores by taking a linear combination of the standardized responses most highly correlated with the respective factors. One downfall with this method is, however, that the factors might be moderately correlated with one another. This could lead to multi-collinearity in any response modeling.

III.6 Logistic Regression for Ordinal Data

III.6.1 The Objectives

Once a set of independent underlying factors has been determined from FA, these factors may be used in a regression model. The response that we are trying to predict is the MLB success defined in Section III.2. Therefore, methods for modeling an ordinal discrete response can be used. The goal, ultimately, is to model the probability of performing at or above each given level within the next three years. These probabilities may be used for ranking the players given the values of their factor scores determined from the FA.

III.6.2 The Model and Notation

Consider first the binary logistic regression model, as discussed by Agresti (1996). Suppose that a generic response variable Y can take on two values, a 1 for a "success" and a 0 for a "failure." The binary logistic regression model is often expressed as a linear form of the logit (log-odds) of the probability of response 1 at given values of the predictors, denoted by $\pi(\mathbf{x}) = P(Y = 1|\ \mathbf{x})$, as shown below:

$$\text{logit}[P(Y = 1|\ \mathbf{x})] = \log\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \alpha + \boldsymbol{\beta}\mathbf{x} \qquad (3.5.1)$$

where α and $\boldsymbol{\beta}$ are unknown parameters.

The quantity $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$ is called the odds of success. It can be shown that holding all other factors constant, as x_k increases by one unit, the odds of success increase by e^{β_k}. Equation (3.5.1) yields the success probability

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}\mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}\mathbf{x})} \tag{3.5.2}$$

We can next extend the relationships in Eqs. (3.5.1) and (3.5.2) to accommodate an ordinal response. This can be done several ways, each of which imposes different assumptions and structures on the model. In this research, it is done by considering the cumulative probabilities, as also discussed by Agresti (1996). First, recall that response variable of interest, MLB success, takes on the ordered values 0, 1, 2, 3, and 4. The cumulative probabilities are then as follows,

$$P(Y \le j) = \pi_0 + \ldots + \pi_j, \qquad j = 0,\ldots,3$$

The *cumulative logit* logistic regression model is then

$$\text{logit}[P(Y \le j)] = \log\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = \alpha_j + \boldsymbol{\beta}\mathbf{x}, j = 0,..., 3 \tag{3.5.3}$$

Probabilities for specific categories can be calculated as differences of the cumulative probabilities. For example,

$$P(Y = 4) = 1 - P(Y \leq 3) = 1 - \frac{\exp(\alpha_3 + \boldsymbol{\beta}\mathbf{x})}{1 + \exp(\alpha_3 + \boldsymbol{\beta}\mathbf{x})}$$

SAS's Proc Logistic computes any cumulative or individual probability that is desired.

III.6.3 Assessing the Model Fit

The cumulative logit model in (3.5.3) is a parallel-lines regression model also known as the *proportional odds model* and is the model fit by SAS's Proc Logistic. (Note that an intercept is not fit for the last response category because $P(Y \leq 4)$ must be one.) It assumes common slopes across cumulative logits. This assumption must be checked, and it must be determined which predictors provide significant information about the response, if there are any interactions between predictors present, and if there are any dispersion problems.

Hosmer and Lemeshow (2000) discuss the following steps for assessing the fit of a proportional odds model:

1. The first step in checking model fit is to remove / add any predictor variables that are found to be insignificant / significant. SAS's Proc Logistic can perform any standard variable selection method. A forward selection method is used in this research.

2. The second step is to determine if any predictors from step 1 have a significant *non-linear* relationship with the logit of the cumulative probabilities. This can be done by

trying transformations such as squares or logs of the predictors. These transformations are tried with J-1 separate binary regressions of $y \leq j$ versus $y > j$.

3. The third step is to check for any significant interactions. An interaction would indicate that the effect that one predictor has on the response changes across levels of another predictor. This can be done by adding cross-products of the predictors to the model.

4. The fourth step is to validate the assumption of proportional odds. SAS's Proc Logistic produces a score test for the proportional odds assumption. If the test is found to be non-significant, this is often considered as sufficient evidence that the proportional odds assumption is appropriate. If the test is found to be significant, the fix that will be considered is to fit either a non-proportional odds model or a partial proportional odds model.

   The non-proportional and partial proportional odds models loosen the constraints of the proportional odds model by no longer assuming that all predictors have common coefficients across the response logits. If there is evidence that all predictors have coefficients that significantly differ across response logits, then the non-proportional odds model is used. If at least one but not all predictors have coefficients that significantly differ across response logits, then the partial proportional odds model is chosen. Stokes, Davis, and Koch (2000) discuss fitting these models using SAS's Proc Genmod.

5. The fifth step is to assess the deviance about the model. If the deviance statistic is large compared to its degrees of freedom, then that suggests poor model fit or over-dispersion is a problem. What is large? The deviance statistic should follow approximately a chi-square distribution with $N - (p + 1)$ degrees of freedom, where N is the sample size times the number of cumulative logits minus one and p is the number of predictors. The standard error is $\sqrt{2df}$. So, a possible interpretation is that if the deviance statistic is more than two standard errors larger than expected, that suggests there may be a problem. If it's more than three standard errors, there is likely a problem. (This asymptotic approximation isn't technically valid unless **x** is discrete with a fixed, finite number of categories. Thus, this "rule" is not to be followed too strictly.)

   Over-dispersion may be explained by players being related with respect to something left unmeasured (for example playing for the same team). This violates the assumption of independence among players' responses. In this case, no remedy is possible based on the data available. Other explanations may be misspecification of the link function or players poorly fit by the model. In this research, the probit and complimentary log-log links are used as alternatives if a poor fit is observed. Also, diagnostics are done to identify any players poorly fit by the model. These players are removed from the analysis and duly noted.

IV. Results

IV.1 Success Rate of Prospects

How likely is it that a minor-league prospect is able to get at least a taste of the majors within three years? Table 1 displays the rate of prospects from 2002 achieving each of the levels of major-league success by the 2005 season. Recall that the levels of major-league success try to mimic the following to the extent of the limited observation time:

0 - Player has not yet reached the major leagues
1 - Player has had a "cup of coffee" in the majors
2 - Player is a major league "journeyman"
3 - Player is a starting position player
4 - Player is a star at the major league level

Less than one-third (28.69%) of 2002 prospects were able to get at least one at-bat in the majors by the 2005 season. Less than 1% played like all-stars in the majors by that time.

Table 1: Rate of Success of 2002 Prospects

MLB Success	Frequency	Percent
0	701	71.31
1	104	10.58
2	90	9.16
3	82	8.34
4	6	0.61

How is that success rate affected by the player's position or by the highest level of the minors the player played at in 2002? A player's position does not have a significant

relationship with their level of major-league success at the .05 level. Table 2 gives the summary statistics for the tests of association between position and major-league success. The row-mean-scores-differ test statistic tests whether or not at least one position is significantly more or less likely to succeed in the majors. (This test is marginally significant. It does appear that shortstops have a slightly better chance of getting to the majors.) The non-zero correlation test statistic tests whether or not with increasing difficulty in position played, success in the major leagues becomes more or less likely. This test is not significant.

Table 2: Summary Statistics for Tests of Association between Position and MLB Success

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	**Alternative Hypothesis**	**DF**	**Value**	**Prob**
1	**Nonzero Correlation**	1	0.6839	0.4082
2	**Row Mean Scores Differ**	5	10.3934	0.0648

Table 3 gives the summary statistic for the test of association between the highest level of the minors at which a player played and their level of major-league success. This test along with Table 4 indicates that players in lower levels of the minors are not as likely to achieve major-league success in only three years thereafter.

Table 3: Summary Statistic for Test of Association between Highest Level of Minors Played and MLB Success

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	**Alternative Hypothesis**	**DF**	**Value**	**Prob**
1	**Nonzero Correlation**	1	107.0268	<.0001

Table 4: Cross-Tabulation of Major-League Success and Highest Level Played in 2002

MLB Success	Level					
Frequency Col Pct	A	A+	AA	AAA	Majors	Total
0	214 88.43	230 82.44	168 69.42	89 66.92	0 0.00	701
1	12 4.96	19 6.81	31 12.81	22 16.54	20 22.99	104
2	7 2.89	16 5.73	22 9.09	9 6.77	36 41.38	90
3	7 2.89	13 4.66	19 7.85	13 9.77	30 34.48	82
4	2 0.83	1 0.36	2 0.83	0 0.00	1 1.15	6
Total	242	279	242	133	87	983

IV.2 Identification of Trends across Levels of MLB Success

Several hitting statistics also have a marginal relationship with major-league success. (All hitting statistics are defined in the Glossary.) Table 5 gives the observed significance level to different contrasts for all hitting statistics under investigation. (In the case of a linear contrast, a '+' indicates that the trend is increasing while a '-'indicates that the trend is decreasing.) Not all four contrasts are orthogonal. Often the linear contrast is significant because the mean response from the lowest major-league success category is so much different from the mean response for the other categories. Hence, the contrast for a linear trend across the last four major-league success categories is also tested.

The only real surprise may be the significant quadratic trend in the number of times caught stealing. I speculate this to be a result of players with a low level of major-league success don't attempt many base-steals while those players with a high level of major-league success are good base runners and don't often get caught. Those players in the middle success categories attempt more base-steals and also get caught more often. (It may also be simply a Type I error.) Otherwise, those players never making it to the majors have significantly different mean responses from players getting at least a taste of the majors for nearly every offensive statistic (only caught stealing is not statistically significant). Often those mean responses continue to increase or decrease across increasing levels of major-league success.

Table 5: Observed Significance Levels of Four Contrasts across Increasing Levels of Major-League Success

	Contrast			
	Linear	**Quadratic**	**0 vs. 1-4**	**Linear Across 1-4**
Runs	<.0001(+)	0.5653	<.0001	0.0482(+)
RBI's	<.0001(+)	0.1867	<.0001	0.0002(+)
Singles	0.0131(+)	0.2385	<.0001	0.5046
Doubles	0.0002(+)	0.3227	<.0001	0.1133
Triples	0.1969	0.2561	<.0001	0.9363
Homeruns	<.0001(+)	0.5081	<.0001	0.0019(+)
Homeruns per At-Bat	<.0001(+)	0.4235	<.0001	0.0037(+)
Strike-outs	0.0435(+)	0.2908	0.0004	0.1838
Walks	<.0001(+)	0.1403	<.0001	0.0034(+)
Strikeout-to-walk ratio	0.0559(-)	0.9782	0.0014	0.1850
Stolen Bases	0.4558	0.3512	0.0008	0.7905
Caught Stealing	0.2411	0.0272	0.1651	0.0399(-)

	Contrast			
	Linear	**Quadratic**	**0 vs. 1-4**	**Linear Across 1-4**
Batting Avg	<.0001(+)	0.3213	<.0001	0.0319(+)
On-base Pct	<.0001(+)	0.6364	<.0001	0.0013(+)
Isolated Power	<.0001(+)	0.8310	<.0001	0.0044(+)

Figures 2 and 3 demonstrate two examples of hitting statistics that have a relationship with major-league success, strikeout-to-walk ratio and on-base percentage. (Figures for the remaining hitting statistics can be found in Appendix A.) The plots display 95% confidence intervals for the mean response at each level of major-league success. The striking width of the intervals at the maximum success level is because only six players achieved that level of major-league success.

Strikeout-to-Walk Ratio

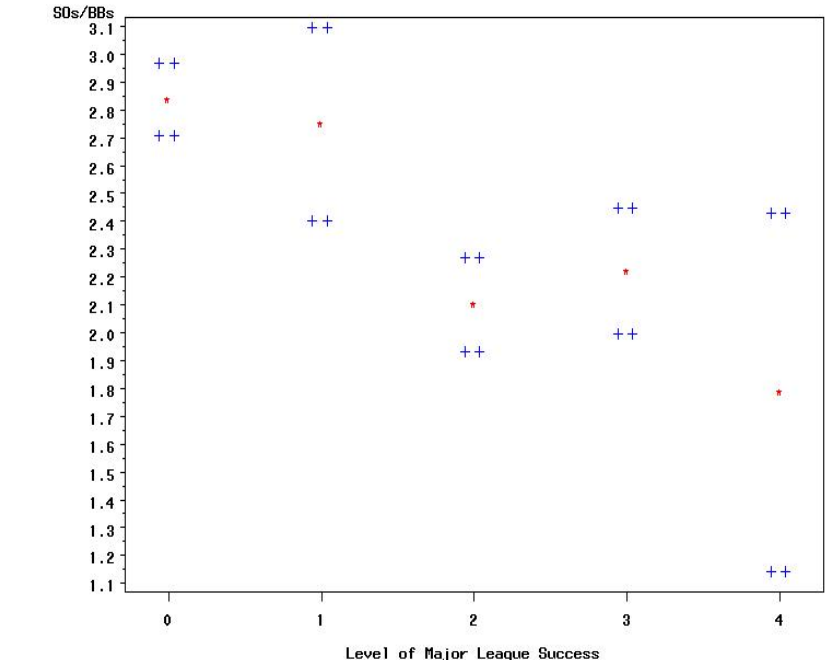

Figure 2

On-Base Percentage

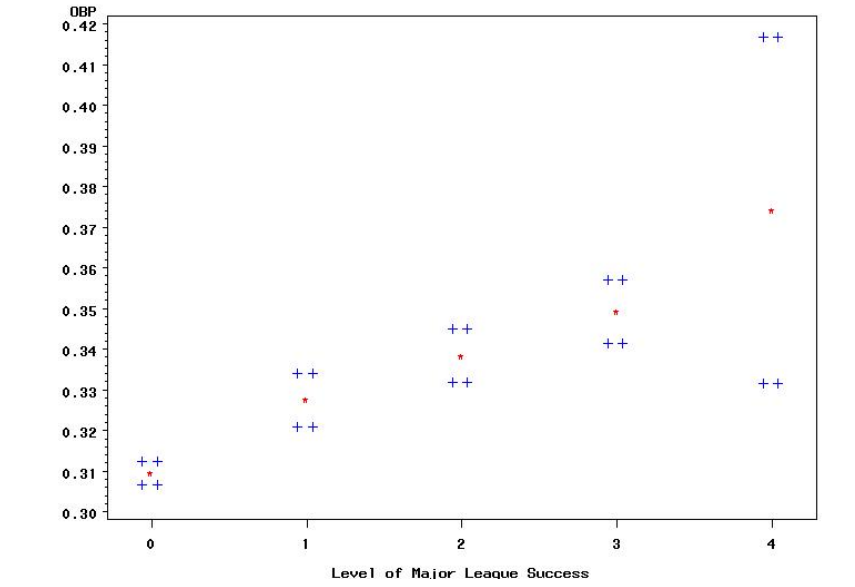

Figure 3

IV.3 Factor Analysis

After looking at trends in offensive minor-league statistics across increasing levels of major-league success, it appears that many of these offensive statistics are related to major-league success. However, modeling is easier if a smaller number of independent, underlying factors can be found. So, the next step is to perform a factor analysis.

The eigenvalues given in Table 6, which are associated with principal components, suggest that there are four underlying factors.

Table 6: Eigenvalues from a Principal Components Analysis of the Offensive Statistics

	Eigenvalue	Difference	Proportion	Cumulative
1	6.78731488	3.75251692	0.4525	0.4525
2	3.03479796	1.48785035	0.2023	0.6548
3	1.54694762	0.53961227	0.1031	0.7579
4	1.00733535	0.20041495	0.0672	0.8251
5	0.80692039	0.21690135	0.0538	0.8789
6	0.59001905	0.20925515	0.0393	0.9182
7	0.38076390	0.12892652	0.0254	0.9436
8	0.25183737	0.02811748	0.0168	0.9604
9	0.22371989	0.07491444	0.0149	0.9753
10	0.14880545	0.04247357	0.0099	0.9852
11	0.10633187	0.05439282	0.0071	0.9923
12	0.05193906	0.01084627	0.0035	0.9958
13	0.04109279	0.02552523	0.0027	0.9985
14	0.01556756	0.00896067	0.0010	0.9996
15	0.00660688		0.0004	1.0000

After extracting four underlying factors from the fifteen original variables, the residual off-diagonal correlations from the estimated correlation matrix are checked. None are greater than 0.25, and the root mean square off-diagonal residual is 0.049. These facts indicate that no more than four factors are needed, and it also suggests that three factors could be tried. However, the four factors extracted have clearly reasonable interpretations and are decided to be necessary and sufficient.

The factor analysis equations are solved using a maximum likelihood method assuming four underlying factors. The factors are then rotated using the varimax method in order to be interpreted. The rotated factor pattern is presented in Table 7. The offensive statistics strongly correlated with each respective factor are bolded. The variables strongly correlated with Factor 1 are characteristics of players with slugging abilities, where homeruns and isolated power really drive the factor. The variables strongly correlated with Factor 2 are characteristics of players with lead-off skills, where singles, runs, and stolen bases really drive the factor. The variables strongly correlated with Factor 3 are characteristics of hitters with "patience" at the plate, where on-base percentage, walks, and strikeout-to-walk ratio really drive the factor. Lastly, the variables strongly correlated with Factor 4 are characteristics of pure hitters, where batting average really drives the factor.

Table 7: Correlations between Offensive Statistics and Respective Underlying Factors after Rotation Using Varimax Method

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
Runs	0.49913	**0.72833**	0.31792	0.20200
RBI's	**0.76729**	0.39355	0.17794	0.22384
Singles	0.20727	**0.78561**	0.22498	0.35518
Doubles	**0.63267**	0.42819	0.14319	0.31459
Triples	0.15241	**0.64310**	-0.04255	0.17510
Homeruns	**0.95376**	0.05974	0.08769	0.08240
Homeruns per At-Bat	**0.90664**	-0.24413	0.03008	0.06664
Strike-outs	**0.62056**	0.50414	0.05038	-0.26281
Walks	0.37240	0.46494	**0.74438**	-0.10329
Strikeout-to-walk ratio	0.07000	-0.06907	**-0.76743**	-0.15713
Stolen Bases	-0.14133	**0.75894**	0.17187	0.01649
Caught Stealing	-0.15988	**0.74316**	0.14986	0.01742
Batting Avg	0.25126	0.26888	0.20724	**0.88097**
On-base Pct	0.25811	0.18933	**0.69713**	**0.58047**
Isolated Power	**0.90621**	-0.11583	0.00314	0.22413

In order to create responses for these new variables for current players and those in the future, scoring coefficients for each of the offensive statistics will be needed. Recall that these can be computed using Thompson's method, if the data are normally distributed. However, the normality assumption does not hold for the data. Thus, the ad-hoc method mentioned as an alternative to Thompson's is implemented. Table 8 presents the linear combinations of standardized responses used to compute the factor scores for each

player. Responses for the original variables and newly computed factor scores for a subset of players are given in Appendix B.

Table 8: Linear Combinations of Standardized Responses for Obtaining Factor Scores for Current and Future Players

Factor	Linear Combination of Standardized Responses
Factor 1	$(Z_{HR} + Z_{HR/AB} + Z_{IsoPower}) / 3$
Factor 2	$(Z_{1B} + Z_{3B} + Z_{Run} + Z_{SB} + Z_{CS}) / 5$
Factor 3	$(Z_{BB} - Z_{SO/BB} + Z_{OBP}) / 3$
Factor 4	Z_{BA}

IV.4 Logistic Regression Modeling of Ordinal Data

The four underlying factors previously computed are not highly correlated with one another and now may be used in a logistic regression model, treating the level of success in the major leagues as an ordinal response. In addition, three more predictors will be considered. Very talented players may move up the levels of the minor leagues faster than other players. For this reason, some players may be very young with respect to the other players at their level. A "bonus" should be given to these players to compensate for any negative effect on their performance due to playing against much older competition. Thus, a discrete variable called Over-His-Head, or OHH, is used, where this variable is the positive difference between the player's age and the maximum prospect age at their

respective level of the minors. Another predictor is the player's position played in the minors. The last additional factor is the highest level played by a player in the minors in 2002. Therefore, the predictors being investigated follow.

- X1: Position played (Catcher, Shortstop, First-base, Second-base, Third-base, and Outfield)
- X2: OHH (0, 1, ..., 6), discrete quantitative
- X3: Highest level played in minors in 2002 (A, A+, AA, AAA, and Majors), nominal
- X4: Factor 1, continuous
- X5: Factor 2, continuous
- X6: Factor 3, continuous
- X7: Factor 4, continuous
- All interactions between significant main factors

The response variable is Y: MLB Success (0, 1, 2, '3 or greater'). The last two categories of the response variable are pooled together because only six players fit into the last category. Stokes, Davis, and Koch (2000) note that if a non-proportional or partial proportional odds model must be fit, all response categories in a cross-tabulation of variables should have a minimum count of five. Collapsing the final two response categories succeeds in meeting this requirement.

The first step in building a logistic regression model is to select statistically significant predictors and interactions. A forward selection of significant predictors is performed with alpha-to-enter equal to 0.05. Table 9 gives a summary of the predictors and interactions significant in explaining the variation in the level of major-league success attained, or those added to the final model. Table 10 then gives the Type III analysis of

the effects in the final model. The player's position is the only main effect not found to be a significant predictor.

Table 9: Predictors and Interactions Added to the Final Logistic Regression Model

Summary of Forward Selection					
Step	**Effect Entered**	**DF**	**Number In**	**Score Chi-Square**	**Pr > ChiSq**
1	Level	4	1	282.1446	<.0001
2	Factor4	1	2	102.6879	<.0001
3	Factor1	1	3	35.6837	<.0001
4	OHH	1	4	34.4677	<.0001
5	Factor2	1	5	17.4357	<.0001
6	Factor2*Level	4	6	18.4808	0.0010
7	OHH*Factor4	1	7	8.1363	0.0043
8	Factor4*Level	4	8	12.2907	0.0153
9	Factor3	1	9	3.9805	0.0460
10	Factor1*Factor3	1	10	4.4558	0.0348

Table 10: Type III Analysis of Predictors and Interactions Selected for Final Logistic Regression Model

Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
OHH	1	34.4031	<.0001
Level	4	132.1940	<.0001
Factor1	1	17.9661	<.0001
Factor2	1	10.4271	0.0012
Factor3	1	2.0093	0.1563
Factor4	1	28.7973	<.0001
OHH*Factor4	1	7.0122	0.0081
Factor2*Level	4	11.8115	0.0188
Factor4*Level	4	10.9543	0.0271
Factor1*Factor3	1	4.2845	0.0385

The next step is to check whether or not the assumption of proportional odds is met. The assumption of proportional odds means that the coefficients on the significant predictors and interactions do not change significantly from one cumulative response logit to another. Table 11 gives the results to the score test of proportional odds. This test is significant at the 0.05 level, and thus, a non-proportional or partial proportional odds model must be fit.

Table 11: Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
76.0793	38	0.0002

In fitting a non-proportional odds model, three binary logits are created for each observation, and two new variables called MLB Success B and Logtype are created as shown below in the SAS Data step. They compare response level 3 versus levels 2, 1, and 0; levels 3 and 2 versus 1 and 0; and levels 3, 2, and 1 versus level 0.

```
data scores1b; set scores1;
      do; if mlb_success=3 then mlb_successb=1;
      else mlb_successb=0; logtype=3; output; end;
      do; if mlb_success=3 or mlb_success=2 then mlb_successb=1;
      else mlb_successb=0; logtype=2; output; end;
      do; if mlb_success=3 or mlb_success=2 or mlb_success=1 then
      mlb_successb=1;
      else mlb_successb=0; logtype=1; output; end;
run;
```

Cross-products between logtype and the other predictors in the model are added to allow for different regression coefficients across cumulative response logits. A repeated statement is also added where a subject is a player, and an independent working correlation matrix is specified. Thus, the GENMOD procedure statements follow.

```
proc genmod data= scores1b descending;
      class last first logtype level;
      model mlb_successb= logtype ohh level factor1 factor2
      factor3 factor4
      ohh*factor4 factor2*level factor4*level factor1*factor3
      logtype*ohh logtype*level logtype*factor1 logtype*factor2
      logtype*factor3 logtype*factor4 logtype*ohh*factor4
      logtype*factor2*level logtype*factor4*level
      logtype*factor1*factor3
            / dist= bin link= logit type3;
      repeated subject= last*first / type= ind;
run;
```

If the cross-product terms that include logtype are found to be statistically insignificant, then the proportional odds assumption holds at least approximately for that particular predictor and a partial proportional odds model is formed. A backward selection method is used to determine the predictors for which the proportional odds assumption did not hold. Table 12 summarizes the factors for which the proportional odds assumption holds at the .05 level. The assumption holds for some variables, and hence, a partial proportional odds model is shown to be adequate. The parameter estimates from the final partial proportional odds model are given in Appendix D.

Table 12: Summary of Factors for which the Proportional Odds Assumption Holds at the .05 Level

Effect	Assumption Holds?
OHH	Yes
Level	Yes
Factor1	No
Factor2	No
Factor3	No
Factor4	Yes
OHH*Factor4	Yes
Factor2*Level	Yes
Factor4*Level	No

Effect	Assumption Holds?
Factor1*Factor3	Yes

Before we move on to the next step, it is important to understand the interactions between any two factors. Figures 4-12 illustrate the significant interactions through two-dimensional plots of the cumulative logits (response level 3 versus level 2, 1, and 0). In the case of the Factor 4-by-Level interaction, three plots are displayed, one for each of the cumulative logits, in order to display how this interactions change across cumulative logits. All predictors not involved in the interaction are set at their mean value except for highest level played in the minors set at double-A. (The corresponding plots of the probability curves rather than the cumulative logits are given in Appendix C.)

The effect of factor 4 (pure-hitting ability) on the chances of performing well in the majors (either success level 3 or 4) becomes larger as players get older with respect to others in their league, as illustrated in Figure 4.

Figure 4

The effects of factor 1 (ability to slug) and factor 2 (lead-off hitting skills) on the chances of getting at least a taste of the majors are larger than the effects on achieving higher levels of success in the majors, as illustrated in Figures 5 and 6. This pattern is reversed for the effect of factor 3 ("patience" at the plate), as illustrated in Figure 7.

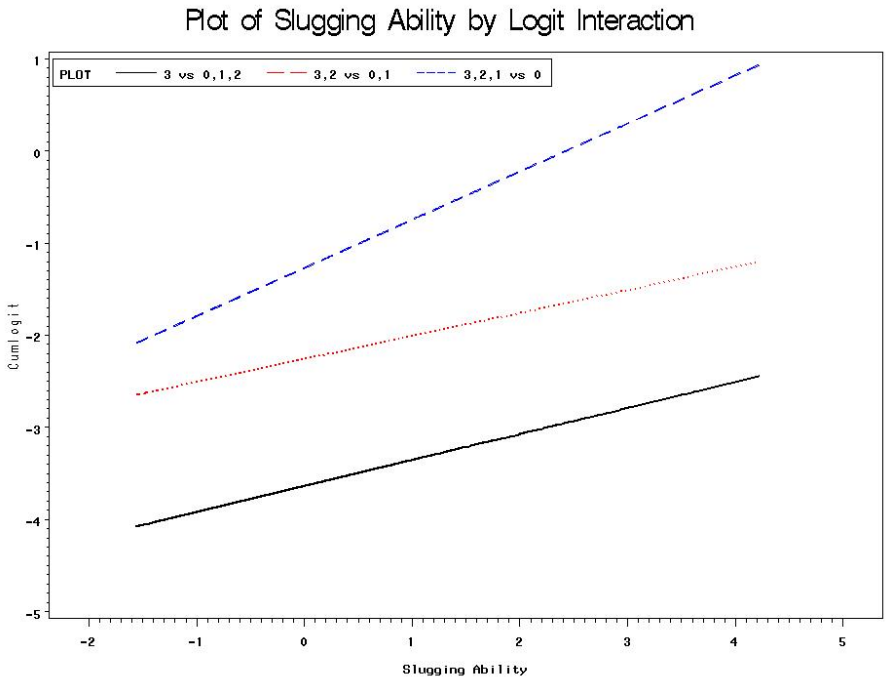

Figure 5

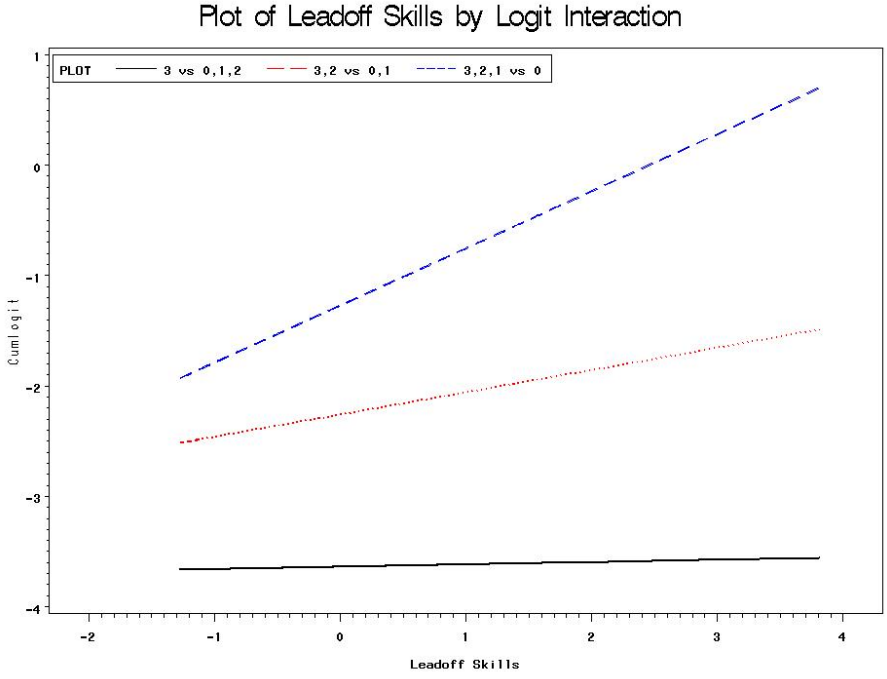

Figure 6

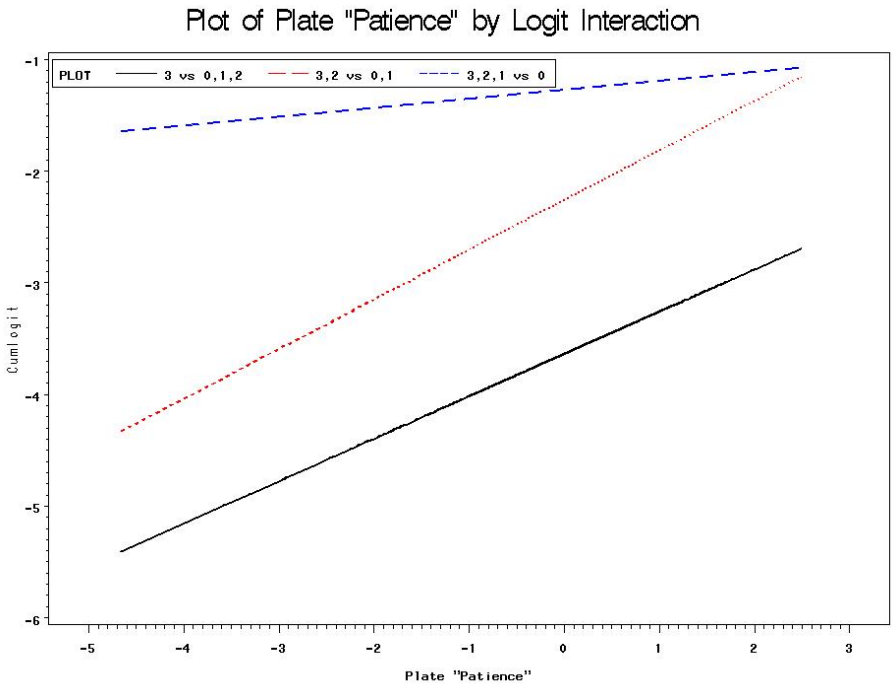

Figure 7

The effect of factor 2 (leadoff hitting skills) on the chances of performing well in the majors (either success level 3 or 4) is very mixed for players in varying levels of the minor leagues, as illustrated in Figure 8. The effect of factor 4 (pure-hitting ability) on the chances of getting at least a taste of the majors is positive for players in all levels of the minor leagues, as illustrated in Figure 11. However, its effect on the players who made it to the majors in 2002 reaching higher levels of success becomes more negative, as illustrated in Figures 9 and 10.

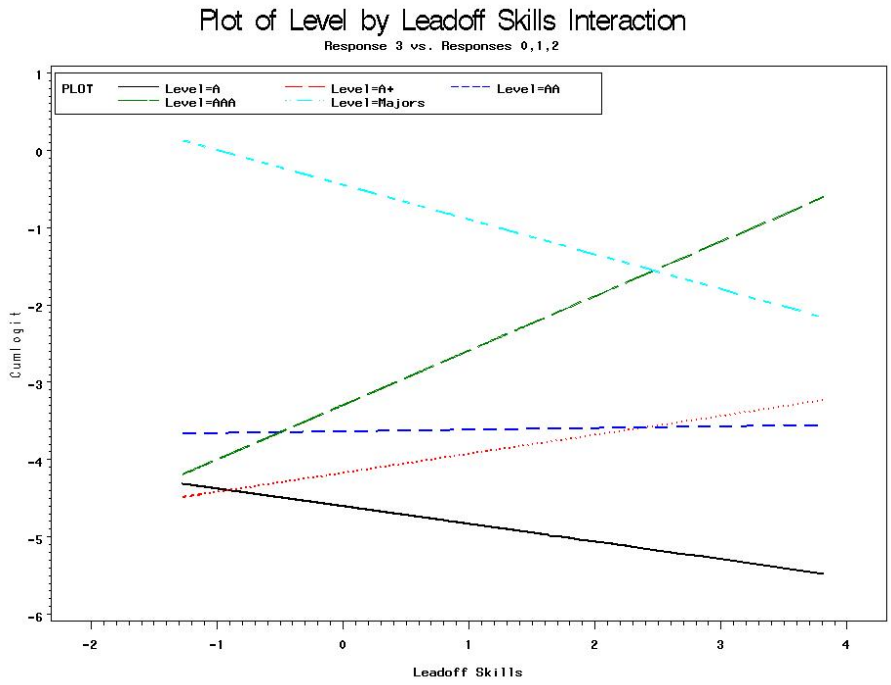

Figure 8

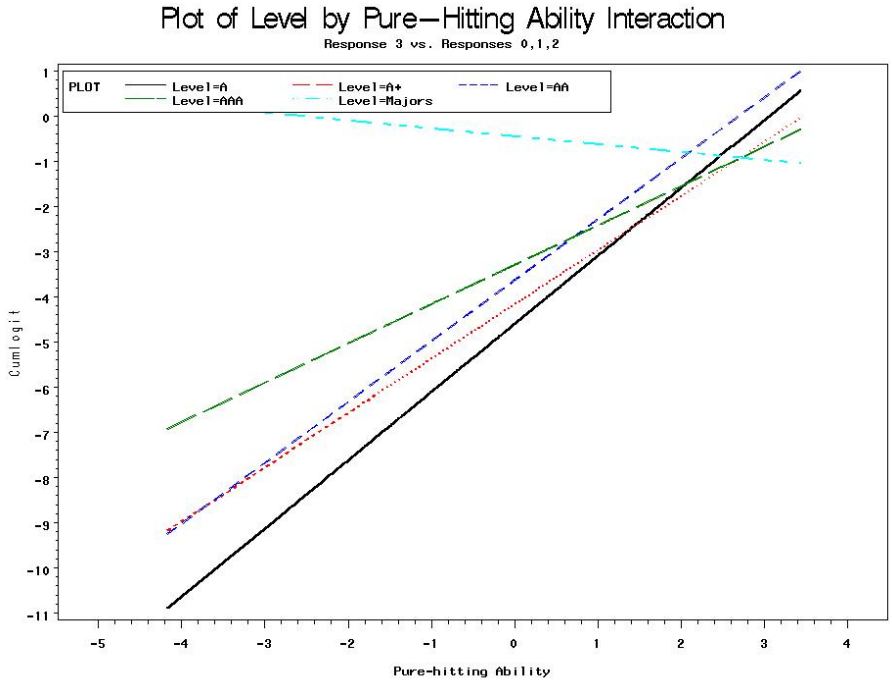

Figure 9

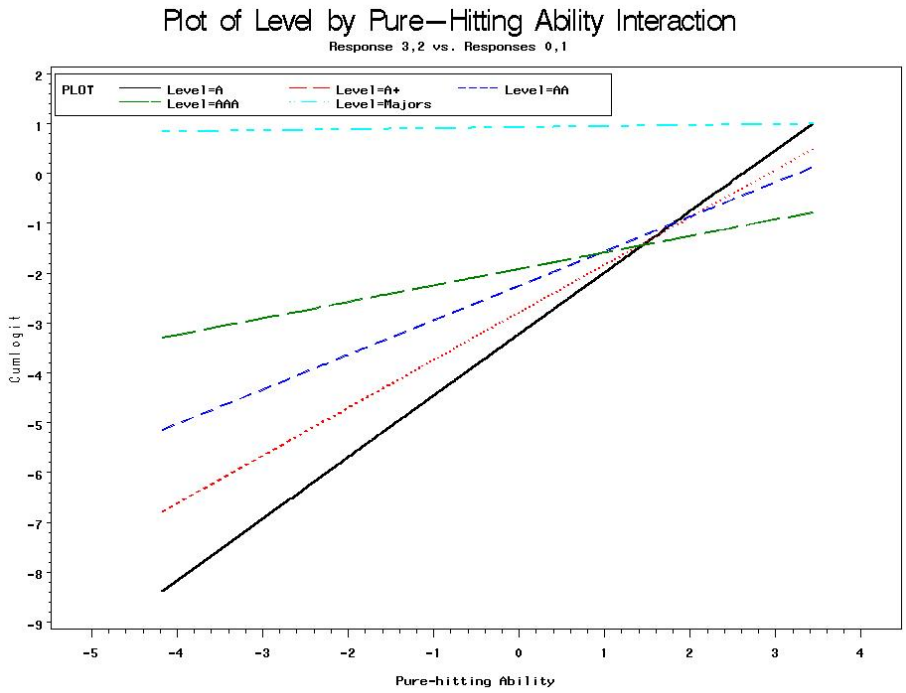

Figure 10

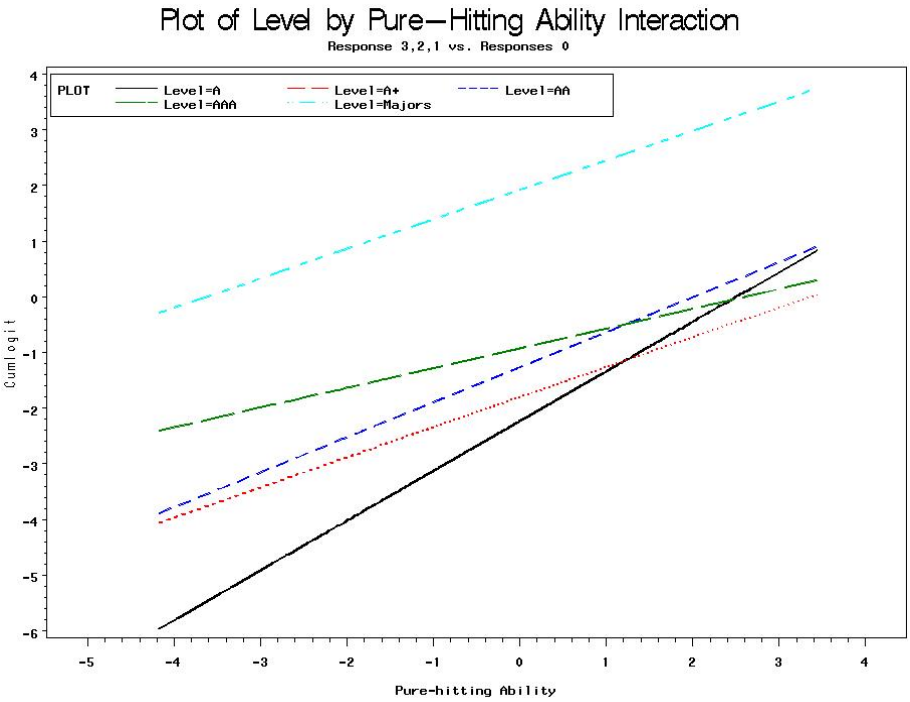

Figure 11

The effect of factor 4 (pure-hitting ability) increases as factor 1 (ability to slug) increases, as illustrated in Figures 12.

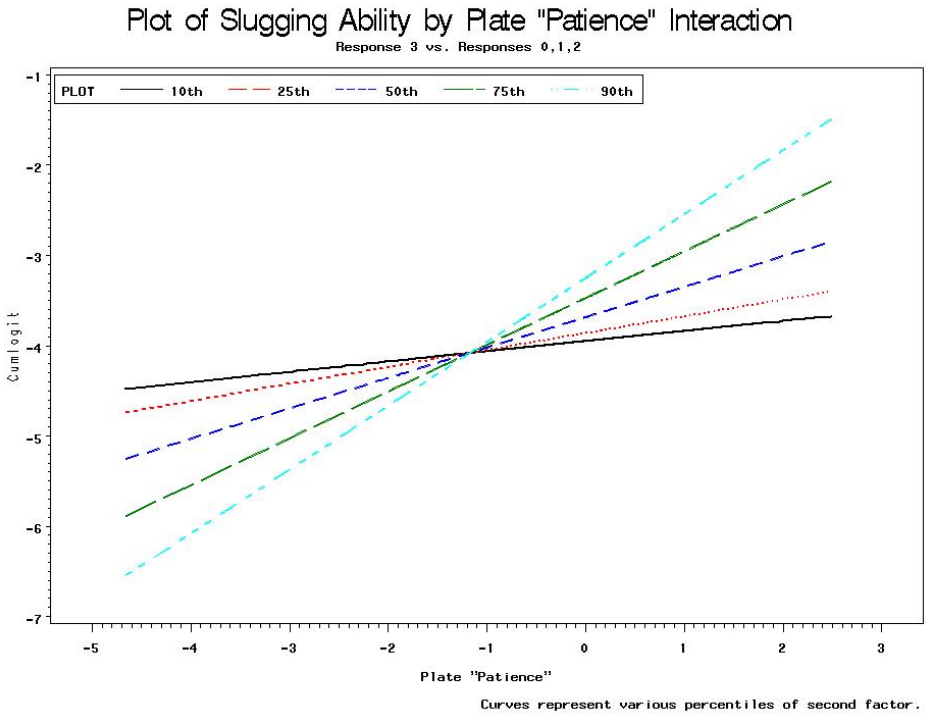

Figure 12

The final step in making sure the model provides a proper fit is to check the deviance about the model. A deviance statistic that is too large suggests over-dispersion problems or other lack-of-fit problems. Table 13 gives the deviance statistic and the associated degrees of freedom. The expected value of the statistic is the number of degrees of freedom, 2919. The standard error is the square root of two times the degrees of freedom, or 76.41. The observed deviance is 1705.86, well below the expected value.

Table 13: Deviance Statistic for Assessing Goodness of Fit

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	2919	1754.7090	0.6011

IV.5 Final Prospect Rankings

With the final logistic regression model found to be an appropriate model, the probability of reaching at least major-league success level three is computed. Then the players are then ranked from the highest probability of success to the lowest. Table 14 displays the Top 50 minor league hitters of 2002 based on these rankings along with their respective ranking (or rating if not ranked) by Sickels (2003) and their observed level of Major League success. (Sickels rates prospects on a scale ranging from "A" to "C-." His Top 50 players typically include all those rated A, A-, B+, and many rated B.) Only three of the Top 50 have not yet played in the majors, Jason Stokes being the most highly ranked. Stokes has had up-and-down seasons subsequent to 2002 mainly due to poor plate discipline that was not observed in 2002 and some injuries. Many players ranked in the Top 50 by our model were not in Sickels' book. (Most of these players had 50+ games in the majors in 2002. We applied the 50+ game rule only to seasons prior to 2002.) At the bottom of Table 14 are the remaining players who have achieved the highest level of success in the majors thus far, level 4, and their rankings.

Table 14: Top 50 Minor League Hitters in 2002 According to Logistic Regression Modeling of MLB Success along with John Sickel's Rankings (or Ratings if not Ranked) (Sickels, 2003)

Sickel's Rank	My Rank	Last	First	P(Success)	MLB Success
B	1	Cust	Jack	0.93108	2
7	2	Choi	Hee Seop	0.90506	3
10	3	Hafner	Travis	0.83581	4
C+	4	Pena	Wily Mo	0.82044	3
8	5	Hairston	Scott	0.79678	3
B-	6	Munson	Eric	0.77927	3
B	7	Hawpe	Brad	0.77014	3
3	8	Martinez	Victor	0.75210	3
C	9	Hart	Jason	0.75086	1
C	10	Langerhans	Ryan	0.74085	3
11	11	Stokes	Jason	0.73841	0
C	12	Duncan	Jeff	0.73586	2
NA	13	Phelps	Josh	0.72519	3
48	14	Borchard	Joe	0.71974	2
B+	15	Werth	Jayson	0.70878	3
B-	16	Shelton	Chris	0.70550	3
1	17	Teixeira	Mark	0.70318	4
C	18	Henson	Drew	0.67427	1
C	19	Quinlan	Robb	0.66658	3
42	20	Restovich	Michael	0.65228	3
NA	21	Crede	Joe	0.64730	3
C+	22	Chen	Chin-Feng	0.64564	1
B	23	Cash	Kevin	0.64526	2
NA	24	Brito	Juan	0.63283	2
NA	25	Pena	Carlos	0.62278	3
B-	26	Ludwick	Ryan	0.62235	3
C	27	Davis	J.J.	0.62209	2
32	28	Harris	Brendan	0.61418	1
NA	29	Burroughs	Sean	0.60818	2

Sickel's Rank	My Rank	Last	First	P(Success)	MLB Success
C+	30	Berroa	Angel	0.60685	3
NA	31	Rivera	Mike	0.60125	2
C+	32	Broussard	Ben	0.58743	3
C+	33	Infante	Omar	0.57282	2
NA	34	Blalock	Hank	0.56537	3
45	35	Tracy	Chad	0.55898	3
C	36	Hall	Bill	0.55853	3
B-	37	Gomes	Jonny	0.55552	3
4	38	Phillips	Brandon	0.55544	2
5	39	Cuddyer	Michael	0.55230	3
25	40	Baldelli	Rocco	0.55211	3
NA	41	Ross	Dave	0.54508	2
B	42	Stanley	Henri	0.52699	0
NA	43	Lunsford	Trey	0.51438	1
NA	44	Lopez	Felipe	0.50795	3
36	45	Linden	Todd	0.50436	2
B	46	Dubois	Jason	0.50108	3
18	47	Overbay	Lyle	0.49728	3
NA	48	Sandberg	Jared	0.48413	3
C	49	Rich	Dominic	0.48373	0
C	50	Torcato	Tony	0.46599	1

B-	141	Bay	Jason	0.19638	4
9	239	Cabrera	Miguel	0.09187	4
26	305	Wright	David	0.05819	4
C+	357	Howard	Ryan	0.04126	4

V. Conclusion

Favorable results have come about as a result of using statistical models to evaluate young minor-league position players. The average response from a number of different hitting categories was noted to be significantly different between minor-leaguers making to the majors and those not making it. In addition, often the average response continued to increase for players showing increased levels of success in the majors. Next, four underlying factors common to all hitting statistics were identified, and these underlying factors were shown to be statistically significant predictors of major-league success. Also found to be significant predictors were the highest level in the minors played in 2002 and a variable measuring a player's age relative to his level. Finally, a partial proportional odds logistic regression model showing adequate fit was built in such a way as to predict major-league success and rank players accordingly.

However, there are limitations to my conclusions due to the data collected. First, the model shows only the degree of success that can be achieved in the succeeding three years. Second, it would have been ideal to consider multiple minor-league seasons. Multiple seasons of data are important because players exhibit variability in performance from year to year, (Jason Stokes is a prime example), and because statistics measuring performance can vary even if the player's true performance level doesn't. James, Albert, and Stern discuss this issue (1993). Players are often assessed by the batting average, which is just the proportion of at-bats with a hit. A player's true batting average can then, for example, be thought of as a probability, being estimated by their proportion of successes each season. There is great variation in estimates of a proportion. Consider a

true .270-hitter. There is approximately a 9-10% chance that he can hit .300 or higher over the course of a minor league season just by luck, assuming 400 at-bats. For these reasons, it would be good to obtain data for at least 2 minor-league seasons per player. Third, it would have been optimal to have obtained data from an earlier point in time. Players from an earlier time-frame should have had ample opportunity to make their way up the levels of the minor leagues and into the majors, if they were ever going to do so. In addition, this would have allowed for more time for the players to truly distinguish themselves in the majors. It would be easier to discern which players really play at an "all-star" level from those who are simply good players who start and those who are marginal major-leaguers who can play when the team needs them.

Fourth, in addition to the variables observed, it would have also been ideal to have information about the players' home ballparks and leagues. There is plenty of evidence discussed by various sabermetricians such as John Sickels and Bill James indicating that some ballparks are hitters' parks and some are pitchers' parks. To slug .500 in one ballpark may be average at best, and it may be eye-popping in another ballpark.

Lastly, even if the ideal data were collected, the reality is that there will still be error in predicted outcomes. There are still variables explaining a player's level of success in the majors, if any, that we have not, and perhaps cannot, measure. Such variables are injury, personal relationships and conflicts, and the loss of confidence. But, the risk of drafting, trading for, or moving up players who don't pan out can be minimized by identifying and using measurable explanatory variables.

VI. Future Research

There are many more routes one could take in order to evaluate and rank minor league hitters. A few in particular pique my interest for future research.

If I could obtain multiple seasons of minor-league data for each player, it would be interesting to do a time-series analysis. The following questions could possibly be answered:

- Do certain offensive statistics trend upward, downward, or vary randomly over the course of a minor-league career?
- Do the trends or patterns of variation of certain offensive statistics differ significantly between players who eventually play in the majors and those who don't?
- Can a smoothing method be used to forecast major-league success?

I would also be interested in trying to model a quantitative response such as runs-created in the majors or win-shares (James & Henzler, 2002) in the majors. This would help take out the subjectivity in choosing categorical levels of major-league success.

Lastly, I would also like to use the model built in this research to rank players from other minor-league seasons to check model adequacy.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. 605 Third Ave., New York, NY: John Wiley & Sons, Inc.

Blake, M. (1991). A brief history of minor league baseball. *The minor leagues* (pp. 23-79). New York, NY: WYNWOOD Press.

Hosmer, D. W., & Lemeshow, S. (2000). In Cressie, Noel A. C., et al (Eds.), *Applied logistic regression* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

James, B. (1985). *Bill james historical baseball abstract*. New York, NY: Villard Books.

James, B., Albert, J., & Stern, H. S. (1993). Answering questions about baseball using statistics. *Chance, 6*(2), 17-22.

James, B., & Henzler, J. (2002). *Win shares*. Morton Grove, IL: STATS Publishing Inc.

Johnson, D. E. (1998). In Crockett C., Velthaus N., Palagi L. and Clark M. (Eds.), *Applied multivariate methods for data analysts*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Krohn, G. A. (1983). Measuring the experience-productivity relationship: The case of major league baseball. *Journal of Business & Economic Statistics, 1*(4), 273-279.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game* (First ed.). New York, NY: W. W. Norton & Company, Inc.

Sickels, J. (2003). *The baseball prospect book 2003* (First ed.). Marceline, MO: Walsworth Publishing Company.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). Using GEE to fit a partial proportional odds model: Univariate outcome. *Categorical data analysis using the SAS system* (2nd ed.) (pp. 533-540). Cary, NC: SAS Institute Inc.

Glossary

At-bat (AB) - A batter has an at bat every time he faces a pitcher except under the following circumstances: he receives a walk, he is hit by a pitch, he hits a sacrifice fly or a sacrifice bunt, he is awarded first base due to interference or obstruction (usually by the catcher), the inning ends while he is still at bat, or he is replaced by another hitter while he is still at bat.

Caught Stealing (CS) – When a player is tagged out by a baseman while attempting to steal a base

Double (D) – A hit by which a player reaches second base

Homerun (HR) – A hit by which a player rounds all three bases and scores

Isolated Power (IsoPower) – The number of extra-base hits (those greater than a single) per at-bat; it is computed by the following formula:

$$\frac{\text{Total bases - Hits}}{\text{Number of At - bats}} = \frac{S + 2D + 3T + 4HR - H}{AB}$$

On-base percentage (OBP) – The percentage of time a player reaches any base; it is computed by the following formula:

$$\frac{\text{Hits plus Walks}}{\text{Number of At - bats plus Walks}} = \frac{H + BB}{AB + BB}$$

On-base percentage plus slugging (OPS) – A statistic used to measure both a player's ability to get on base and to drive players around the bases with extra-base hits; it is computed by the following formula:

$$\text{OBP} + \text{Slugging}$$

Runs created (RC) – A measure created by Bill James to measure a player's total offensive production; it is generically the number of runs he creates for his team and is computed by the following formula:

$$\frac{(H + BB - CS) + (S + 2D + 3T + 4HR + .55SB)}{AB + BB}$$

Single (S) – A hit by which a player reaches first base

Slugging percentage (Slugging) - The total number of bases per at-bat; it is computed by the following formula:

$$\frac{\text{Total bases}}{\text{Number of At - bats}} = \frac{S + 2D + 3T + 4HR}{AB}$$

Stolen base (SB) - When a baserunner successfully advances to the next base while the pitcher is delivering the ball to home plate

Strikeout – When a hitter receives three strikes during his time at bat

Triple (T) – A hit by which a player reaches third base

Walk (BB) – An advance to first base that is awarded to a batter who takes four pitches that are balls, also known as a base-on-balls

Appendix A

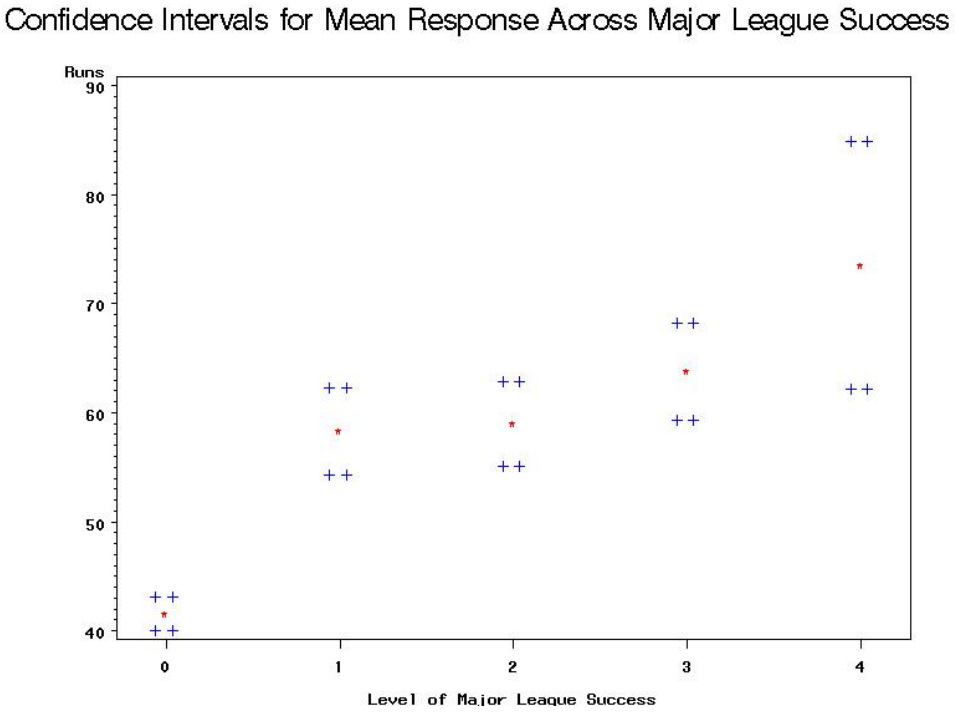

Figure 13

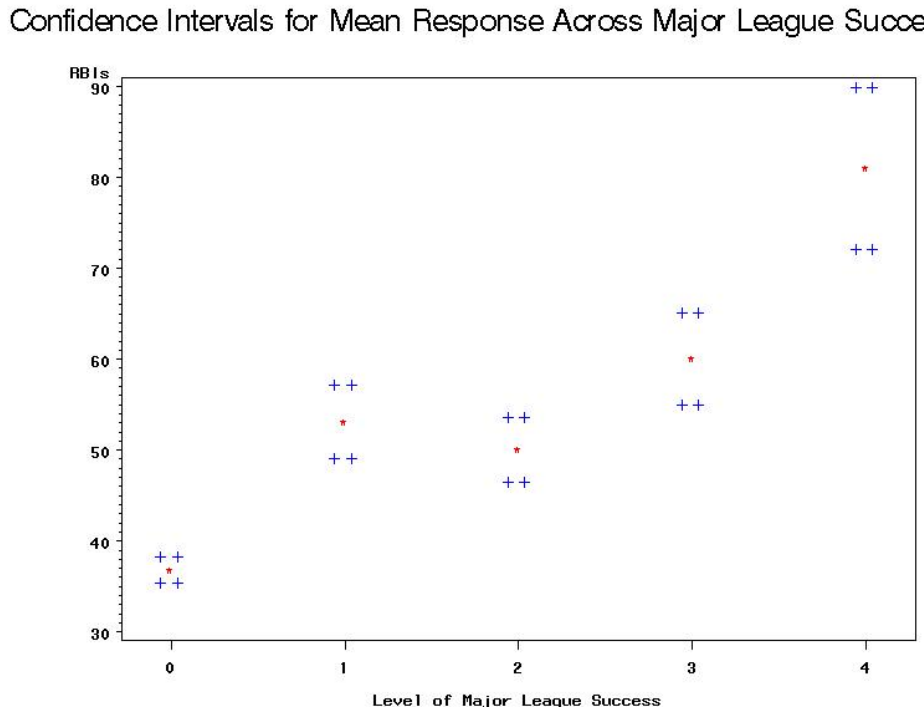

Figure 14

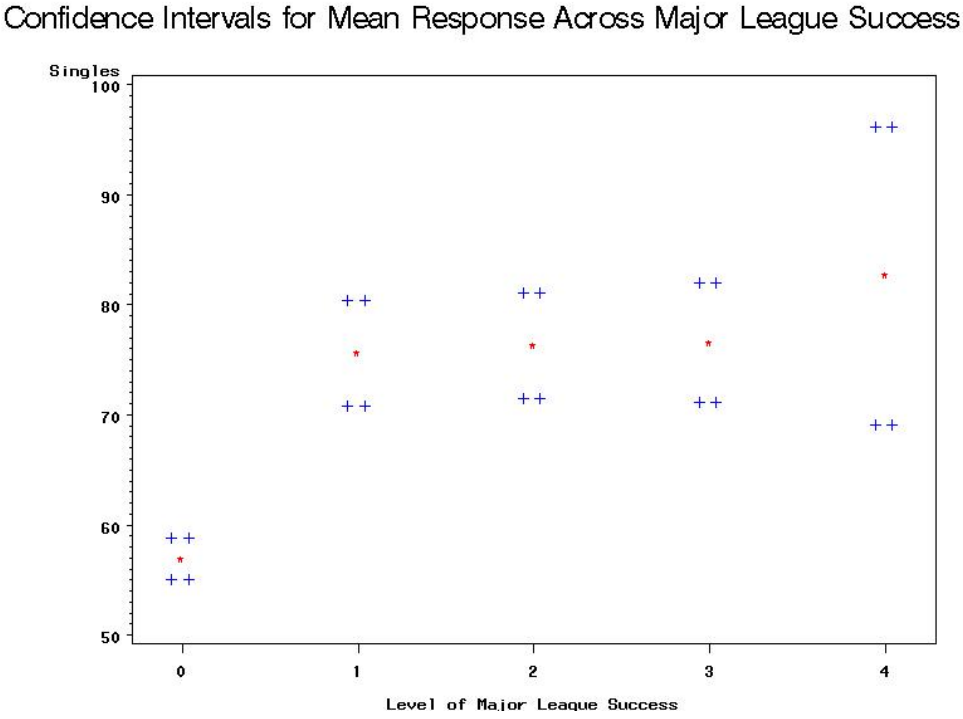

Figure 15

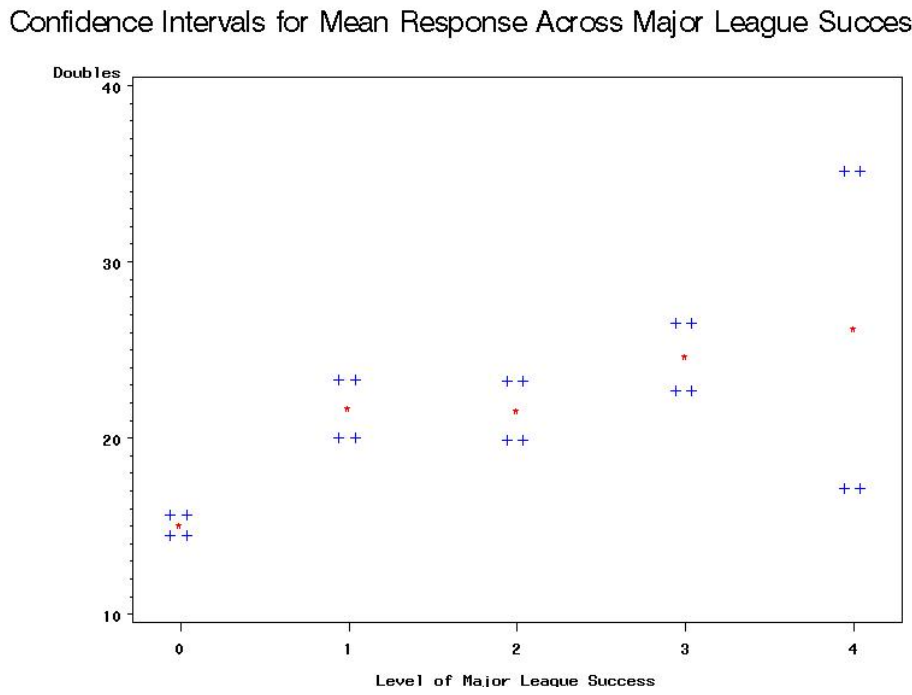

Figure 16

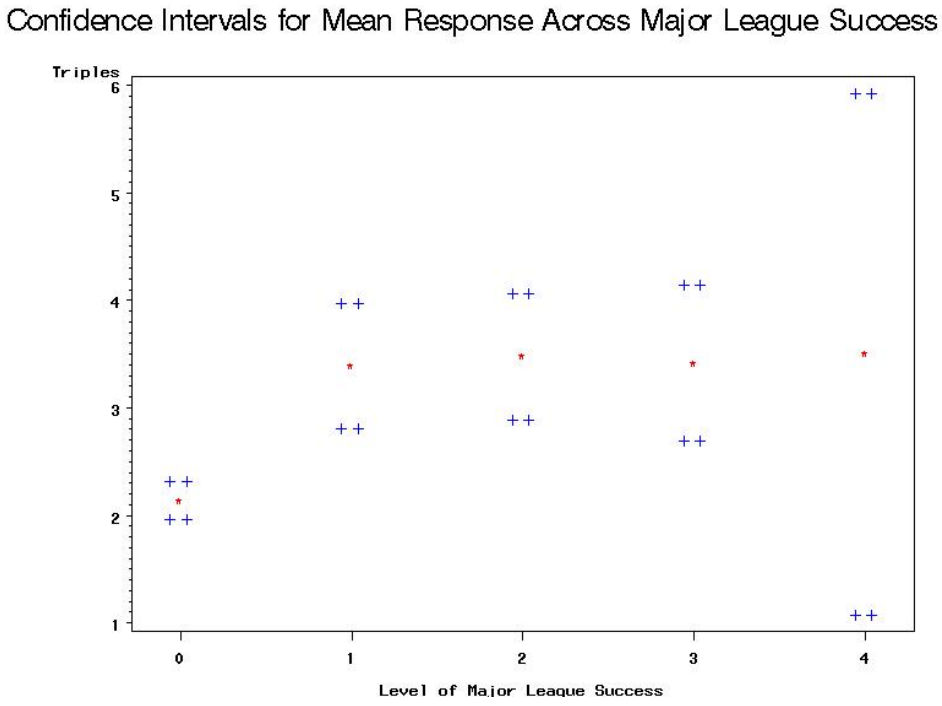

Figure 17

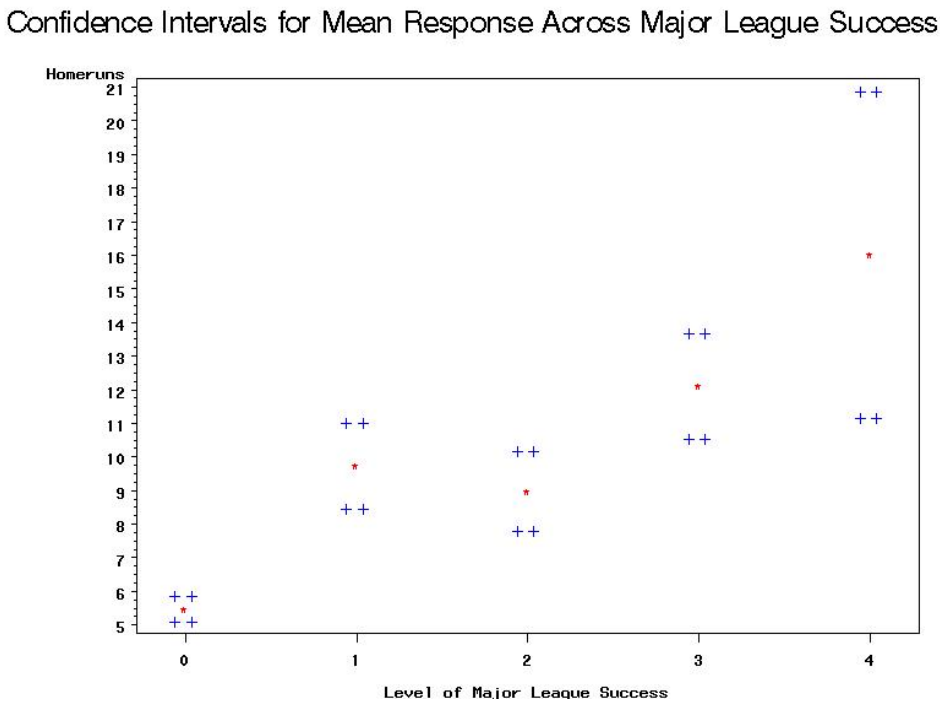

Figure 18

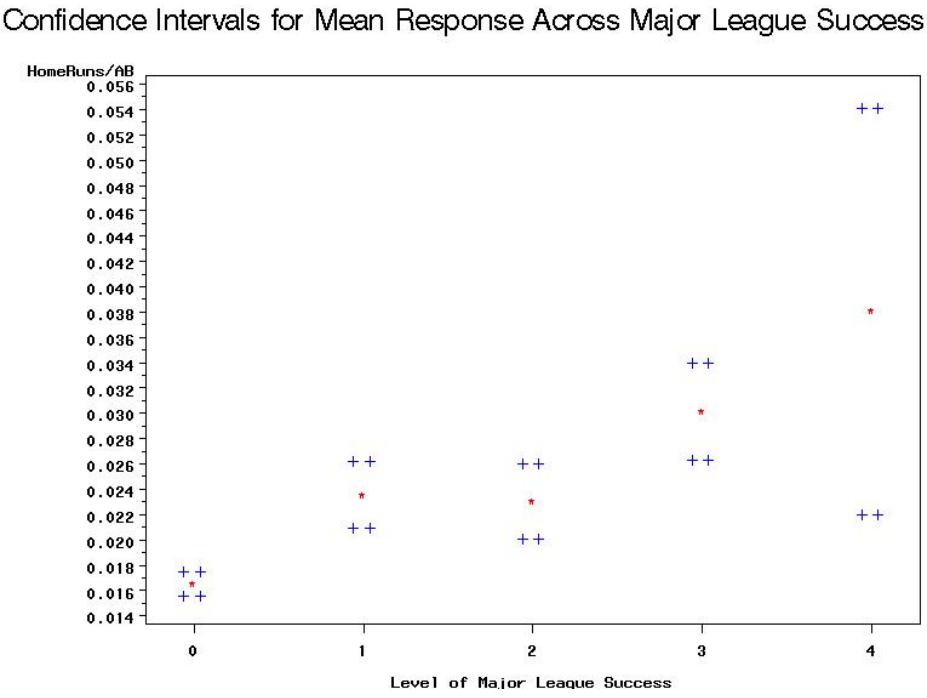

Figure 19

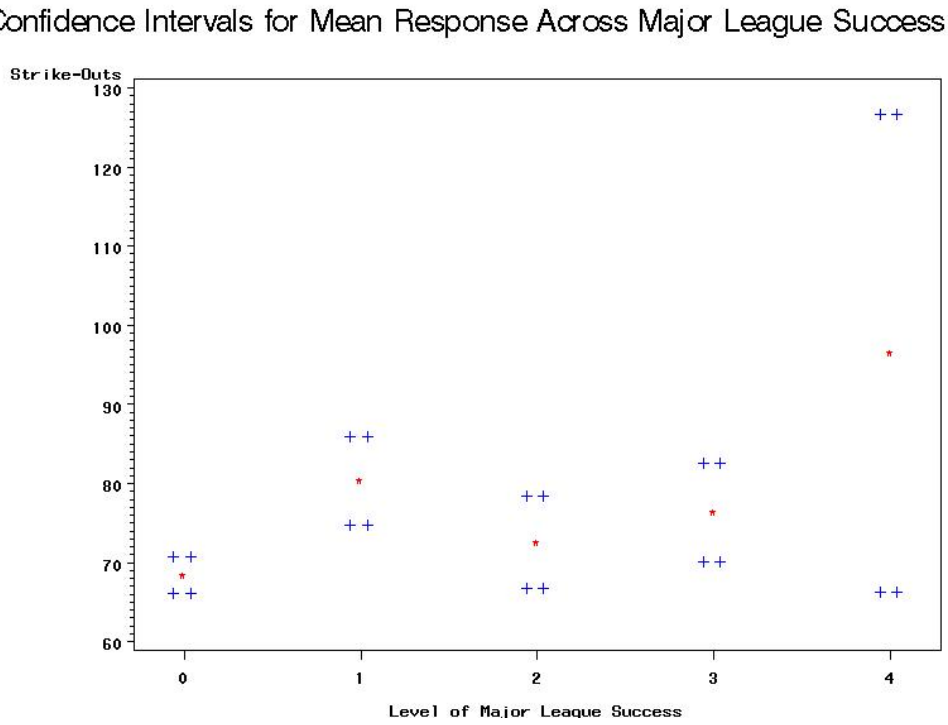

Figure 20

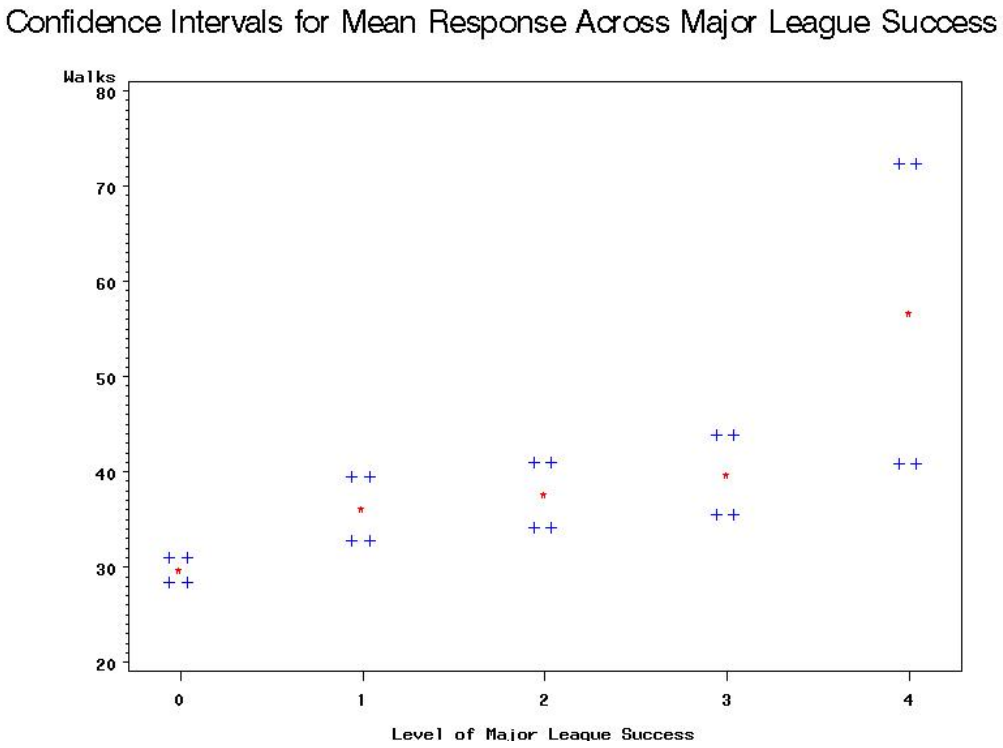

Figure 21

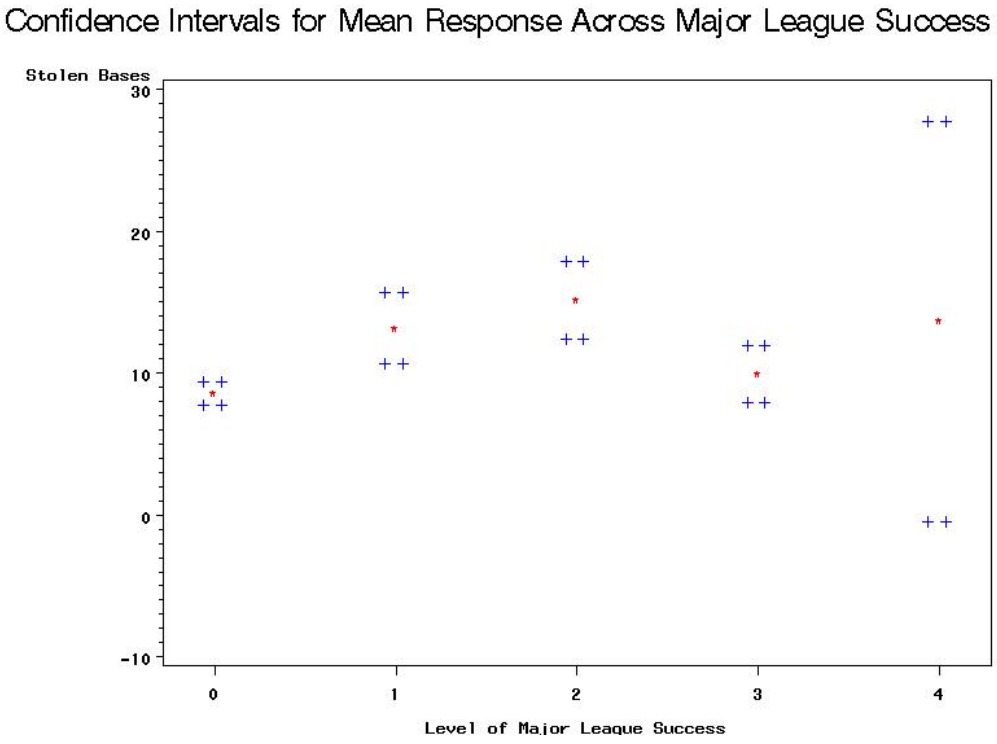

Figure 22

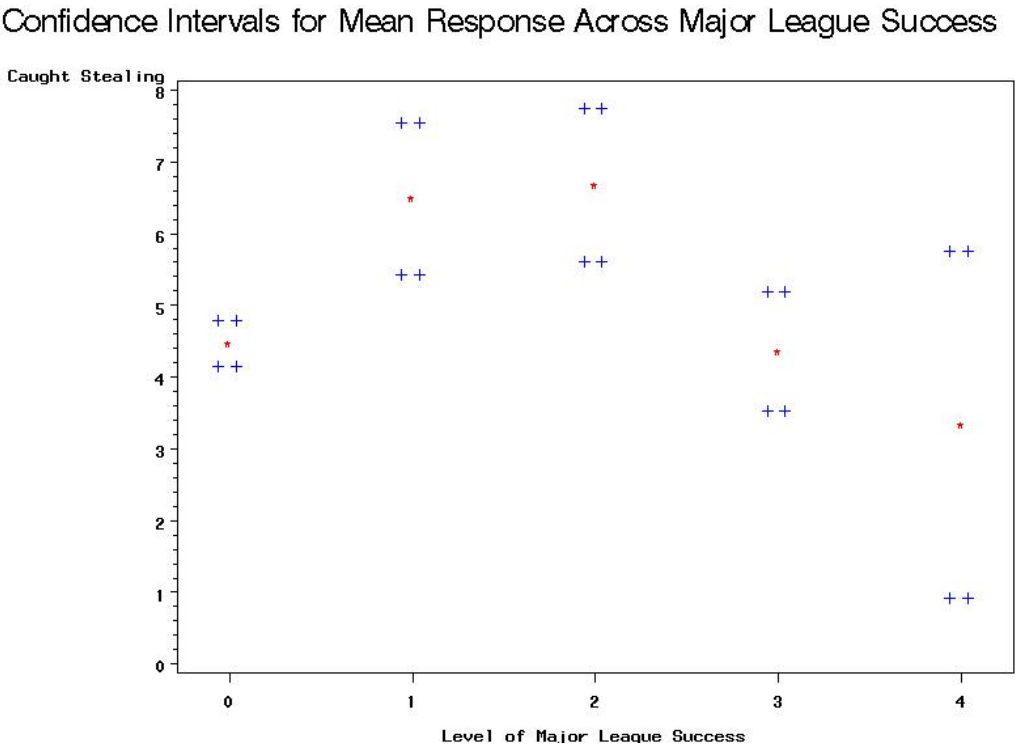

Figure 23

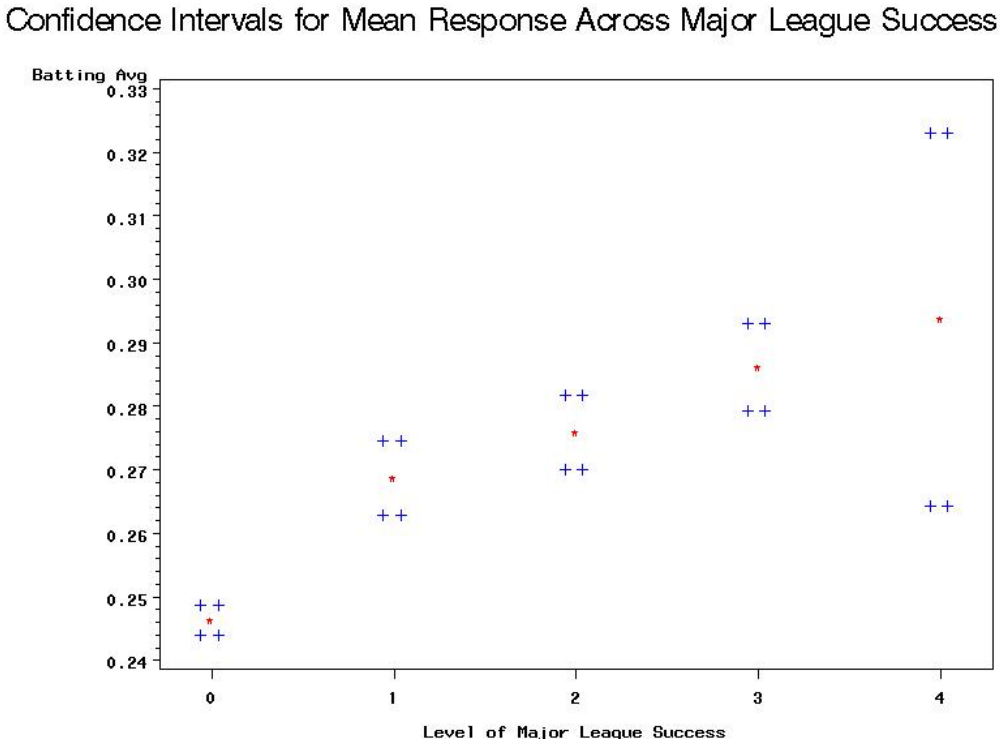

Figure 24

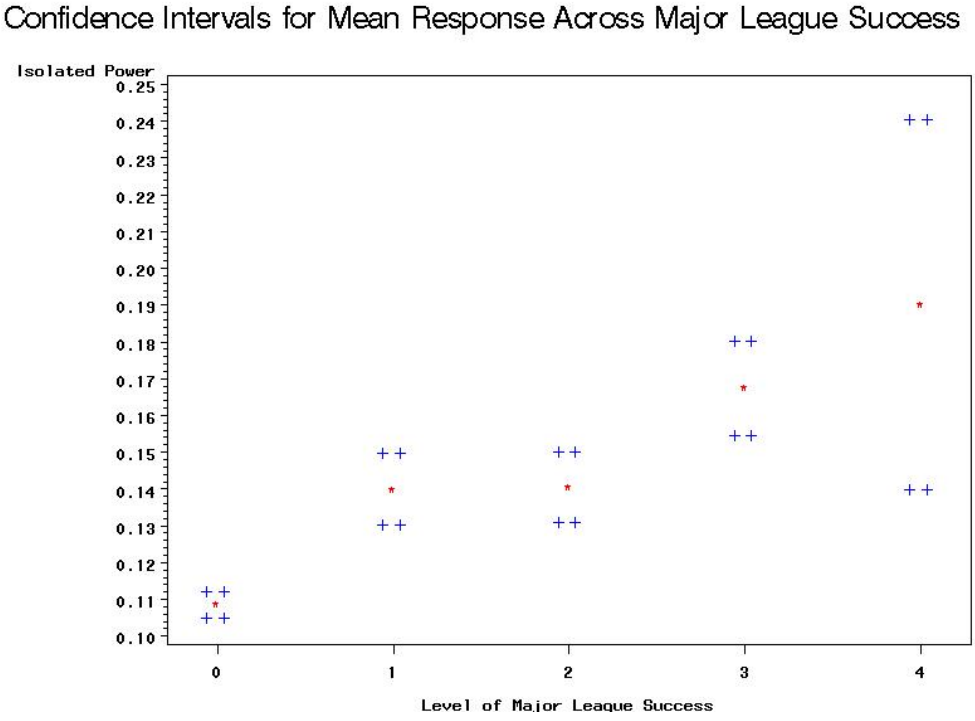

Figure 25

Appendix B

Obs	Last	First	League	Age	Pos	Level	Runs	RBIs	S	D	T	HR
1	Abercrombie	Reggie	SOU	22	OF	AA	81	56	100	23	13	10
2	Abram	Matt	MWL	22	OF	A	23	32	39	9	0	6
3	Abreu	Dennis	SOU	24	2B	AA	45	51	87	17	4	7
4	Abruzzo	Jared	CAL	20	C	A+	53	53	51	27	0	16
5	Acevas	Jon	SOU	24	C	AAA	26	16	26	8	1	3
6	Acevedo	Carlos	FSL	21	OF	A+	8	6	19	6	0	0
7	Acevedo	Inocencio	FSL	23	3B	A+	28	13	36	5	1	2
8	Ackerman	Scott	FSL	23	C	AAA	29	44	53	12	2	9
9	Acuna	Ron	FSL	21	OF	AA	70	57	108	20	7	2
10	Adams	Russ	FSL	21	SS	A+	23	12	27	4	2	1
11	Aguila	Chris	EAS	23	OF	AA	62	46	88	28	4	6
12	Ahumada	Alex	EAS	23	SS	AA	42	28	53	5	3	2
13	Alcala	Juan	CAL	24	C	AA	9	7	19	1	0	1
14	Alexander	Kevin	SAL	21	2B	A	59	28	75	24	0	4
15	Alfaro	Jason	TEX	24	3B	AA	71	74	89	36	2	16
16	Alfonzo	Eliezer	SOU	23	C	AA	30	47	51	17	1	9
17	Allegra	Matt	CAL	21	OF	A+	74	93	81	35	3	20
18	Allen	Luke	PCL	23	OF	Majors	85	78	122	28	3	12
19	Alleva	Joe	MWL	22	C	A	15	18	36	6	0	0
20	Almonte	Erick	INT	24	SS	AAA	53	61	62	17	1	17
21	Alvarez	Jimmy	SOU	22	2B	AA	83	69	95	32	3	8
22	Alvarez	Tony	EAS	24	OF	Majors	79	59	108	37	1	15
23	Amador	Chris	SAL	19	2B	A	60	26	72	5	4	3
24	Ambres	Chip	FSL	22	OF	A+	88	37	79	25	7	9
25	Ambrosini	Dominick	MWL	21	OF	A	53	50	85	29	4	5
26	Amezaga	Alfredo	PCL	24	SS	Majors	77	51	92	25	7	6
27	Anderson	Keith	SAL	23	C	AA	20	20	29	12	0	3
28	Anderson	Travis	CAR	22	C	A+	19	12	25	4	0	3
29	Ansman	Craig	CAL	24	C	AA	73	64	74	24	7	21

Obs	Last	First	League	Age	Pos	Level	Runs	RBIs	S	D	T	HR
30	Aquino	Jackson	MWL	20	SS	A	33	25	52	12	1	1
31	Aracena	Sandy	FSL	21	C	A+	19	15	40	3	0	1
32	Arauja	Ramon	CAR	21	2B	A+	22	11	31	6	0	0
33	Arnerich	Tony	CAR	22	C	A+	33	46	71	12	0	2
34	Arroyo	Will	MWL	20	2B	A	30	22	46	5	2	0
35	Arteaga	Joshua	MWL	22	3B	A+	24	16	38	10	2	1
36	Asadoorian	Rick	MWL	22	OF	A	70	55	87	12	11	8
37	Asche	Kirk	TEX	24	OF	AA	50	56	56	13	10	12
38	Aspito	Jason	CAR	23	OF	A+	40	36	58	14	2	6
39	Asprilla	Avelino	SAL	21	SS	A+	31	21	38	4	1	3
40	Athas	Jamie	CAL	22	SS	A+	65	40	94	15	7	1
41	Atkins	Garrett	SOU	22	3B	AA	71	61	96	27	3	12
42	Avila	Rob	FSL	23	1B	A+	26	26	36	11	0	5
43	Ayala	Eliott	MWL	23	3B	A+	26	18	46	6	2	0
44	Ayala	Odannys	MWL	22	OF	A	68	61	84	22	7	6
45	Aybar	Willy	FSL	19	3B	A+	56	65	49	18	2	11
46	Badeaux	Brooks	INT	25	2B	AAA	45	29	67	14	2	2
47	Bailey	Jeff	EAS	23	1B	AA	45	52	56	17	1	13
48	Bailie	Stefan	FSL	22	1B	A+	16	26	29	8	2	2
49	Baldelli	Rocco	INT	20	OF	AAA	86	71	108	28	3	19
50	Ball	Jarred	MWL	19	OF	A	48	23	58	13	4	2

Obs	SO	SB	CS	BB	OBP	Avg	IsoPower	SO/BB
1	159	42	17	27	0.31059	0.27547	0.14906	5.8889
2	51	1	3	17	0.27413	0.22314	0.11157	3.0000
3	102	18	14	21	0.32151	0.28607	0.11443	4.8571
4	124	1	1	30	0.29880	0.24416	0.19481	4.1333
5	36	0	1	20	0.31351	0.23030	0.11515	1.8000
6	19	0	2	5	0.25000	0.21739	0.05217	3.8000
7	36	11	3	9	0.27320	0.23784	0.07027	4.0000
8	49	1	0	13	0.32014	0.28679	0.16226	3.7692

Obs	SO	SB	CS	BB	OBP	Avg	IsoPower	SO/BB
9	79	37	13	39	0.34783	0.29336	0.08565	2.0256
10	17	5	2	18	0.31515	0.23129	0.07483	0.9444
11	101	14	8	48	0.36478	0.29371	0.12587	2.1042
12	60	13	5	24	0.30851	0.24419	0.06589	2.5000
13	24	2	0	2	0.21495	0.20000	0.03810	12.0000
14	65	12	5	44	0.36567	0.28771	0.10056	1.4773
15	75	11	9	44	0.37475	0.31429	0.19341	1.7045
16	69	2	3	12	0.30100	0.27178	0.16028	5.7500
17	160	9	9	46	0.34259	0.28138	0.20445	3.4783
18	77	4	6	53	0.39350	0.32934	0.13972	1.4528
19	24	0	0	20	0.29952	0.22460	0.03209	1.2000
20	119	12	3	43	0.31042	0.23775	0.17157	2.7674
21	121	20	11	77	0.37456	0.27767	0.12475	1.5714
22	71	29	18	26	0.35084	0.31755	0.16568	2.7308
23	142	56	15	38	0.26754	0.20096	0.05263	3.7368
24	98	23	8	57	0.31272	0.23576	0.12967	1.7193
25	113	7	5	30	0.32484	0.27891	0.11791	3.7667
26	100	23	14	45	0.31083	0.25097	0.11004	2.2222
27	37	0	1	14	0.33526	0.27673	0.13208	2.6429
28	26	0	4	14	0.34328	0.26667	0.10833	1.8571
29	119	3	3	39	0.35408	0.29508	0.23653	3.0513
30	70	12	10	28	0.27011	0.20625	0.05313	2.5000
31	50	2	0	20	0.26230	0.19643	0.02679	2.5000
32	34	7	3	6	0.24022	0.21387	0.03468	5.6667
33	53	2	2	20	0.31157	0.26814	0.05678	2.6500
34	38	11	10	47	0.37037	0.23767	0.04036	0.8085
35	34	1	4	15	0.27848	0.22973	0.07658	2.2667
36	96	14	8	43	0.32992	0.26517	0.13034	2.2326
37	132	9	6	38	0.31159	0.24202	0.18351	3.4737
38	83	2	5	26	0.30994	0.25316	0.11392	3.1923
39	49	7	2	13	0.29353	0.24468	0.07979	3.7692

Obs	SO	SB	CS	BB	OBP	Avg	IsoPower	SO/BB
40	124	14	6	44	0.31569	0.25107	0.06867	2.8182
41	77	6	6	57	0.34392	0.27059	0.13529	1.3509
42	41	4	2	24	0.29008	0.21849	0.10924	1.7083
43	36	4	1	17	0.28629	0.23377	0.04329	2.1176
44	70	5	3	61	0.38961	0.29676	0.13466	1.1475
45	54	15	8	65	0.33181	0.21505	0.14785	0.8308
46	45	8	2	28	0.30707	0.25000	0.07059	1.6071
47	78	3	3	59	0.39674	0.28155	0.18770	1.3220
48	33	1	0	16	0.32571	0.25786	0.11321	2.0625
49	97	26	13	21	0.35872	0.33054	0.19038	4.6190
50	85	12	1	41	0.32597	0.23988	0.08411	2.0732

Obs	HR/AB	Factor 1	Factor 2	Factor 3	Factor 4
1	0.018868	0.34920	2.48930	-0.79747	0.59464
2	0.024793	0.02771	-0.82562	-0.71442	-0.92839
3	0.017413	-0.06966	0.82013	-0.61190	0.90307
4	0.041558	1.49602	-0.54788	-0.48712	-0.31677
5	0.018182	-0.26877	-0.92365	-0.08385	-0.71994
6	0.000000	-1.25412	-1.17408	-1.30686	-1.09571
7	0.010811	-0.77867	-0.54990	-1.07532	-0.50064
8	0.033962	0.72674	-0.64347	-0.55675	0.92411
9	0.004283	-0.83245	1.72664	0.52132	1.11531
10	0.006803	-0.89797	-0.73562	0.06396	-0.69114
11	0.013986	-0.13193	0.65160	0.81959	1.12533
12	0.007752	-0.87717	-0.01877	-0.19079	-0.31589
13	0.009524	-1.06738	-1.21696	-3.30798	-1.60186
14	0.011173	-0.46793	0.05456	0.87714	0.95080
15	0.035165	1.33919	0.57937	0.90806	1.72426
16	0.031359	0.65393	-0.57745	-1.13545	0.48711
17	0.040486	1.75371	0.58840	0.31814	0.76649

Obs	HR/AB	Factor 1	Factor 2	Factor 3	Factor 4
18	0.023952	0.51855	0.77753	1.28790	2.16244
19	0.000000	-1.38120	-1.06983	-0.08136	-0.88594
20	0.041667	1.40689	-0.10842	0.13255	-0.50334
21	0.016097	0.02054	1.05629	1.56022	0.65850
22	0.029586	0.97929	1.42719	0.15806	1.81939
23	0.007177	-0.91899	1.55690	-0.51915	-1.57401
24	0.017682	0.14371	1.21054	0.62863	-0.56122
25	0.011338	-0.29895	0.29277	-0.19362	0.69475
26	0.011583	-0.28773	1.46859	0.28363	-0.11859
27	0.018868	-0.14582	-1.03277	-0.18376	0.63124
28	0.025000	-0.15410	-0.94108	0.04189	0.33838
29	0.049180	2.21327	0.46602	0.36790	1.16537
30	0.003125	-1.12040	-0.06148	-0.43881	-1.41997
31	0.004464	-1.25603	-0.96805	-0.65679	-1.70580
32	0.000000	-1.36477	-0.78922	-1.74560	-1.19811
33	0.006309	-0.96819	-0.52038	-0.27110	0.38122
34	0.000000	-1.32886	-0.07492	1.10823	-0.50558
35	0.004505	-0.94011	-0.62721	-0.56829	-0.73662
36	0.017978	0.09943	1.25343	0.40460	0.29478
37	0.031915	0.97988	0.58630	-0.09452	-0.37889
38	0.018987	-0.09178	-0.26958	-0.27989	-0.05458
39	0.015957	-0.54398	-0.62132	-0.78132	-0.30149
40	0.002146	-1.04472	0.86652	0.18577	-0.11546
41	0.023529	0.48077	0.48965	0.96594	0.45251
42	0.021008	-0.12998	-0.81144	-0.18722	-1.06381
43	0.000000	-1.31032	-0.62705	-0.43444	-0.61914
44	0.014963	-0.05373	0.53018	1.46842	1.21415
45	0.029570	0.64463	0.16907	1.12031	-1.16374
46	0.005882	-0.89073	-0.18170	0.05249	-0.14668
47	0.042071	1.29684	-0.38489	1.45550	0.77164
48	0.012579	-0.46612	-0.94235	-0.10966	0.08212

Obs	HR/AB	Factor 1	Factor 2	Factor 3	Factor 4
49	0.039749	1.59223	1.37378	-0.25008	2.19744
50	0.006231	-0.79712	-0.04173	0.36530	-0.44134

Appendix C

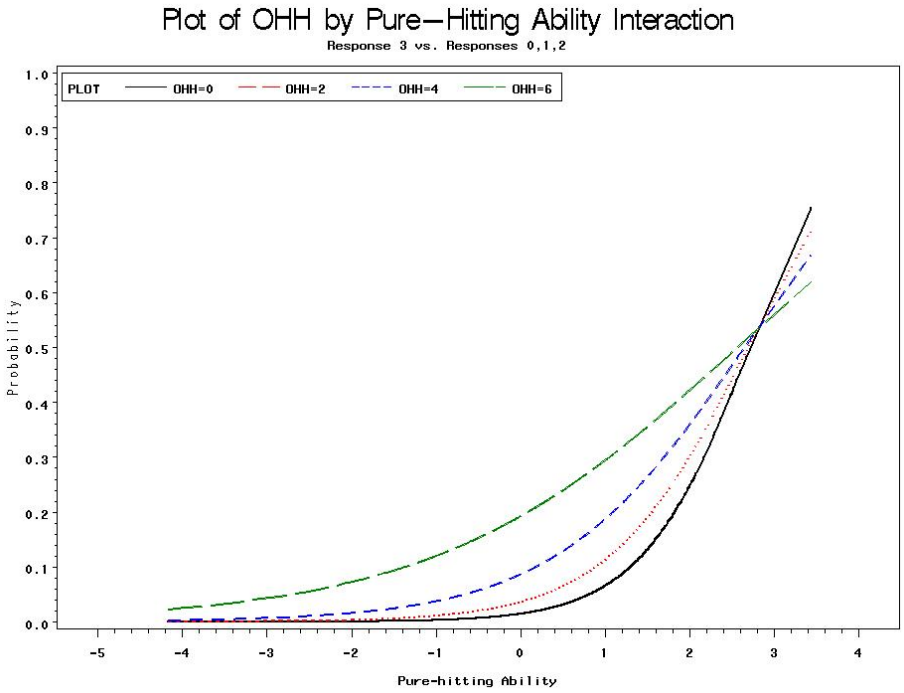

Figure 26

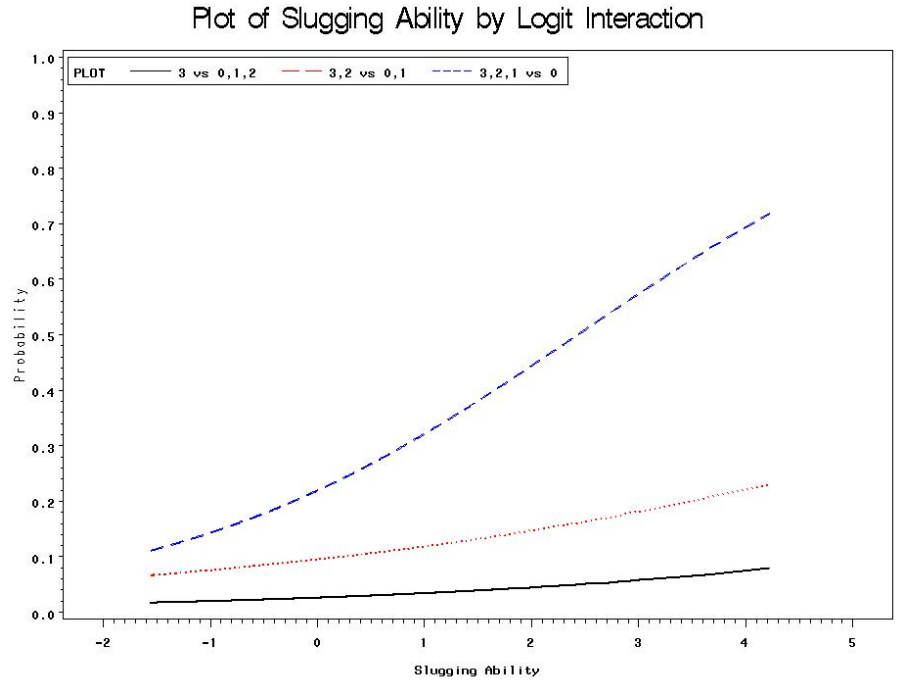

Figure 27

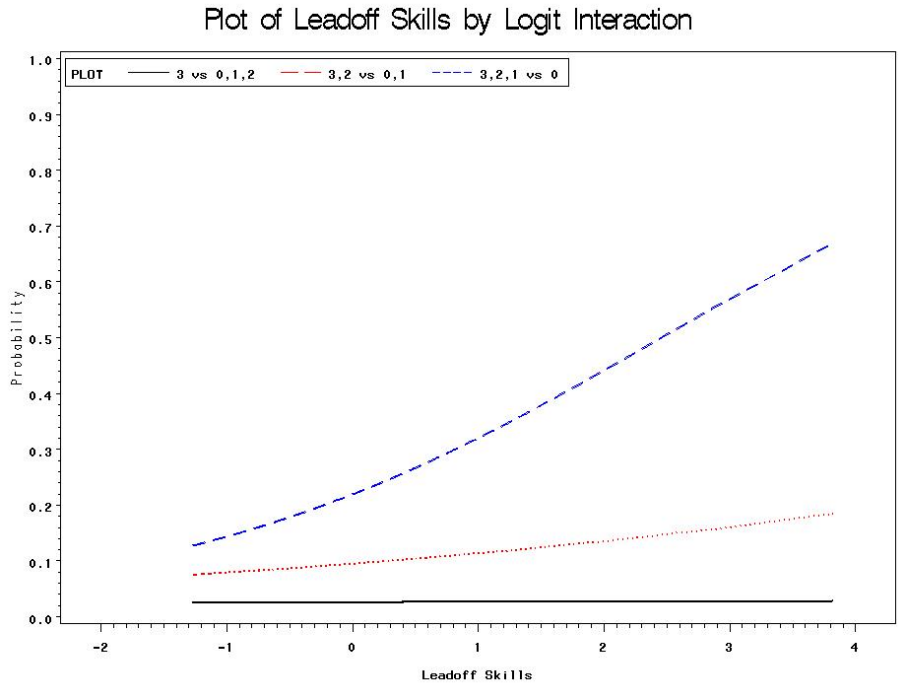

Figure 28

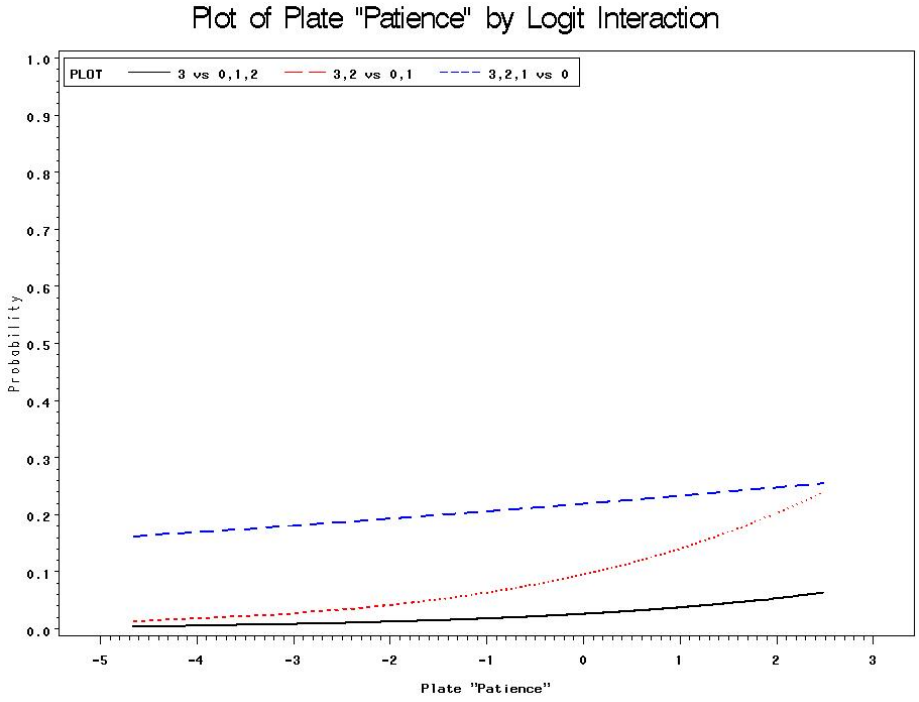

Figure 29

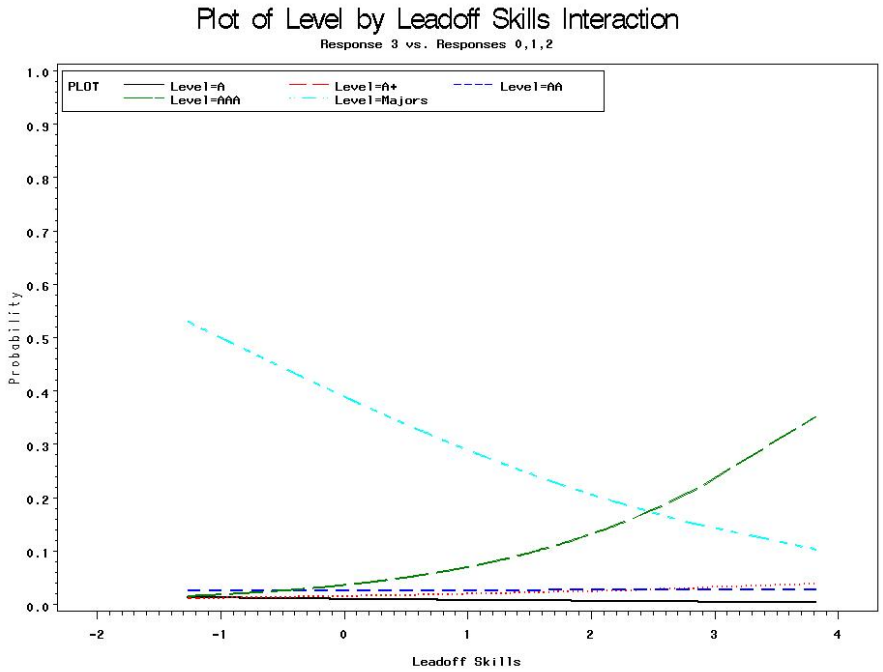

Figure 30

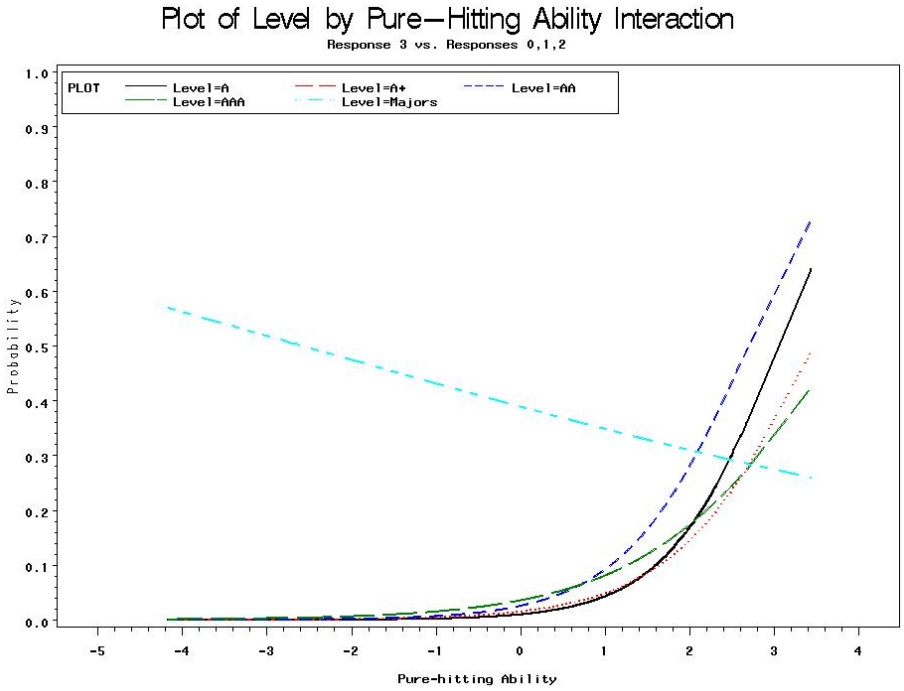

Figure 31

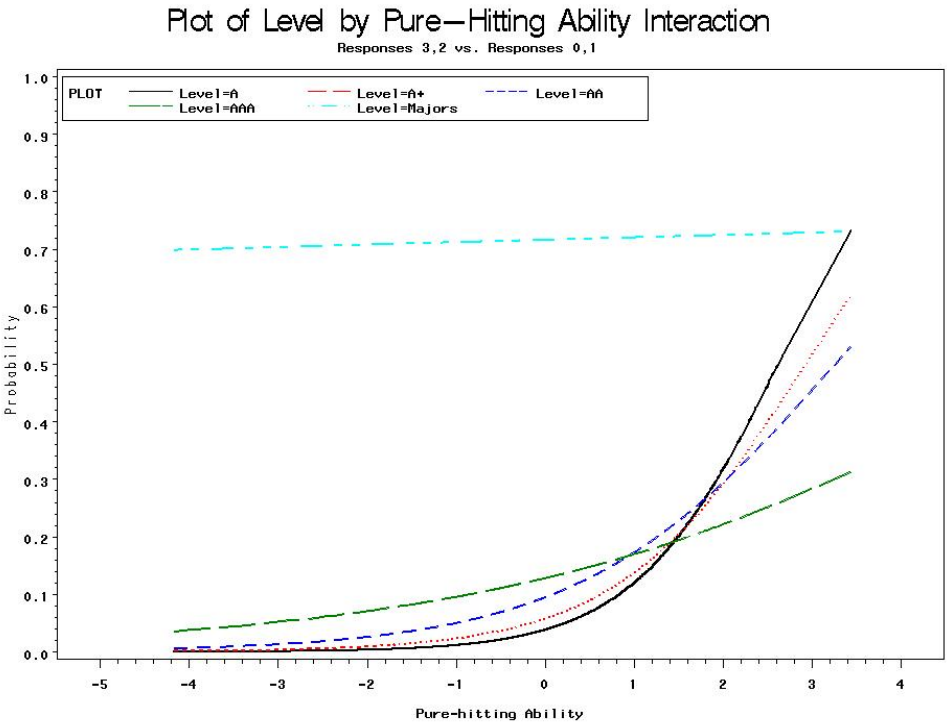

Figure 32

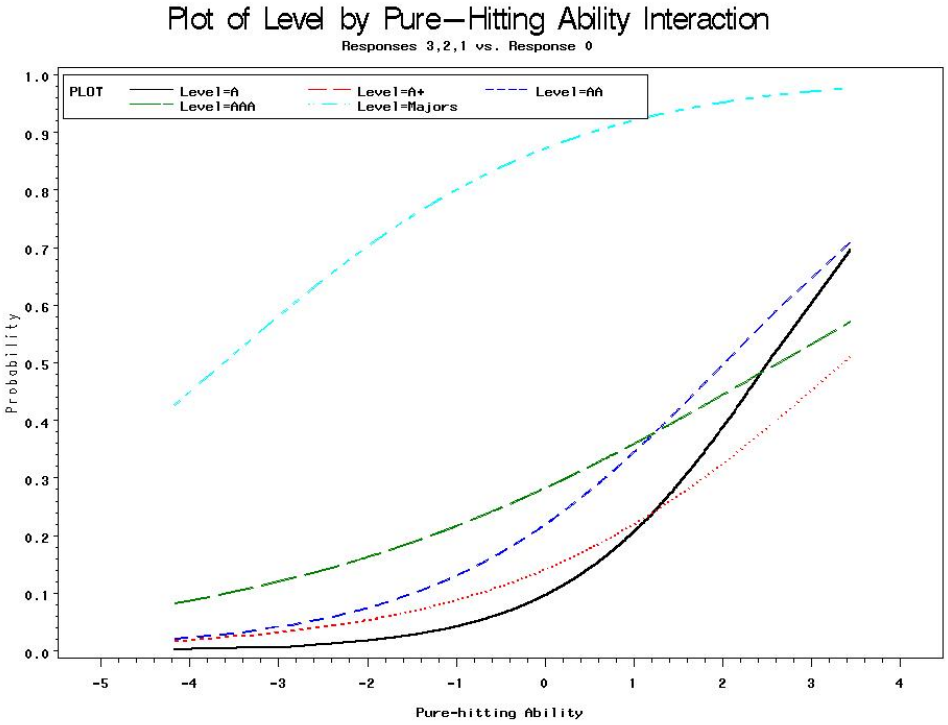

Figure 33

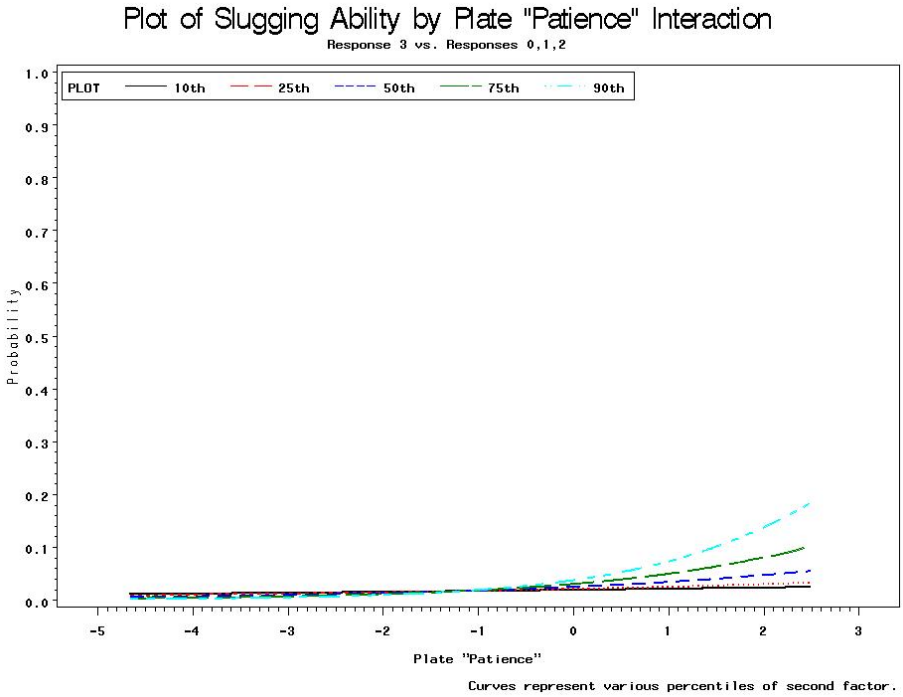
Plot of Slugging Ability by Plate "Patience" Interaction
Response 3 vs. Responses 0,1,2
PLOT
10th
25th
50th
75th
90th
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-5
-4
-3
-2
-1
0
1
2
3
Plate "Patience"
Curves represent various percentiles of second factor.

Figure 34

Appendix D

Analysis Of GEE Parameter Estimates								
Empirical Standard Error Estimates								
Parameter			Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept			-1.0225	0.2504	-1.5134	-0.5316	-4.08	<.0001
Logtype	1		2.3652	0.2388	1.8972	2.8331	9.91	<.0001
Logtype	2		1.3781	0.2026	0.9810	1.7753	6.80	<.0001
Logtype	3		0.0000	0.0000	0.0000	0.0000	.	.
OHH			0.4619	0.0748	0.3153	0.6084	6.18	<.0001
Level	A		-4.1518	0.3341	-4.8066	-3.4970	-12.43	<.0001
Level	A+		-3.7219	0.3203	-4.3496	-3.0942	-11.62	<.0001
Level	AA		-3.1860	0.3102	-3.7939	-2.5781	-10.27	<.0001
Level	AAA		-2.8467	0.3448	-3.5225	-2.1709	-8.26	<.0001
Level	Majors		0.0000	0.0000	0.0000	0.0000	.	.
Factor1			0.2825	0.1445	-0.0007	0.5657	1.96	0.0506
Factor2			-0.4499	0.2604	-0.9602	0.0604	-1.73	0.0840
Factor3			0.3790	0.2234	-0.0589	0.8168	1.70	0.0898
Factor4			0.0286	0.2390	-0.4398	0.4969	0.12	0.9048
OHH*Factor4			-0.1654	0.0641	-0.2911	-0.0397	-2.58	0.0099
Factor2*Level	A		0.2224	0.3568	-0.4769	0.9216	0.62	0.5331
Factor2*Level	A+		0.6961	0.3368	0.0361	1.3562	2.07	0.0387
Factor2*Level	AA		0.4713	0.2852	-0.0876	1.0303	1.65	0.0984
Factor2*Level	AAA		1.1540	0.4503	0.2714	2.0367	2.56	0.0104
Factor2*Level	Majors		0.0000	0.0000	0.0000	0.0000	.	.
Factor4*Level	A		1.6835	0.4439	0.8134	2.5537	3.79	0.0001
Factor4*Level	A+		1.3755	0.3904	0.6103	2.1407	3.52	0.0004
Factor4*Level	AA		1.5235	0.3610	0.8159	2.2310	4.22	<.0001
Factor4*Level	AAA		1.0452	0.4173	0.2272	1.8632	2.50	0.0123
Factor4*Level	Majors		0.0000	0.0000	0.0000	0.0000	.	.
Factor1*Factor3			0.2414	0.1040	0.0376	0.4451	2.32	0.0202

Analysis Of GEE Parameter Estimates								
Empirical Standard Error Estimates								
Parameter			Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Factor1*Logtype	1		0.2402	0.1382	-0.0306	0.5111	1.74	0.0822
Factor1*Logtype	2		-0.0329	0.1118	-0.2520	0.1861	-0.29	0.7684
Factor1*Logtype	3		0.0000	0.0000	0.0000	0.0000	.	.
Factor2*Logtype	1		0.4956	0.1892	0.1248	0.8665	2.62	0.0088
Factor2*Logtype	2		0.1801	0.1653	-0.1438	0.5040	1.09	0.2757
Factor2*Logtype	3		0.0000	0.0000	0.0000	0.0000	.	.
Factor3*Logtype	1		-0.2988	0.2197	-0.7293	0.1318	-1.36	0.1738
Factor3*Logtype	2		0.0654	0.1763	-0.2801	0.4108	0.37	0.7107
Factor3*Logtype	3		0.0000	0.0000	0.0000	0.0000	.	.
Factor4*Logtype*Level	1	A	-0.6156	0.3389	-1.2799	0.0487	-1.82	0.0693
Factor4*Logtype*Level	1	A+	-0.6624	0.3529	-1.3541	0.0294	-1.88	0.0605
Factor4*Logtype*Level	1	AA	-0.7199	0.2843	-1.2772	-0.1626	-2.53	0.0113
Factor4*Logtype*Level	1	AAA	-0.5157	0.3929	-1.2858	0.2544	-1.31	0.1893
Factor4*Logtype*Level	1	Majors	0.7042	0.3786	-0.0379	1.4463	1.86	0.0629
Factor4*Logtype*Level	2	A	-0.2750	0.2651	-0.7946	0.2447	-1.04	0.2997
Factor4*Logtype*Level	2	A+	-0.2448	0.3134	-0.8591	0.3695	-0.78	0.4348
Factor4*Logtype*Level	2	AA	-0.6557	0.2137	-1.0745	-0.2368	-3.07	0.0022
Factor4*Logtype*Level	2	AAA	-0.5392	0.3051	-1.1373	0.0588	-1.77	0.0772
Factor4*Logtype*Level	2	Majors	0.1954	0.2460	-0.2867	0.6776	0.79	0.4269
Factor4*Logtype*Level	3	A	0.0000	0.0000	0.0000	0.0000	.	.
Factor4*Logtype*Level	3	A+	0.0000	0.0000	0.0000	0.0000	.	.

Analysis Of GEE Parameter Estimates								
Empirical Standard Error Estimates								
Parameter			**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Factor4*Logtype*Level	3	AA	0.0000	0.0000	0.0000	0.0000	.	.
Factor4*Logtype*Level	3	AAA	0.0000	0.0000	0.0000	0.0000	.	.
Factor4*Logtype*Level	3	Majors	0.0000	0.0000	0.0000	0.0000	.	.