It May Be a Great Day for Baseball, but Is It a Great Day for a Knuckleball?

Robert H. Carver[1]
[1]Stonehill College, 320 Washington Street, Easton, MA 02357

Abstract
The knuckleball is one of the rarest pitches in the repertoire of major league pitchers. It has the potential to confound batters with its unpredictable movement, which results from the absence of rotation on the ball and the interplay of the air turbulence and pressure differentials on the stitches and smooth surface of the baseball. The pitch is notoriously difficult to control, but when effective its slow speed and wide arc leaves batters extremely frustrated. Given the crucial role of air pressure and movement, one wonders if atmospheric and climatalogical variation influence a skilled pitcher's ability to control the knuckleball in a given game. This paper examines game day data for veteran knuckleballer Tim Wakefield of the Boston Red Sox and finds that at least part of the answer may be blowin' in the wind. The analysis is accessible to undergraduate students, illustrating the managerial utility of multivariate models.

Key Words: Baseball, variation, scatter plots, multiple regression models.

1. Introduction

Many statistics educators have advocated the use of sports-related data in teaching introductory statistics (Albert, 2003, 2010; Cochran, 2005; Minton, 1994). Certainly examples from sports appeal to many college-aged students, and come from familiar territory in contrast to many real-world alternatives. Sports data sets are readily available in the public domain and students often understand their context, origins, and meaning. In the United States, baseball is arguably the richest source of granular sports-related data. Admittedly, baseball examples are not for everyone and could alienate many students particularly in classes with an international student body. However, in the right setting a good baseball example can motivate the reluctant statistics student and can illustrate both the utility of statistical methods and the meaning of certain critical statistical concepts.

This paper focuses on one narrow aspect of the game: the curious behavior of a pitch known as the knuckleball. The erratic movement of the knuckleball is the result of pressure differentials on opposing sides of the baseball in flight. When thrown by a skilled pitcher the ball barely rotates as it makes its way towards home plate. The very absence of rotation leaves the ball under the control of the air flows around it which buffet the ball and cause irregular motion. As described below, the knuckleball is a rare pitch in major league play; only a few pitchers use it and it is notoriously difficult to control. As such, it provides an ideal setting for consideration of the meaning and implications of variability as the term is used in statistics. Moreover, students can readily understand the managerial and strategic advantages that lie in the control or failure to control the variability of the pitch.

As a fan of the Boston Red Sox and their veteran knuckleballer Tim Wakefield, I have long been curious about the volatility of this pitch. Casual observation suggests that there are games when the pitch is ineffective almost from the outset while on other days, the pitch is as good as any combination thrown by major league pitchers. Given the extent to which the ball's movement is influenced by air flow and pressure differentials, one wonders whether environmental factors can predict whether a given outing will be disastrous or successful. This curiosity led to two basic questions:

- How much, if at all, do atmospheric conditions affect knuckleball movement?
- Can pre-game weather conditions help to predict pitching performance for a knuckleball pitcher?

Because the questions in this study involve the relationship between weather and the behavior of a knuckleball, the study also provides a context in which to teach the underlying ideas of *conditional* distributions. It's a small step for students to internalize the idea that this particular pitch might behave differently depending on wind, humidity, barometric pressure—collectively known as atmospheric conditions.

Finally, because baseball fans in the class are presumably comfortable with the notion that a pitcher's effectiveness on a given day depends on a variety of factors, this example is a prime candidate for introducing the basic multiple regression model. Additionally the dataset and study described here provide opportunities to illustrate techniques and concepts relevant to study design, measurement, data preparation, proxy variables, multivariate modeling, and interpretation of model estimates.

2. Literature Review

The theme of the 2011 Joint Statistical Meetings is "Statistics: An All Encompassing Discipline." One might apply an analogous phrase to baseball. Scholarly and literary interest in baseball has deep roots, with analyses coming from an array of disciplines. Novelists and journalists have long found baseball to provide rich context for fiction and for engaging behind-the-scenes reporting (Goodwin, 1998; R. Kahn, 1987; King, 1999; Lewis, 2004; Malamud, 2003; Will, 1991). Economists have built valuation models to assess the contributions of players to their team and statisticians have examined the long-term performance of individual players (Depken, 2000; L. M. Kahn, 1993; Koop, 2002; Scahill, 1990; Schall & Smith, 2000; Scully, 1974).

Physicists have taken special interest, both from the standpoint of opportunities for engaging students in familiar real-world phenomena–baseball is just one favorite sport among physicists–and for improving the game. The now-classic source is R.K. Adair's little book *The Physics of Baseball* (Adair, 1990). Adair's book covers all aspects of the game. As more data have become available, others have zeroed in on the flight of the pitched or batted ball, typically with an eye to optimizing performance (Bahill & Baldwin, 2007; Clark, 2007; Kagan, 2009; Minton, 1994; Sawicki, Hubbard, & Stronge, 2003).

Statisticians have taken special interest in baseball for many years, often focusing on offensive statistics – perhaps because of their wide availability since the earliest days of the game, perhaps because offense is exciting and so many of the game's icons made their contributions as batters (Albert, 1994; Albright, 1993; Bennett & Flueck, 1983).

At the 1988 Joint Statistical Meetings the Section on Statistical Graphics sponsored an exposition on graphical analyses, emphasizing the promise of integrating then-developing computer graphics and more traditional analysis. Hoaglin and Velleman surveyed the submissions in the exposition and reported that "the full dataset [for the expo] included data on pitchers separately, but most respondents dealt only with the hitters; we likewise restrict our attention to hitters here." (Hoaglin & Velleman, 1995)

More recently, statistics educators have expressed considerable interest in the potential of baseball data to excite and engage students of statistics (Albert, 2003; Cochran, 2005). The emphasis on offense has been quite broad and persistent, probably due to the greater availability of hitting statistics. Watnik's dataset of MLB salaries is explicitly restricted to non-pitchers (Watnik, 1998). In several cases, it seems to go almost without saying that studies should rely on offensive data: Berry *et al.* express the "blind spot" in classic fashion: "Baseball is rich in data. We have data on every player (non-pitcher) who has batted in Major League Baseball (MLB) in the modern era (1901-1996)" (Berry, Reese, & Larkey, 1999). Even as pitching statistics have gained attention the overwhelming volume of batting records often keeps the pitchers' statistics warming the bench, so to speak. Consider, for example Koop's 2002 article comparing players' performance (p. 711): "Data limitations make it difficult to model defensive performance. Accordingly, all pitchers are omitted from the sample…"(Koop, 2002).

Kahn's analysis treats pitchers and non-pitchers separately, as do Schall & Smith (L. M. Kahn, 1993; Schall & Smith, 2000). Lackritz incorporated some pitching statistics in his models of salaries(Lackritz, 1990). Rosner *et al.* did ground-breaking work about 15 years ago, but the more recent explosion of data collection has stimulated much more work on pitching, thanks to the availability of play-by-play data for every game starting in 1984 (Rosner, Mosteller, & Youtz, 1996).

1984 was a watershed year for pitching data when Dick Cramer, in collaboration with Bill James, established STATS, Inc. This new enterprise began gathering and processing unprecedented bodies of play-by-play information which in turn they sold to ball clubs (Lewis, 2004). STATS began to provide support to major league teams as well as broadcast networks (beginning with ESPN), and the data deluge also led to increased exploration by legions of amateur analysts and to the adoption of sabermetrics-based management by a few teams, most notably the Oakland Athletics.

The data stream has become far deeper in recent years, so that we now have access not just to play-by-play observations, but to *pitch-by-pitch* data. Highly detailed information about the physical and game-relevant attributes of every pitch, every swing, every umpire's call are now streaming in real time. Through the services of retrosheet.org and MLB's Pitchf/x database (described fully in Section 4), there is now a dazzlingly rich body of data with which to excite and engage students interested in baseball.

3. Pitch Movement and Pitcher Performance

The encounter between pitcher and batter is a major source of drama in baseball. This is a complex matchup of physical and psychological mettle as well as game theory and strategy. The effective pitcher typically tries to combine physical execution with smart pitch selection, while the best hitters study the pitchers, looking for "tells" to anticipate the next pitch. Successful hitting appears to be a delicate combination of cognition and athleticism (Gray, 2002); the same can be said of successful pitching.

Pitchers control the ball and vary pitch selection, speed, and location in an effort to fool and/or overwhelm the batter. A full account of the physics underlying the flight of different pitches is well beyond the scope of this article, but among the accessible sources are Adair, Clark and Minton (Adair, 1990; Clark, 2007; Minton, 1994). For the purposes of this paper and of the class presentation it is sufficient for students to understand that by adjusting the speed, orientation and spin on the ball, a pitcher can cause the pitched ball to deviate from a straight line and from the predictable effects of gravity.

In the discussion and analysis that follow, we adopt this standard convention to represent the three-dimensional space between pitcher and batter:

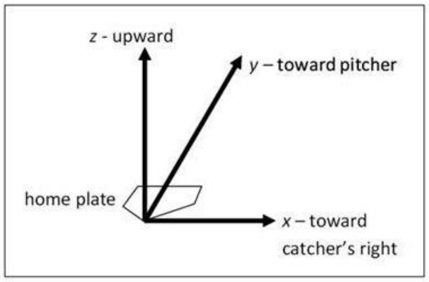


Figure 1: Axis Orientation to Describe Pitch Movement
Image source: Kagan, 2009.

If a pitcher were to throw a "straight line" pitch, it would move in the Y direction towards home plate, would not deflect at all on the X axis, but gravity would account for calculable negative movement along the Z axis. By spinning the ball as it is released, the pitcher induces additional movement in the X direction.

Because the stitches on a baseball protrude from the surface, orientation of the ball axis combines with the irregular surface of the ball to cause asymmetric disturbances in the air as the ball flows through. By varying the spin rate and orientation, these asymmetric and changing disturbances in turn cause differences in air pressure on opposing sides of the ball and the ball's trajectory bends. This is the Magnus Effect (see Figure 2).

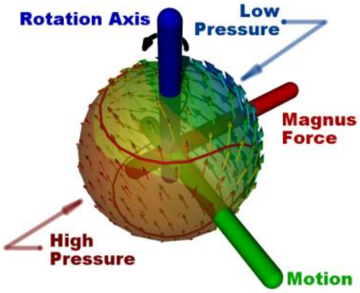
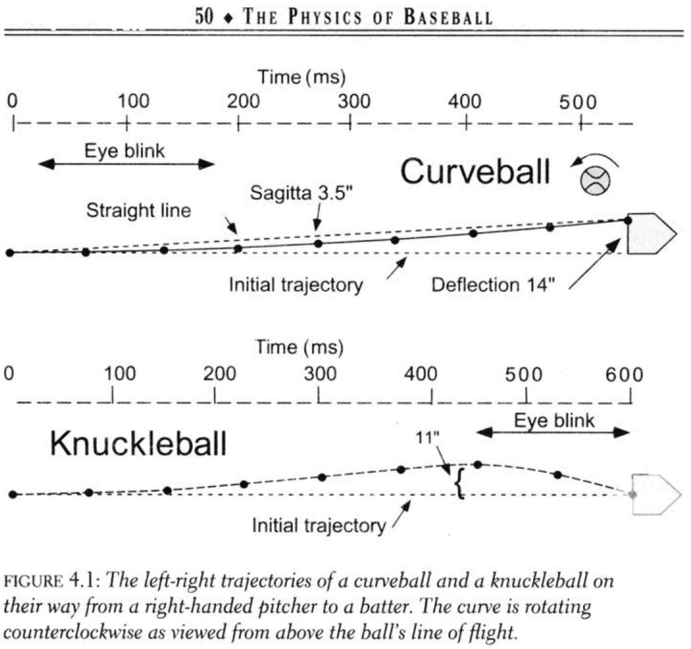


Figure 2: Pressure Differentials leading to the Magnus Effect, the Source of Pitch Movement
Image source: http://sv.wikipedia.org/wiki/Fil:Curveball-magnus-effect.jpg

When throwing a knuckleball, the pitcher imparts minimal spin on a relatively slow pitch. The initial speed of a knuckleball at 60-70 mph, spin rate of 30 revolutions per minute (rpm), which equals 1/2 a revolution and 1/4 of a revolution between pitcher's release and bat-ball contact point (Bahill & Baldwin, 2007). The very *absence* of rotation magnifies the Magnus Effect, and the flight of the knuckleball becomes quite unpredictable to observers—particularly to batters, catchers and umpires. Figure 3, below, from Adair (2002, p.50) compares the typical flight of a curveball and knuckleball. In the lower part of the figure we see the *break* of the knuckleball, defined as the maximum X-axis deflection from its initial straight trajectory. The unpredictability of the knuckleball accounts both for its rewards and risks, according to Adair (p. 55), and opens the door for very useful class discussions of the foundational concept of *variability* and of its importance both in baseball and in statistical investigations:

> "The disadvantage of the knuckleball … is that the forces can vary strongly with very small differences in orientation of the ball; hence the pitcher, however skilled, finds it very difficult to control the pitch. If it breaks sharply, it is difficult to catch and leads to too many passed balls. If it doesn't break, it is no more than a batting practice pitch, which is even worse."



50 ♦ THE PHYSICS OF BASEBALL

FIGURE 4.1: *The left-right trajectories of a curveball and a knuckleball on their way from a right-handed pitcher to a batter. The curve is rotating counterclockwise as viewed from above the ball's line of flight.*

Figure 3: Typical Trajectories of Curveball and Knuckleball

On some days in some innings the knuckleball baffles and humiliates good batters. As noted earlier, the pitch is notoriously difficult to hit except when it is notoriously easy to hit (garik16, 2010; Kottke, 2008).

Both the movement of the ball and the *dispersion* of the movement are key to its effectiveness. Consider the scatterplots in Figure 4. The left panel shows the X-Z movement for pitches thrown by starting pitchers Tim Wakefield (left panel) and Josh Beckett during the 2009 season.

The scales on both graphs are identical though the color coding is not. In the left panel, the blue points are knuckleballs—Wakefield's near exclusive pitch. In contrast, the right panel represents several pitch varieties offered by Beckett. The blue, green, and brown points are fastball varieties. The yellow are changeups and the red are curve balls.

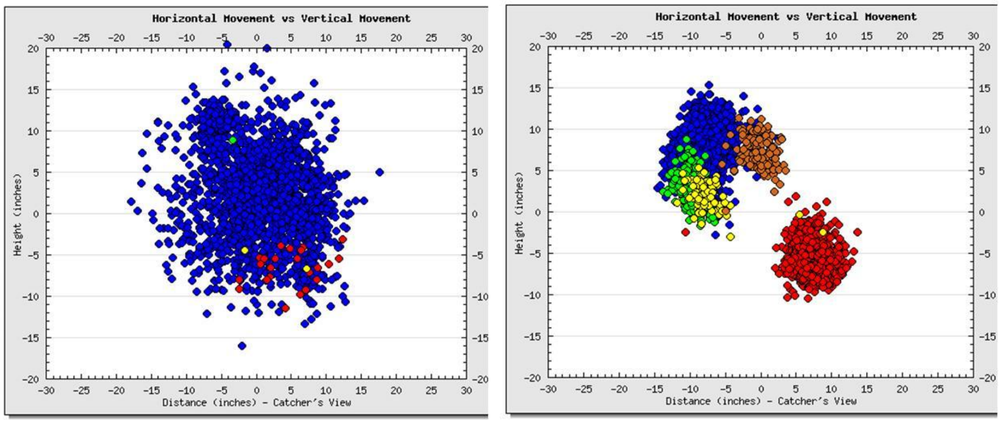


Figure 4: Pitch Movement and Variability—Tim Wakefield and Josh Beckett, 2009 Season

If a batter correctly predicts that a curve ball is on the way, his prediction is loaded with useful information. Curve balls tend to move in comparatively consistent ways. The batter has the chance to adjust his swing to accommodate deviations from the initial trajectory, albeit with very little time to do so.

With Wakefield, batters almost know for sure that the next pitch will be a knuckleball, but this knowledge has almost no practical utility. In classroom discussion, Figure 4 provides a rich set of teaching opportunities to stimulate and direct student thinking about bivariate distributions and their comparisons, about conditional distributions, about informative and non-informative priors, about common-cause and special-cause variation, and about just what those special causes might be in a given situation.

When the ball lacks its unpredictable movement the advantage goes to the hitter, leading many observers and participants to wonder just what accounts for sudden changes in fortune. The apparent unpredictability of the knuckler is the stuff of baseball lore, journalistic flights of fancy, even poetry (McCaffery, 1998):

A drop of sweat left on the ball? A stray mosquito?
The fat guy in the fourth row who blew on his coffee?
Who knows? Who knows how this pitch makes itself?

To reframe these verses in terms of statistics education, we have a response variable and an explanatory theory based on several factors including speed, rotation, and orientation. Physicists even can supply a functional form. Taken together, this model would appear to account for some of the variation in the response variable, but what accounts for the residual variation.

This can lead to a very fruitful class brainstorming session among knowledgeable or opinionated student fans seeking candidate factors that might reduce the unexplained variation. Earlier this season following a particularly effective day’s work, Tim Wakefield told a local sports reporter “The more humid the air is, the more resistance [the knuckleball] has against it, so it's going to move a little more” (brackets in original) (Abraham, 2011). Wakefield’s recent memoir sheds a bit more light: “Whether the result of humidity, wind, or maybe nothing at all, the knuckleball is can move dramatically on one pitch and do relatively little on the next, giving managers little advance warning that a pitcher is about to implode" (p. 19) (Wakefield & Massarotti, 2011).

In an account of the first game of the 2004 World Series, in which Wakefield was the starting pitcher, the weather figured significantly in his thinking:

> By the time Game 1 finally started at precisely 8:09 PM, the New England weather, precisely as one would expect, was something of a factor. The game-time temperature was a crisp 49 degrees, but of far greater concern to Wakefield was the wind blowing in [from center field]… **Wakefield prefers a slight wind blowing against him**... **Wakefield found that wind resistance increased the movement on his pitches. Again, it was a matter of physics. A well-executed knuckler, thrown against the wind, resulted in more acute movement of the pitch** (*ibid.,* p. 208, emphasis added).

In other words, this accomplished practitioner has found that his effectiveness depends in part on some combination of things under his control and on wind and humidity.

4. Data Sources, Study Design, and Teaching Opportunities

The research questions set the stage for an in-class discussion of measurement and study design. The first question might be pursued via experimental design but the second question implies realistic game conditions and therefore demands observational data. Introductory students rather quickly discover the many issues and tradeoffs involved, and the need for operational compromises becomes clear. Is it possible to obtain accurate real-time weather data for each pitch? What kinds of measurements are available for the movement of pitches? What are the best or most suitable measures of pitch movement and pitcher performance? Is it important to control for stadium factors like altitude, domes, or orientation?

Like any *post hoc* observational study, we are limited here by the availability of data, but fortunately the available datasets are quite extensive and easy to access. it will help to start with a description of four primary data sources and then lay out the critical study design choices that I made.

4.1 Data Sources

This study relies on four major sources of data, briefly described here:

- **www.retrosheet.org**: This organization captures play-by-play data for every game in Major League Baseball.
- **www.baseball-reference.com**: This site tabulates data initially captured by retrosheet, providing game-by-game summaries for each MLB ball game.
- **MLB Pitchf/x** database: This database contains pitch-by-pitch observations for every pitch thrown in a major league game starting in 2008 (expanded data available in 2009).
- **National Climactic Data Center**'s Hourly Surface Data: This GIS-based searchable database provides detailed weather and climatalogical observations gathered by National Weather Service stations and other observatories around the world ("National Climactic Data Center (NCDC) Geodata Portal," 2011).

With the exception of the Pitchf/x data, all of these sources provide a straightforward interface to generate or capture downloads in *.csv or *.xls format. Though one can obtain the Pitchf/x data directly, manipulation requires custom programming. Fortunately, several website developers have built free front-end query tools and these are quite easy to navigate. Among the best are those provided by Kalk, Lefkowitz, and Brooks (Brooks, 2011; Kalk, 2007; Lefkowitz, 2011). For extremely helpful introductions to the data available in the pitchf/x database, see Nathan and Albert (Albert, 2010; Nathan, 2010).

Variables from these four sources were combined into a single JMP data table, joining the four original data tables by date and game starting time. NCDC observations are recorded approximately hourly; the observations most closely preceding game time were used. For most introductory students, the details of downloading and integrating data from disparate sources would be excessively confusing. Students may benefit from a description of the process of locating, selecting, and consolidating the data to understand that this is not a mysterious process.

4.2 Design Considerations

To investigate the two research questions it was necessary to make choices, and these too are worth discussing in class. Some these choices were driven by data availability, while others were matter of limiting the scope of the project.

During his career, Wakefield has worked both as a starting pitcher and a reliever. Given the differences in preparation for starters as well as the managerial nature of research questions, I limited the analysis to the performance of starting pitchers in each game omitting relief appearances.

Though we have detailed data for every pitch of every major league game, the data about weather conditions are typically available hourly and there is no reliable way to identify the prevailing conditions in a given ballpark at a particular moment in a game. Hence I decided to record a single set of weather-related variables at game time, and apply them to the entire game. Clearly, this is not an ideal design but as a starting point in the investigation seems reasonable. This choice is closely related to the decision to make the **game** the basic unit of analysis, rather than the pitch, the at-bat, or the inning.

Each stadium is slightly different, and some of those differences may have implications for a knuckleball pitcher. There is some variation in the compass orientation of ballparks, so that a wind from one direction might be blowing in at one park, but across the diamond at another. Each stadium lies at a different altitude, affecting air density and barometric pressure, and wind patterns within ballparks also vary. To control for ballpark variation, the analysis is restricted to home games played by the Red Sox at Fenway Park.

Baseball teams play 81 games each season at home, and at most a pitcher enjoying uninterrupted health might expect to appear as a starter in perhaps one-fourth of those games. As such, a single season provides a small sample of games for any individual pitcher. For this reason, I decided to observe Wakefield's performance over the course of several seasons.

The first phase of the analysis focused on knuckleball movement. I restricted the sample to Wakefield's knuckleballs during his home starts for seasons 2008 through 2010.

According to Pitchf/x database in this period Wakefield threw a knuckleball 99.6% of the time. This analysis omits the non-knucklers.

For the second phase of the analysis dealing with pitching effectiveness, I wanted to compare Wakefield's performance to all of the other starters on the Red Sox staff. For this model, I used combined data from 2007-2009, holding the 2010 season out for later validation. In that three-year span, Wakefield started 40 games and the rest of the staff started 203 home games.

Across the data sources there is some redundancy in the sense that particular variables appear in two sources. In most instances, the observations are consistent and can be readily reconciled. One notable exception comes when we look at the weather data from the NCDC and retrosheet.org. In the discussion that follows, the ballpark-specific weather conditions come from Retrosheet, following the logic that the NCDC observations are made at Logan Airport, several miles from Fenway and situation at the edge of Boston Harbor where both temperature and wind direction may well vary from the stadium. On the other hand, Retrosheet does not report relative humidity, dew point or barometric pressure so those come from the NCDC.

5. Models and Variable Definitions

This section presents two models corresponding to the two research questions. The analysis is essentially exploratory. The physics involve well-developed theory, but it is not clear whether or how atmospheric and weather-related variables should appropriately enter either model. In some instances we have several plausible measures of a single construct. *A priori* there is little reason to expect one measure to be superior to another. Hence, the models are presented below in terms of constructs and candidate measures. The estimates presented in Section 6 were developed through a combination of stepwise regression and author judgment.

5.1 Pitch Movement Model and Variables

The earlier discussion suggests that pitch movement is a function of the pitcher's skill in launching the knuckleball as well as various atmospheric and climatalogical conditions. The model presented here includes three measures of pitcher skill and five environmental condition variables. As noted above, other factors are controlled by the design of the study. The Pitchf/x database includes several measurements of pitch movement In this part of the study, "break length" is the dependent variable. Break-length "is the largest deviation, in inches, of the actual from the straight-line trajectory" (Nathan, 2010). For each game which Wakefield started, I computed the mean break length of all pitches during his starting appearances. The candidate explanatory factors are listed in Table 1, with a short rationale for their inclusion in the model:

Table 1: Candidate Explanatory Variables for Average Break Length

Variable	*Definition and Rationale for inclusion*
Mean(Spin Axis)	Pitchf/x records axis of rotation (1-360 deg) for each pitch. This is the simple arithmetic mean of that measure. Pitcher attempts to optimize break by orienting seams in a specific way.
StdDev(Spin Axis)	Standard deviation of spin axis. As deviation increases, break length may decline from maximum.
Mean (Spin Rate), rpm	The ideal knuckleball barely spins. Increases in spin rate should

	diminish mean break length, other things being equal.
Wind Speed, mph	NCDC-reported wind speed prior to game; *a priori* it is difficult to say whether increased wind speed would have a positive or negative impact on pitch movement
Crosswind Dummy	Equals 1 if wind is blowing across the diamond.
Dewpoint (° F)	Dewpoint is a measure of humidity, representing the temperature at which the air is saturated. Higher dewpoints correspond to higher humidity; conventional baseball wisdom is that higher humidity is associated with more movement.
Temperature (° F)	Theory is not clear as to whether temperature should effect movement, but is included as a control.
Barometric Pressure at Sea Level (inches)	NCDC reports sea-level pressure; Fenway Park lies nearly at sea-level. Given the physics of the knuckleball, it is reasonable to expect some impact of high pressure, but it is not clear whether that would increase or decrease movement.

5.2 Pitching Effectiveness Model and Variables

In contrast to pitch movement, the choice of an appropriate dependent variable is more challenging in the case of pitching effectiveness. The traditional measure is game Earned Run Average (ERA). Section 10.18 of the Official Rules of MLB define an earned run in explicit detail, enumerating the conditions under which a run is considered earned ("Official Rules of Baseball," 2010). Essentially, runs scored are earned by the starting pitcher unless they come as the result of an error or interference. Game ERA is a normalized figure extrapolating the number of earned runs over 9 innings. Higher ERAs indicate comparatively ineffective pitching performance.

The trouble with ERA as a metric is that earned runs do not depend exclusively on the performance of the pitcher. Once a ball is put into play by a batter then many other factors beyond the pitcher's control enter the picture. In particular, the consequences of a batted ball depend heavily on the positioning and effectiveness of defensive play. The sabermetrics literature is replete with lengthy discussions of the shortcomings of pitching metrics, because of the interdependence of defense. Many "defensive independent pitching statistics" (DIPS) have been proposed, but none are yet widely accepted as standard (Basco & Davies, 2010).

The difficulties of devising a "pure" measure of pitching effectiveness can form the basis for useful discussions in the classroom—what exactly do we measure? What are the strengths and weaknesses of alternative measurements and summary statistics? To what degree does a particular statistic authentically capture a specific construct?

Section 6 reports and interprets the estimated model. In this phase of the analysis, I estimated the model twice. First I used the games which Wakefield started, and then I used games started by the rest of the pitching staff.

In this model, there are three groups of variables listed in Table 2. The first (Days Rest) should capture a pitcher's readiness to pitch. The second (Opp Team Avg) is the end-of-season team batting average of the opposing team. Clearly this is unknown at game time, but this measure serves as a proxy for the quality of the opposing team hitters. Other things being equal, most pitchers would tend to have a lower game ERA pitching against a weak (low average) lineup. The remaining five variables all measure different aspects of the weather and atmospheric conditions.

Table 2: Candidate Explanatory Variables for Pitching Performance

Variable	*Definition and Rationale for inclusion*
Days Rest	The number of days elapsed since a pitcher's previous start. For most pitchers, performance should improve with additional rest, so the expected sign is negative: ERA should be lower with more rest.
Opp Team Batting Avg	End-of-season team batting average of the opposing team, which serves as a proxy for the quality of the opposing team hitters. Other things being equal, the sign of this coefficient should be positive.
Wind Speed, mph	NCDC-reported wind speed prior to game; *a priori* it is difficult to say whether increased wind speed would have a positive or negative impact on pitch movement
Windout Dummy	Equals 1 if wind is from home towards centerfield.
Relative Humidity	Whereas dewpoint is an absolute measure (controlling for pressure) this is a measure relative to temperature. Conventional baseball wisdom is that higher humidity is associated with more movement.
Barometric Pressure at Sea Level (inches)	NCDC reports sea-level pressure. Given the physics of the knuckleball, it is reasonable to expect some impact of high pressure, but it is not clear whether that would increase or decrease movement.

6. Findings

6.1 Knuckleball Movement Model

Table 3 summarizes the regression estimates for the first model. The table presents two iterations of the model, where the first set of estimates include all of the variables shown in Table 1, and the second set of estimates drops the insignificant variables.

The magnitude of coefficients—inches of mean maximum lateral break per unit change in the factor—are less important to this discussion than the sign and significance of the estimates. As noted earlier, the dataset is based on 22 starting appearances by Tim Wakefield between 2008 and 2010, analyzing only knuckleballs.

Table 3: OLS Estimates of the Average Break Length Model

Variable	*Estimates — full model*	*P-Value*	*Estimates— reduced model*	*P-Value*
Intercept	16.1697	0.1807	10.3162	< 0.0001
Mean(Spin Axis)	0.0235	0.0003	0.0233	< 0.0001
StdDev(Spin Axis)	0.0149	0.0201	0.0144	0.0136
Mean (Spin Rate), rpm	–0.0016	0.1214	–0.0017	0.0574
Wind Speed, mph	–0.0154	0.3761		
Crosswind Dummy	–0.7052	0.0040	–0.7000	0.0014
Dewpoint (° F)	–0.0110	0.3003	–0.0122	0.0796
Temperature (° F)	0.0032	0.7565		
Barometric Pressure at Sea Level (inches)	–0.2048	0.5935		
R^2 | Adj R^2	0.79 | 0.66		0.77 | 0.70	
F	6.017	0.0023	10.787	0.0001

In the reduced model, we find that four factors are significant at the 0.05 level and the other two are significant at the 0.10 level. The model accounts for roughly 77% of the variation in break length and includes variables related to Wakefield's actions and to weather conditions. Both the orientation of the pitch (mean of spin axis) and the *variability* of the orientation (standard deviation of the axis) positively affect the break length. It is mildly surprising that varying the spin axis increases the break length. As expected, lower spin rates are associated with larger break other things being equal.

The presence of a crosswind appears to reduce the mean break length regardless of wind speed, and higher humidity (dewpoint) is associated with lower break length after controlling other factors. This is surprising in light of the belief—held by Wakefield and other baseball insiders—that higher humidity enhances the effectiveness of the knuckler.

The residuals (not shown here) reveal no important violations of regression assumptions and there is no evidence of collinearity. This model provides opportunities to illustrate several important concepts in multiple regression modeling, including hypothesis development, model specification, interpretation of coefficients, and the practical and managerial implications of estimates. It also nicely exemplifies a situation in which multiple factors simultaneously influence the variation in a response variable.

6.2 Pitching Effectiveness Model

For this model, I divided the dataset into two subsets: games started by Wakefield and those started by all other pitchers. Table 4 reports the estimation results. In contrast to the pitch movement model, this model has relatively little explanatory power especially for the games started by pitchers other than Wakefield.

Table 4: OLS Estimates of the Game ERA Model

Variable	*Wakefield*	*P-Value*	*Other Starters*	*P-Value*
	Starts (n=40)		*(n=203)*	
Intercept	–19.4312	0.1700	–7.5236	0.6983
Days Rest	0.0103	0.4781	**0.0148**	**0.0139**
Opposing Team Batting Avg	**–25.2558**	**0.0106**	–0.6370	0.9620
Wind Speed, mph	**–0.0692**	**0.0056**	–0.0023	0.9515
Wind Out Dummy (=1 when wind blows out to center field)	**–0.4201**	**0.0410**	–0.0035	0.9796
Relative Humidity	0.0082	0.1609	–0.0042	0.9492
Barometric Pressure at Sea Level (inches)	**1.0136**	**0.0340**	0.3992	0.5419
R^2 | Adj R^2	0.46 | 0.36		0.03 | 0.001	
F	4.709	0.0015	1.049	0.3947

One immediately striking result is that the variables which are significant in Wakefield's case are *not* significant for the other pitchers[1] and likewise, the one variable that may have explanatory power for the others is irrelevant in the Wakefield case. Other pitchers' Game ERAs appear to be insensitive to weather conditions, and may be increased slightly with a longer hiatus between appearances. Given the weakness of the second set of estimates, there was little point in evaluating residuals. For the estimates of Wakefield's performance, residual analysis raises no questions about OLS assumptions, and again there is no indication of collinearity.

Wakefield's effectiveness seems to be unrelated to frequency of starts and this aligns with conventional wisdom that knuckleballers can take the mound regardless of rest. More interestingly, his Game ERA benefits from stronger winds, from wind blowing out (recall that he prefers to pitch into the wind), and from lower barometric pressure. The estimates also indicate that after accounting for these atmospheric conditions, he is more

[1] This applies both to the other starters collectively and individually. Only the aggregate results are reported here.

effective (lower ERA) against teams with better batting averages. This result is surprising; one can speculate about a reasonable causal chain but it is difficult to offer a definitive plausible explanation.

The important takeaway message here is that three atmospheric factors *do* have predictive power for the knuckleball pitcher but *not* for the other members of the pitching staff during the study period. It is also very much worth noting that the atmospheric variables that seem to influence pitch movement are different from the ones that influence ERA. Several factors might account for the differences: ERA is defense-dependent and not as much under the pitcher's control as pitch movement; batters and pitchers do attempt to adjust to expected pitch movements, and those adjustments may be imperfect; the two models are estimated using different seasonal samples.

7. Discussion and Further Research

This paper has laid out a comprehensive example of a research investigation targeted to statistics students who share a common interest in baseball. The study draws on common knowledge among baseball fans and players, illustrating the use of massive observational databases that have recently become available. It leads students through one iteration of the entire process of a statistical investigation.

Along the way, there are numerous opportunities to engage students in serious thought about critical concepts such as study design, representative sampling, variable selection, and interpretation of multiple regression coefficients. The dataset that supports this study draws on several sources and though it should be and has been sanitized for use by students, there are still real-world complications embedded in the dataset that can enrich and deepen understanding of statistical practice. The study reported here has several limitations, already discussed, that provide opportunities for critical thinking and open the door for students to suggest ways to mitigate their impact or to design an improved follow-on study.

Some sensible improvements and enhancements include the following:

- Apply the findings to later seasons to test predictive value.
- Estimate similar models for away games as well as home games, controlling for stadium variables such as elevation, orientation, and presence of a dome.
- Examine pitch-by-pitch movement data, rather than using summary statistics per game, integrating NCDC data.
- Further examine "pitcher-only" (*i.e.* defense-independent) measures of effectiveness.
- Further examine relationship of pitch movement to outcomes.

Acknowledgements

Many thanks to Jim Albert for his very constructive and helpful suggestions, and to Tim Wakefield for inspiration.

References

Abraham, P. (2011, May 28, 2011). Wakefield Spot-on Again. *The Boston Globe*, p. 1.

Adair, R. K. (1990). *The Physics of Baseball* (3rd (2002) ed.). New York: HarperCollins.

Albert, J. (1994). Exploring Baseball Hitting Data: What About Those Breakdown Statistics? *Journal of the American Statistical Association, 89*(427), 1066-1074.

Albert, J. (2003). *Teaching Statistics Using Baseball*. Washington, DC: The Mathematical Association of America.

Albert, J. (2010). Baseball Data at Season, Play-by-Play, and Pitch-by-Pitch Levels. *Journal of Statistics Education [online], 18*(3), 27. Retrieved from www.amstat.org/publications/jse/v18n3/albert.pdf

Albright, S. C. (1993). A Statistical Analysis of Hitting Streaks in Baseball. *Journal of the American Statistical Association, 88*(424), 1175-1183.

Bahill, A. T., & Baldwin, D. G. (2007). Describing baseball pitch movement with right-hand rules. *Computers in Biology and Medicine, 37*, 1001-1008.

Basco, D., & Davies, M. (2010). The Many Flavors of DIPS: A History and an Overview. *Baseball Research Journal, 39*(2).

Bennett, J. M., & Flueck, J. A. (1983). An Evaluation of Major League Baseball Offensive Performance Models. *The American Statistician, 37*(1), 76-82.

Berry, S. M., Reese, C. S., & Larkey, P. D. (1999). Bridging Different Eras in Sports. *Journal of the American Statistical Association, 94*(447), 661-676.

Brooks, D. (2011). BrooksBaseball.net. from http://www.brooksbaseball.net/content.php

Clark, D. (2007). K-ball physics explained. Retrieved 7/10/2011, from http://www.oddball-mall.com/knuckleball/mego.htm

Cochran, J. J. (2005). Can You Really Learn Basic Probability by Playing a Sports Board Game? *The American Statistician, 59*(3), 266-272.

Depken, C. A. (2000). Wage disparity and team productivity: evidence from major league baseball. *Economics Letters, 67*, 87-92.

garik16. (2010, June 11, 2010). Why is it so hard to hit the knuckleball? http://www.amazinavenue.com/2010/6/10/1511336/why-is-it-so-hard-to-hit-the

Goodwin, D. K. (1998). *Wait Till Next Year*. New York: Simon & Schuster.

Gray, R. (2002). Markov at the Bat: A Model of Cognitive Processing in Baseball Batters. *Psychological Science, 13*(6), 542-547.

Hoaglin, D., & Velleman, P. (1995). A critical look at some analyses of major league baseball salaries. *The American Statistician, 49*, 277-284.

Kagan, D. (2009). The Anatomy of a Pitch: Doing Physics with PITCHf/x Data. *The Physics Teacher, 47*, 412-416.

Kahn, L. M. (1993). Free Agency, Long-term contracts and compensation for major league baseball: estimates from panel data. *The Review of Economics and Statistics, 75*(1), 157-164.

Kahn, R. (1987). *The Boys of Summer*. New York: Harper Collins.

Kalk, J. (2007). PITCHf/x tool by Josh Kalk. 2011, from http://baseball.bornbybits.com/php/combined_tool.php

King, S. (1999). *The Girl Who Loved Tom Gordon*. New York: Scribner.

Koop, G. (2002). Comparing the Performance of Baseball Players: A Multiple-Output Approach. *Journal of the American Statistical Association, 97*(459), 710-720.

Kottke, J. (2008). The break on the knuckleball. Retrieved 7/15/2011, 2011, from http://kottke.org/08/04/the-break-on-the-knuckleball

Lackritz, J. (1990). Salary evaluation for professional baseball players. *The American Statistician, 44*, 4-8.

Lefkowitz, J. (2011). Joe Lefkowitz's PitchF/X Tool. 2011, from http://www.joelefkowitz.com/pitch.php

Lewis, M. (2004). *Moneyball: The Art of Winning and Unfair Game*. New York: Norton.

Malamud, B. (2003). *The Natural*. New York: Farrar, Straus and Giroux.

McCaffery, R. J. (1998). Chaos Theory and the Knuckleballer. *Ploughshares, 24*(4), 125-126.

Minton, R. (1994). A Progression of Projectiles: Examples from Sports. *The College Mathematics Journal, 25*(5), 436-442.

Nathan, A. M. (2010, March 8, 2010). MLB PITCHf/x Data. Retrieved 7/25/2011, from http://webusers.npl.illinois.edu/~a-nathan/pob/tracking.htm

National Climactic Data Center (NCDC) Geodata Portal. (2011). from http://gis.ncdc.noaa.gov/map/isd/

Official Rules of Baseball. (2010). Available from http://mlb.mlb.com/mlb/downloads/y2011/Official_Baesball_Rules.pdf

Rosner, B., Mosteller, F., & Youtz, C. (1996). Modeling Pitcher Performance and the Distribution of Runs per Inning in Major League Baseball. *The American Statistician, 50*(4), 352-360.

Sawicki, G. S., Hubbard, M., & Stronge, W. J. (2003). How to hit home runs: Optimum baseball bat swing parameters for maximum range trajectories. *American Journal of Physics, 71*(11), 1152-1162.

Scahill, E. M. (1990). Did Babe Ruth Have a Comparative Advantage as a Pitcher? *The Journal of Economic Education, 21*(4), 402-410.

Schall, T., & Smith, G. (2000). Do Baseball Players Regress Toward the Mean? *The American Statistician, 54*(4), 231-235.

Scully, G. (1974). Pay and Performance in Major League Baseball. *The American Economic Review, 64*(6), 915-930.

Wakefield, T., & Massarotti, T. (2011). *Knuckler: My Life with Baseball's Most Confounding Pitch*. Boston: Houghton Mifflin Harcourt.

Watnik, M. R. (1998). Pay for Play: Are Baseball Salaries Based on Performance. *Journal of Statistics Education [online], 6*(2).

Will, G. F. (1991). *Men at Work: The Craft of Baseball*. New York: Harper.