2011

Using Local Correlation to Explain Success in Baseball

Jeff Hamrick

University of San Francisco, jhamrick@usfca.edu

John Rasp

Follow this and additional works at: http://repository.usfca.edu/msan_fac

Part of the [Statistics and Probability Commons](#)

Recommended Citation

Hamrick, J., Rasp, J. Using local correlation to explain success in baseball. (2011) *Journal of Quantitative Analysis in Sports*, 7 (4), art. no. 5. DOI: 10.2202/1559-0410.1278

This Article is brought to you for free and open access by the College of Arts and Sciences at USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. It has been accepted for inclusion in Master of Science in Analytics (MSAN) Faculty Research by an authorized administrator of USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. For more information, please contact repository@usfca.edu.

Journal of Quantitative Analysis in Sports

Volume 7, Issue 4

2011

Article 5

Using Local Correlation to Explain Success in Baseball

Jeff Hamrick, *Rhodes College*
John Rasp, *Stetson University*

Recommended Citation:

Hamrick, Jeff and Rasp, John (2011) "Using Local Correlation to Explain Success in Baseball," *Journal of Quantitative Analysis in Sports*: Vol. 7: Iss. 4, Article 5.

©2011 American Statistical Association. All rights reserved.

Using Local Correlation to Explain Success in Baseball

Jeff Hamrick and John Rasp

Abstract

Statisticians have long employed linear regression models in a variety of circumstances, including the analysis of sports data, because of their flexibility, ease of interpretation, and computational tractability. However, advances in computing technology have made it possible to develop and employ more complicated, nonlinear, and nonparametric procedures. We propose a fully nonparametric nonlinear regression model that is associated to a local correlation function instead of the usual Pearson correlation coefficient. The proposed nonlinear regression model serves the same role as a traditional linear model, but generates deeper and more detailed information about the relationships between the variables being analyzed. We show how nonlinear regression and the local correlation function can be used to analyze sports data by presenting three examples from the game of baseball. In the first and second examples, we demonstrate use of nonlinear regression and the local correlation function as descriptive and inferential tools, respectively. In the third example, we show that nonlinear regression modeling can reveal that traditional linear models are, in fact, quite adequate. Finally, we provide a guide to software for implementing nonlinear regression. The purpose of this paper is to make nonlinear regression and local correlation analysis available as investigative tools for sports data enthusiasts.

KEYWORDS: nonparametric statistics, nonlinear regression, local correlation, peak performance, heteroscedasticity, sabermetrics, sports statistics

Author Notes: The first author would like to acknowledge valuable discussions with and feedback from Murad S. Taqqu regarding the techniques employed in this paper. The authors also thank an anonymous referee for helpful comments.

1 Introduction

Billy Beane made quite a splash in the public view when his use of statistical models for evaluating player talent, as general manager of the Oakland A's, was profiled in Michael Lewis' bestselling book *Moneyball* (Lewis, 2003). However, the use of statistical tools in the analysis of sports long predates Beane's much-publicized ventures.

In fact, statistical measures of player performance have existed in baseball virtually from the beginning of the game (Schwartz, 2004). They have greatly shaped public perception of player ability. The tenacity of traditional statistics has been challenged recently by the advent of modern quantitative approaches to the game. Bill James is often viewed as having inaugurated this "sabermetric revolution" with the publication of his *Baseball Abstract* (James, 1984). The popular press has welcomed this analytical trend: a quick search at Amazon.com for books on the general topic of "baseball statistics" yields over 900 titles. Other sports have followed in baseball's wake in this move toward quantification, and bookstores' shelves are filled with volumes on statistics related to a variety of sports, including football and basketball. For example, Kuper and Szymanski's *Soccernomics* claims to do for soccer what *Moneyball* did for baseball—namely, contradict many conventional wisdoms via statistical analysis (Kuper and Szymanski, 2009).

Academic researchers have been active in the quantitative analysis of baseball as well. Lindsey (1959, 1961, 1963) is the pioneer in this movement. Since that time the literature has burgeoned, expanding to consider a variety of aspects of the game. One key area of interest is descriptive measures of player performance. Baumer (2008), for example, compares the predictive ability of two standard measures of hitter proficiency: batting average and on base percentage. Anderson and Sharp (1997) develop a "composite batter index" of overall hitter ability. Koop (2002) employs economic models of multiple output production to compare player performance. Kaplan (2006) uses variance decomposition to assess player offensive productivity. Schell (1999, 2005) makes comparisons across historical eras to identify baseball's best hitters and sluggers. This interest in applying quantitative tools to historical issues is also seen in Bennett (1993), who investigates the "Black Sox" scandal of 1919.

The game of baseball lends itself well to mathematical modeling. Bennett and Flueck (1983) survey a variety of models used to evaluate offensive production in baseball. Since then the field has greatly expanded. Albright (1993), for example, examines hitting streaks in baseball. Rosner, Mosteller, and Youtz (1996) develop models for pitcher performance and the distribution of runs in an inning. Keller (1994) examines models, based upon the Poisson distribution, for the probability of winning a game, applying them to soccer as well as baseball. Yang and Swartz

(2004) use Bayesian methods for predicting the winner of a baseball game, while Koop (2004) uses Bayesian techniques for examining changes in the distribution of team performance over time. Simon and Simonoff (2006) model the “last licks” effect (the advantage the home team accrues by virtue of batting last, separate from home audience and field benefits). Barry and Hartigan (1993) use choice models to predict baseball divisional winners.

Business and economic aspects of the game have also been widely investigated, with salary issues having been a principal area of interest. Hadley and Gustafson (1991) and Hadley and Ruggiero (2006) consider the impact that free agency and arbitration have had on player compensation. Horowitz and Zappe (1998) examine salary structures of players at the end of their careers. The impact of salary growth on club finance is studied by Lacritz (1990). Hakes and Sauer (2006) explore the economic implications of the best-seller *Moneyball*, while Winfree (2009) takes a legal and economic approach to investigate quantitatively just what constitutes a baseball team’s “market.”

Within the sports data community, both the popular and the academic literatures have primarily approached statistical investigations using fairly basic tools. Descriptive techniques dominated the earlier efforts. Later approaches incorporated elementary inferential tools such as linear regression. However, analysis of sports-related questions has increasingly been done with sophisticated statistical methods. For example, Gibbons, Olkin, and Sobel (1978) use ranking and subset selection procedures to investigate the length of the baseball season and postseason. Investigation into membership in the Baseball Hall of Fame, which might be modeled as a standard logistic regression (Markus and Rasp, 1994), has been investigated using neural networks (Young, Holland, and Weckman, 2008), radial basis function networks (Smith and Downey, 2009), and random forests and simulated annealing (Freiman, 2010).

This paper is another effort to introduce advanced statistical techniques to the analysts of sports data. We propose an alternative to the linear parametric or nonlinear parametric regression models that are often used to analyze various issues in sports (Fair, 2008). In the presence of a single dependent variable and a single independent variable, the analysis of these models often reduces to a study of the Pearson correlation coefficient between the dependent and independent data. We propose instead to use a fully *nonlinear nonparametric regression model*. These models, which we generally call *nonlinear regression models* (despite the risk of confusion with parametric nonlinear regression models) admit an analogue of the usual Pearson correlation coefficient called the *local correlation function*. Nonlinear regression models are well-suited to a much richer class of joint distributions of independent and dependent variables and, as we will see, the local correlation function can facilitate a much more insightful examination of sports-related data.

While the techniques presented here are, of course, applicable to data in the context of any sport, we illustrate them by examining three research questions concerning baseball. We first consider the relationship between a pitcher's command of the strike zone and his ability to limit the number of runs allowed. Our second analysis examines the trajectory of a player's hitting ability throughout his career as he grows older. In the third analysis, we consider whether team success is more closely related to pitching, hitting, or defense.

Our analyses do create insight into these issues. However, the primary focus of the paper is to demonstrate how nonlinear regression models can be used in sports analysis. In particular, these models give rise to a local correlation function that is analogous to the usual Pearson correlation coefficient. We show how nonlinear regression models and the local correlation function can be used as descriptive tools and inferential tools. We also give an instance in which these tools are unnecessarily complications for analyses that can be done with simpler techniques.

In Section 2, we motivate the nonlinear regression model, define the local correlation function as an analogue of the Pearson correlation coefficient, and briefly explain how the estimation procedures for nonlinear regression are undertaken. In Section 3, we present three applications of nonlinear regression and the local correlation function. In Section 4, we offer concluding remarks and suggestions for future research. Finally, in Appendix A, we provide instructions on how to obtain and use software to fit nonlinear regression models.

2 A Nonlinear Regression Model

2.1 From Linear To Nonlinear Regression Models

While Pearson correlation is the most common statistical tool for measuring the strength of association between two random variables, it is often misunderstood and misused. Correlation is only one of many possible measures of association, including Kendall's τ and Spearman's ρ . See Joe (1997) and Drouet-Mari and Kotz (2001) for more information about these measures, and other measures, of association. In particular, Pearson correlation is a measure of *linear* association, i.e., two random variables X and Y with finite second moments are perfectly correlated if and only if $\mathbb{P}(Y = \alpha + \beta X) = 1$ for some $\alpha, \beta \in \mathbb{R}$. The Pearson correlation determines the joint distribution of X and Y if and only if the distribution of the random vector (X, Y) is jointly Gaussian. More generally, analogues of the Pearson correlation play a similar role for distributions whose constant density contours are ellipsoidal. See Bradley and Taqqu (2003) and Embrechts, McNeil, and Straumann (2002) for

more details. In cases in which the random vector (X, Y) is non-elliptical, correlation can be a misleading measure of dependence and using it as part of a simple linear regression analysis can be problematic. To address these problems, we suggest that quantitative sports analysts consider using *nonlinear* regression models and the resulting analogue of the usual Pearson correlation called *local correlation*.

The concept of local correlation was first proposed in papers by Bjerve and Doksum (1993) and Doksum, Blyth, Bradlow, Meng, and Zhao (1994). Their goal was to extend the connection between regression slopes, correlation, and the variance explained by regression models to analogous but nonlinear models. Unlike the Pearson correlation, local correlation is useful for data that are not jointly Gaussian. We now motivate the concept. Suppose that

$$(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho). \quad (1)$$

It is well-known that

$$Y|X = x \sim N\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right). \quad (2)$$

For the details of the supporting calculations see, for example, Ghahramani (2005, pg. 450-451). For data satisfying Equation (1), the usual linear regression model is

$$Y = \alpha + \beta X + \sigma \varepsilon \quad (3)$$

where $\varepsilon \sim N(0, 1)$ is independent of X and $\sigma > 0$. In this case, the local mean function is

$$m(x) := \mathbb{E}[Y|X = x] = \alpha + \beta x \quad (4)$$

with regression slope $m'(x) = \beta$. If we denote the covariance of X and Y by σ_{XY} , the regression slope $\beta = \sigma_{XY}/\sigma_X^2$, and $\sigma_{XY} = \rho\sigma_X\sigma_Y$. It therefore follows that the regression slope $\beta = \rho\sigma_Y/\sigma_X$ and that

$$\rho = \beta \frac{\sigma_X}{\sigma_Y}. \quad (5)$$

These results hold for any pair of random variables X and Y with finite second moments, though the true local mean function is only rarely—such as in the case of jointly Gaussian random variables—of the form in Equation (4).

Now recall from the theory of least-squares regression that σ_Y^2 , the variance of the dependent random variable Y , can be written as the sum of the variance explained by the regression (namely, $\beta^2\sigma_X^2$) and the residual (or unexplained) variance σ^2 . In other words,

$$\sigma_Y^2 = \beta^2 \sigma_X^2 + \sigma^2 \quad (6)$$

and thus

$$\rho = \frac{\sigma_X \beta}{\sqrt{\sigma_X^2 \beta^2 + \sigma^2}}. \quad (7)$$

Notice in Equation (7) that ρ is determined by the unconditional variance of the independent random variable σ_X^2 , the regression slope β , and the residual variance of the regression σ^2 . In particular, the correlation is an increasing function of β and a decreasing function of σ^2 .

Now suppose that Equation (3) does not govern the joint distribution of X and Y and that we replace the particular regression function $m(x) = \alpha + \beta x$ and the particular scedastic function $\sigma(x) \equiv \sigma$ by arbitrary functions $m : \mathbb{R} \rightarrow \mathbb{R}$ and $\sigma : \mathbb{R} \rightarrow \mathbb{R}^+$. Now, Equation (3) becomes

$$Y = m(X) + \sigma(X)\varepsilon, \quad (8)$$

where again ε is assumed to be a standard normal random variable that is independent of X . Moreover, notice that $Y|X = x$ is a typical linear regression relationship, even though Y (by itself) is not. We are then motivated by Equation (7) to define *local correlation* as

$$\rho = \frac{\sigma_X \beta(x)}{\sqrt{\sigma_X^2 \beta^2(x) + \sigma^2(x)}}, \quad (9)$$

where $\beta(x) := m'(x)$ is the slope of the nonparametric regression function $m(x) = \mathbb{E}[Y|X = x]$ and $\sigma^2(x) = \text{Var}(Y|X = x)$ is the nonparametric residual variance. Note that defining the function $\beta(x)$ requires us to assume that the covariate X is a continuous random variable; this assumption is used in Bjerve and Doksum (1993) and Doksum et al. (1994).

The viewpoint that must be taken if Equation (8) governs the relationship between X and Y is quite different from the viewpoint that must be taken if Equation (3) governs the relationship between X and Y . With the linear regression model, X affects Y through the intercept α , the slope β and the constant σ associated with the noise term ε . With the nonlinear regression model, X affects Y through the mean level $m(X)$ and the standard deviation $\sigma(X)$ associated with the noise term ε . These values may be different for different values of $X = x$. There are commonalities between the two models, however: for a given slope value $\beta(x)$, the larger $\sigma(x)$ is, the smaller $|\rho(x)|$ is. In the usual linear regression model, the Pearson correlation

$\rho > 0$ if and only if $\beta > 0$. Similarly, in a nonlinear regression model the sign of $\rho(x)$ at $X = x$ is determined by the negativity or positivity of $\beta(x)$ at $X = x$.

Notice that for the analysis of sports-related data, the standard linear regression model in Equation (3) is especially restrictive. There are very few joint distributions that admit a linear regression function $m(x) = \alpha + \beta x$ and a constant scedastic function $\sigma(x) = \sigma^2$. The Gaussian distribution is one such distribution. However, not even the multivariate Student's t distribution has these properties. Though its regression function is linear, its scedastic function is non-constant (Spanos, 1999). The powerful generality afforded by the nonlinear regression model in Equation (8) is much better suited to address a variety of statistical research questions in the quantitative analysis of sports data.

Finally, we note that local correlation satisfies many of the same properties satisfied by the usual Pearson correlation, including the following

1. $-1 \leq \rho(x) \leq 1$ for every x in the domain of ρ .
2. $\rho(x)$ is invariant with respect to location and scale changes in both X and Y .
3. In the case of a true linear regression model, $\rho(x)$ reduces to the usual Pearson correlation.
4. Equivalently, $\rho(x) \equiv \rho$ for all bivariate normal distributions.
5. If X and Y are independent, then $\rho(x) \equiv 0$, but the converse statement is not true.
6. If $Y = m(X)$, i.e., there is no noise in the model, then $\rho(x) = \pm 1$ for every x in the domain of the local correlation function. The sign of ρ at x is the same as the sign of $\beta(x) := m'(x)$. Notice that this property contrasts strongly with the analogous property for linear regression model: $\rho = \pm 1$ if and only if the model has no noise *and* the relationship between Y and X is perfectly linear.

For a more complete list of properties possessed by the local correlation function, see Bjerre and Doksum (1993).

Having motivated the local correlation function as the appropriate analogue of the usual Pearson correlation if one adopts a nonlinear regression model instead of a linear regression model, we now briefly discuss the problem of estimating m , σ , and ρ when Equation (8) is assumed to govern the relationship between a dependent variable Y and a covariate X .

2.2 Estimation of the Local Mean, Scedastic, and Correlation Functions

In order to estimate the local correlation function $\rho(x)$ at some particular point x_0 , we first assume that our N observations $\{(X_i, Y_i)\}_{i=1}^N$ are an independent sample

from a population (X, Y) and apply methods similar to those established by Bjerve and Doksum (1993) and Mathur (1998) and used previously by Bradley and Taqqu (2005) and Hamrick and Taqqu (2008b). The strategy is to estimate the values of $m(x_0)$, $\sigma(x_0)$, and $\rho(x_0)$ through consecutive local polynomial regressions. To estimate $\hat{\rho}(x_0)$, Bjerve and Doksum (1993) use a local linear regression to estimate $\beta(x_0)$ with a choice of bandwidth equal to the unconditional standard deviation of the covariate, denoted again by σ_X . Then, they perform a local linear regression on the squared residuals using a bandwidth selection technique based on σ_X and ultimately obtain an estimate of the scedastic function, which we denote by $\hat{\sigma}(x_0)$.

Instead, we use a suggestion of Mathur (1998). First, we apply a local quadratic regression to estimate $\beta(x_0)$ using a separate estimate of the asymptotically optimal bandwidth for that particular regression. This choice of bandwidth is designed to reduce the bias in our estimator $\hat{\beta}(x_0)$. Second, we apply a local linear regression on the squared residuals to estimate $\sigma(x_0)$. We use yet another technique to estimate the asymptotically optimal bandwidth (Ruppert, Wand, Holst, and Hössjer, 1997).

To estimate m at a target point x_0 , we assume that m is analytic about x_0 and therefore admits a Taylor expansion, i.e.,

$$m(x) \approx m(x_0) + m^{(1)}(x_0)(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \dots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \quad (10)$$

This polynomial estimate of the regression function is fit locally at x_0 using weighted least squares regression. That is, the terms $m^{(k)}(x_0)/k!$, $k = 0, \dots, p$, are estimated as the coefficients of the weighted least squares problem

$$\min_{(\beta_0(x_0), \dots, \beta_p(x_0)) \in \mathbb{R}^p} \sum_{i=1}^n \left\{ Y_i - \sum_{k=0}^p \beta_k(x_0)(X_i - x_0)^k \right\}^2 w_i(x_0, h), \quad (11)$$

where the weights of the regression at x_0 are given by a kernel function

$$w_i(x_0, h) := K_h(X_i - x_0) := \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right). \quad (12)$$

Solving this least-squares problem yields the estimators

$$\hat{m}^{(k)}(x_0) = k! \hat{\beta}_k(x_0). \quad (13)$$

As noted above, the bandwidth h in the regression weights $w_i(x_0, h)$ in Equation (12) can be chosen in an asymptotically optimal way. By taking the derivative of

Figure 1: A plot of the Epanechnikov kernel $K(u) := \frac{3}{4} \max\{0, 1 - u^2\}$.

the Taylor expansion in Equation (10), we obtain the value of $\beta(x_0) = m'(x_0)$. In practice, these values of the local mean function $m(x)$ and the local slope function $\beta(x)$ are obtained along a relatively fine mesh of equally spaced points along the x -axis. For purposes of visualization and prediction in between the mesh points, one can use the built-in features of MATLAB to fit a spline to the pairs $(x_0, \hat{m}(x_0))$ while additionally using the information affected by the estimates $\hat{m}'(x_0)$.

For additional details about the optimal selection of the bandwidth h or the estimation of the scedastic function σ , see the explanation in Bradley and Taqqu (2005). For more information about local polynomial regression, see Fan and Gijbels (1996).

Consistent with Bjerve and Doksum (1993), we choose the Epanechnikov kernel

$$K(u) := \frac{3}{4} \max\{0, 1 - u^2\}. \quad (14)$$

The kernel is plotted in Figure 1. This choice of kernel is quite typical in local polynomial modeling. Remarkably, for local polynomial estimators it can be shown that the Epanechnikov kernel—rather than a Gaussian kernel—is optimal in the sense that regardless of the degree of the local polynomial fit, the Epanechnikov kernel minimizes the asymptotic mean squared error. For more information about the Epanechnikov kernel and other choices of kernel functions, see Fan and Yao (2005).

Additionally, the monograph by Fan and Gijbels (1996) has detailed information about local polynomial regression. There is additional information about optimal bandwidth selection for a variety of nonparametric models in Fan and Yao (2005) and, in the special case of estimating $\sigma(x_0)$, in Ruppert et al. (1997). From the perspective of actually implementing estimation procedures for $\sigma(x)$, this distinction is important because Ruppert et al. (1997) showed that the asymptotically optimal bandwidth for the estimate of $\sigma(x_0)$ is different than the asymptotically optimal bandwidth for the corresponding estimate of $m(x_0)$.

2.3 Asymptotic Distribution of Estimated Functions

Though more complicated than their linear analogues, the constants α and β , the asymptotic distributions of the estimated local mean, slope, scedastic, and correlation functions are known and provide a foundation for confidence interval construction and hypothesis testing. For example, Fan and Gijbels (1996) show and then Hamrick and Taqqu (2008a) use the following theorem.

Theorem 2.1. *Suppose that $\hat{\beta}$ is the estimator first described in Equation (13). Additionally, suppose that the following regularity conditions hold: the density of the continuous random variable X , $f_X(x)$, the third derivative of the true local mean function, $m^{(3)}(x)$, and the first derivative of the true local scedastic function $(d/dx)(\sigma^2(x))$ are continuous; the residual variance $\sigma^2(x)$ is positive and finite; and the conditional fourth moment of Y given $X = x$ is finite, i.e., $\mathbb{E}[Y^4|X = x] < \infty$. Then for any x_0 such that $f_X(x_0) > 0$, we have*

$$\left(\frac{7f_X(x_0)nh^3}{15\sigma^2(x_0)}\right)^{1/2} [\hat{\beta}(x_0) - \beta(x_0) - (h^2m^{(3)}(x_0)/14) + o(h^2)] \longrightarrow N(0, 1) \quad (15)$$

as $n \longrightarrow \infty$, $h \longrightarrow 0$, and $nh \longrightarrow \infty$.

In particular, Fan and Gijbels (1996) show that if the bandwidth h tends to zero at the rate of $n^{-1/7}$, i.e., $h = o(n^{-1/7})$, then

$$\left(\frac{7f_X(x_0)nh^3}{15\sigma^2(x_0)}\right)^{1/2} [\hat{\beta}(x_0) - \beta(x_0)] \longrightarrow N(0, 1) \quad (16)$$

and, as a result, $\hat{\beta}(x_0)$ is asymptotically unbiased. Asymptotic unbiasedness is, of course, not necessarily desirable since it is generally traded off against asymptotic variance. The MATLAB software used for our analysis and that we describe in Appendix A chooses a bandwidth that optimizes the bias-variance tradeoff as determined by Fan and Gijbels (1996) and Ruppert et al. (1997).

Recall that the local correlation estimator of the form

$$\hat{\rho}(x_0) := \frac{s_X \hat{\beta}(x_0)}{\sqrt{s_X^2 \hat{\beta}^2(x_0) + \hat{\sigma}^2(x_0)}}, \quad (17)$$

where $s_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$ is the sample standard deviation of the covariate data. The following result holds.

Theorem 2.2. *Suppose that (1) x_0 is an interior point of the support of $f_X(x)$, the density of the covariate X ; (2) $m(x)$ has four continuous derivatives in some neighborhood of x_0 ; (3) $f_X(x)$ and $\sigma^4(x)$ are differentiable in a neighborhood of x_0 ; (4) the estimation of the local mean function $\hat{m}(x)$ is locally quadratic with bandwidth $h_1 = O(n^{-1/7})$; (5) the estimation of the local scedastic function $\hat{\sigma}(x)$ is locally linear with bandwidth $h_2 = O(n^{-1/5})$. Then for the estimator defined in Equation (17),*

$$\left(\frac{7f(x_0)nh_1^3}{15\sigma^2(x_0)} \right)^{1/2} [1 - \rho^2(x_0)]^{-3/2} [\hat{\rho}(x_0) - \rho(x_0)] \longrightarrow N(0, 1) \quad (18)$$

as $h_1 \longrightarrow 0$, $h_2 \longrightarrow 0$, $nh_1 \longrightarrow \infty$, and $nh_2 \longrightarrow \infty$.

Our aim is neither to state nor prove all results related to nonlinear regression theory; for a survey of the field, see Fan and Yao (2005). This result facilitates the construction of confidence intervals around estimates of the local correlation function $\hat{\rho}$ (see, for example, Figure 3) and the execution of hypothesis tests, as outlined in Section 3.2.

3 Sabermetric Applications

In this section we apply the local correlation techniques developed in Section 2 to three particular sabermetric research questions. These examples do shed light upon the game of baseball. However, the primary purpose of these examples is to illustrate the advantages as well as the limitations of nonlinear regression and the local correlation function. Accordingly, we will first examine a situation in which local correlation methods are useful as a descriptive statistical tool. Second, we discuss an example in which we employ the local correlation for inferential purposes. Finally, we give a situation in which standard linear correlation models are adequate, and local correlation techniques are an unnecessary complication.

All data used in these examples were extracted from Sean Lahman's Baseball Archive, an online database of baseball statistics. This archive contains a wealth of sabermetric data dating back to the beginning of professional baseball,

Figure 2: A scatterplot of pitcher *Command* versus *ERA* for 2008 Major League Baseball.

and is located at <http://www.baseball11.com/>. The subsets of the data used in each analysis are described in the corresponding section below. They are also available from the authors, at

<http://www2.stetson.edu/~jrasp/research/localcorrelation.htm>

3.1 Local Correlation as a Descriptive Tool

To what extent does a pitcher's command of the strike zone influence his ERA?

Ultimately, a pitcher's job is to minimize the number of runs scored by the opposing team. The Earned Run Average (denoted *ERA*) is one traditional measure of his success in doing so. It is computed as the number of earned runs allowed, per nine innings pitched.

The pitcher has a variety of tools at his disposal in his efforts to muffle hitter effectiveness—pitch speed, location, technique, etc. One key variable is simply the pitcher's ability to throw strikes. Sabermetric analysis has focused on *Command* as a measure of strike zone control. *Command* is the ratio of strikeouts to walks allowed (Shandler, 2010, pp. 20, 278). What can we say about the relationship between *Command* and *ERA*?

For this analysis, we computed the *ERA* and *Command* for all major league pitchers during the 2008 Major League Baseball season. Pitchers with fewer than 100 innings pitched were excluded from the analysis. The resulting sample size is

Figure 3: The estimated local correlation function $\hat{\rho}$, local mean function $\hat{m}$, local slope function $\hat{m}'$, and local scedastic function $\hat{\sigma}$ when nonlinearly regressing *ERA* against *Command*.

$n = 142$ observations. A typical scatterplot, as shown in Figure 2, illustrates the relationship between *Command* and *ERA*.

The Pearson correlation coefficient here is -0.578 , indicating a moderately strong inverse relationship between the two variables. However, this conclusion is misleading. The graph suggests that the true relationship is nonlinear—the relationship is more strongly negative for smaller values of *Command*, with a weaker (and possibly negligible) relationship for higher values of *Command*. Moreover, there may be some heteroscedasticity in the data. The variance appears smaller for higher values of *Command*, though it is difficult to say for certain, given the small number of data. Figure 3 illustrates the results obtained by analyzing these same data using the tools discussed in Section 2 using the software described in Appendix A.

The estimated local correlation function $\hat{\rho}(x)$ (in the upper left of Figure 3) appears approximately linear. While confidence bands become quite wide as *Command* increases, it is nevertheless apparent that the relationship between *Command* and *ERA* is strongly negative for low values of *Command*, and becomes considerably weaker for higher values of *Command*.

The estimated local mean function $\hat{m}$ (in the upper right of Figure 3) indicates that the relationship between *Command* and *ERA* is nonlinear, perhaps an exponential function with a negative growth rate. The estimated local slope function $\hat{m}'$ (in the lower left of Figure 3) reinforces this notion. It shows that the local mean is rapidly declining for lower values of *Command*, but that the rate of decline slows and essentially stops for higher values of *Command*. The estimated local scedastic function $\hat{\sigma}$ (in the lower right of Figure 3) indicates a much higher variability in *ERA* for lower values of *Command* than for higher values.

A traditional parametric analysis of these data might examine their scatterplot (see Figure 3), and then proceed to fit a function such as $Y = \alpha + \beta/X$ or $Y = \alpha + \exp\{-\beta X\}$. There are advantages to this approach: it is computationally simpler, and it is fairly easy to make global statements about the relationship between independent and dependent variables. For example, using the model $Y = \alpha + \beta/X$, one could say that as *Command* becomes larger, *ERA* tends to α .

But there are disadvantages to this approach as well. One problem is its essentially *ad hoc* nature—there is no *a priori* reason (in this example) to believe that any particular parametric model is appropriate. Furthermore, as has already been noted, standard parametric assumptions such as homoscedasticity may well be violated. The nonlinear nonparametric analysis, while more involved computationally and requiring choices of kernel and bandwidth, does give a more detailed picture of the relationship between *Command* and *ERA*. It goes beyond the basic conclusion of a moderately strong inverse correlation between these variables. It further suggests, for example, that for the best pitchers (or, at least, for those pitchers with the highest levels of *Command*), there is little to be gained by further enhancing this skill. Diminishing returns are at work. Rather, it is the less accomplished pitchers (those pitchers with lower levels of *Command*) whose *ERA* is most improved by modest improvements in *Command*.

3.2 Local Correlation as an Inferential Tool

One issue that has become of interest lately to sabermetricians (and fantasy baseball leagues) is the trajectory of a player's career. When might a player be said to "peak" in ability? How long does this peak last, and how abrupt will the decline be? Fair (2008) employs a parametric nonlinear regression approach to investigate these matters. We use nonparametric nonlinear regression techniques and the local correlation function to investigate this issue for hitters in major league baseball.

Sabermetricians have come to recognize *OPS* (on base percentage plus slugging percentage) as one of the best overall measures of a hitter's productivity (Shandler, 2010, pp. 277). This statistic combines the player's ability to get on base with

Figure 4: A scatterplot of Age versus *OPS* for selected major league baseball players for the period from 1872 to 2009.

his ability to hit for power. Since we are interested in career trajectory of players, we restrict this analysis to those major league baseball players with at least 5000 times at bat. We compute the *OPS* for each of these players for each season in which he competed. To avoid outliers or high-variance observations caused by an artificially shortened season (such as a season in which the player was injured), we restrict the analysis only to those seasons in which the player had at least 200 times at bat. The independent variable in the analysis is the player's age (years and portions of a year), as of April 1 of the baseball season in question. Overall we have $n = 9603$ data points from 706 different players.

The scatterplot in Figure 4 illustrates the relationship between *Age* and *OPS*. The graph suggests little, if any, structure in the data. There is possibly a peak in *OPS* around the age of 26 or 27, which is consistent with the common wisdom that hitters tend to peak in ability at around this age (Shandler, 2010, pp. 19). A few outliers are conspicuous, and are possibly associated with various controversies over use of performance-enhancing drugs. But on the whole, there is little evidence of a relationship between *Age* and *OPS*. This fact can further be seen by the small Pearson correlation of -0.024 .

Estimation of the local correlation function reveals substantially more information about the data. The local mean function highlights the fact that *OPS* does indeed seem to peak around age 28 (somewhat higher than previously suggested in the literature), although the effect is clearly small in magnitude. Moreover, we note that the local correlation between *Age* and *OPS* is fairly positive and strong (approximately 0.5) for younger players (close to the age of 20). As players age,

Figure 5: The estimated local correlation function $\hat{\rho}$, local mean function $\hat{m}$, local slope function $\hat{m}'$, and local scedastic function $\hat{\sigma}$ when nonlinearly regressing *OPS* against *Age*.

the relationship between *Age* and *OPS* weakens considerably; the two variables are essentially uncorrelated past the age of 28 or so. There is a slight, and barely significant, negative local correlation between *Age* and *OPS* for older players, but it is clearly not particularly strong.

One problem with this analysis is that it in some sense treats all players as equals. Players of widely differing abilities (e.g., Babe Ruth and Mario Mendoza) are graphed together, and thus some of the “signal” reflected in the trajectory of ability is lost in the “noise” of the confounding variable of player ability.

One way to remedy this complication is to standardize the data by subtracting each player’s career average *OPS* from his *OPS* value for each particular year. This standardization centers each player’s data around zero. *Standardized OPS* data thus give information about the trajectory of a player’s career above or below his career average. A scatterplot of these *Standardized OPS* values by age is given in Figure 6.

This standardization reveals a bit more structure in the relationship between hitting performance and age. The Pearson correlation is still slightly negative (-0.081). However, the nonlinearity of the data, and hence the limited usefulness of the standard correlation measure, are readily seen. Traditional parametric

Figure 6: A scatterplot of Age versus *Standardized OPS* for selected major league baseball players for the period from 1872 to 2009.

methods fall far short here. No parametric model is suggested by eye-balling a scatterplot of the data, unlike perhaps the example in Section 3.2. Examination of the local correlation function, and related functions, shown in Figure 7 tells a more complex story.

The differences between this analysis and the previous analysis are small but pronounced. The relationship between *Standardized OPS* and Age is somewhat stronger than before, particularly for younger players. For older players, the relationship is more pronounced in the negative direction. From this perspective as well, players seem to peak in ability around the age of 28.

The pointwise asymptotic properties of our estimator of the local correlation function are known and were discussed in Section 2.3. Hence standard inferential techniques, including confidence interval construction and hypothesis testing, are readily executable. One key distinction between nonlinear regression analysis and the usual linear regression analysis must be kept in mind, however. With the traditional linear model, it is meaningful to talk about *the* slope (or *the* correlation), since a single value is valid for the entire range of the covariate. With local correlation analysis, however, one speaks of the local correlation *at a particular value of the covariate*.

Since we expect player ability to peak around the age of 28, we would expect the *Standardized OPS* to level off at this value, and hence the local correlation to be zero when the variable Age is 28 (mostly due to the slope the local mean function being approximately equal to zero). We therefore want to test the null hypothesis $H_0 : \rho(28) = 0$ against the two-sided alternative.

Figure 7: The estimated local correlation function $\hat{\rho}$, local mean function $\hat{m}$, local slope function $\hat{m}'$, and local scedastic function $\hat{\sigma}$ when nonlinearly regressing *Standardized OPS* against *Age*.

The estimator in Equation (17) of the local correlation function is point-wise asymptotically Gaussian, as stated in Theorem 2.2. To further establish the notion that, for example, the estimate $\hat{\rho}(28)$ is approximately Gaussian even when asymptotic conditions do not formally hold, we generate quantile-quantile (QQ) and probability-probability (PP) plots for bootstrapped distributions of (the normalized value of) $\hat{\rho}(28)$ against a standard normal distribution. The QQ and PP plots adhere closely to the identity functions that are represented by dashed lines in those plots. Hence, the bootstrapped values of $\hat{\rho}(28)$ are plausibly normal. For further information about QQ and PP plots, see the papers by Gnandesikan and Wilk (1968) and Michael (1983), respectively, or a modern econometrics textbook like Thode (2002).

Now that we are particularly confident that the test statistic for determining if there is evidence in favor of $H_1: \rho(28) \neq 0$ is approximately standard normal, we proceed. For this hypothesis test, the test statistic is equal to approximately 1.65, which is less than the critical value of 1.96 necessary to reject H_0 at the 5% level of significance. We do not reject the null hypothesis of zero local correlation between *Age* and *Standardized OPS* for baseball players who are 28 years old. In fact, a more careful inspection of the graph of the local correlation function suggests

Figure 8: Quantile-quantile (QQ) and probability-probability (PP) plots for the bootstrapped distribution of (normalized values of) $\hat{\rho}(28)$ against a standard normal distribution to support the hypothesis testing in Section 3.2.

that $\hat{\rho}(28.25)$ is approximately equal to zero. In fact, a test of the null hypothesis $H_0: \rho(28.25) = 0$ against the two-tailed alternative is not rejected—the test statistic is -0.0279 . In other words, there is fairly strong evidence that baseball players peak in terms of offensive productivity around age 28.25.

Two issues should be noted about these hypothesis tests. First, consider a test of the null hypothesis that the local slope is equal to zero for some value x_0 (i.e., $H_0: \beta(x_0) = 0$). The associated test statistic will *not*, in general, be the same as the test statistic for a test that local correlation is zero at x_0 . In this regard, statistical inference differs from the linear case.

Second, this analysis suggests that players tend to peak in ability around the age of 28.25, somewhat later than the age of 26 that is often reported in the literature. If we consider baseball players approximately one year older and younger than, respectively, 28.25, the null hypotheses $H_0: \rho(27.25) = 0$ and $H_0: \rho(29.25) = 0$ are strongly rejected with a test statistics of approximately 12.9585 and -6.3978 , respectively.

We have illustrated a single inferential procedure. Additional inferential techniques are also available, for example, on β using the asymptotics in Theorem 2.1. Additionally, it is possible to test $H_0: \rho(x_1) = \rho(x_2)$ for $x_1 \neq x_2$; see, for example, Hamrick and Taqqu (2008b) or Hamrick and Taqqu (2008a). For more information, see Fan and Gijbels (1996) and Mathur (1998).

3.3 Local Correlation as an Unnecessary Complication

There's an old adage that championships in baseball are won on "pitching and defense," rather than hitting. Is this claim really true? How well does team success

Figure 9: A scatterplot of team *Earned Run Average* versus team *Winning Percentage* for major league baseball teams over the period from 1921 to 2009.

correlate with measures of team pitching, defense, and hitting? What insights, if any, can the local correlation function add to standard linear parametric methods?

For this analysis, we use data for every team in major league baseball, for every season from 1921 (when full abandonment of “dead ball era” practices occurred) through 2009. For each team, we computed the proportion of games won for the season (*Win Percentage*). We also compiled statistics to capture team pitching, fielding, and hitting ability during the season. We used team *Earned Run Average* (*ERA*) as our measure of pitching success. For fielding, we use the (admittedly imperfect) measure of the team’s *Fielding Percentage* (*FP*). The hitting measure was *OPS*, on base plus slugging percentage. It is generally accepted as the best single statistic for capturing overall hitting performance. In each of these three analyses, the sample size is $n = 1906$ observations.

The scatterplot in Figure 9 above illustrates the relationship between team *Win Percentage* and our pitching measure, team *ERA*. The plot indicates that the relationship between the two variables is plausibly linear. The standard Pearson correlation coefficient here is -0.532 , indicating a moderately strong inverse relationship between the two variables. Unsurprisingly, the more earned runs a team allows, the fewer games they tend to win.

Computing the local correlation function, as well as the affiliated local mean, local slope, and local standard deviation functions, yields results as illustrated in Figure 10. Note in this case that the local correlation function is essentially flat; the correlation is basically constant across the entire range of the *ERA* data. The

Figure 10: The estimated local correlation function $\hat{\rho}$, local mean function $\hat{m}$, local slope function $\hat{m}'$, and local scedastic function $\hat{\sigma}$ when nonlinearly regressing team *Winning Percentage* against team *Earned Run Average*.

indicated local correlation for all values of *ERA* of approximately -0.50 is in rough accord with the computed Pearson correlation of -0.532. The local mean function appears almost exactly linear, indicating that a linear model is appropriate for these data. While the local slope and local scedastic functions do show that the relationship between team *Win Percentage* and team *ERA* has some minor deviations from strict linearity and homoscedasticity, quick perusal of the vertical scales of these graphs shows that such departures from the standard assumptions are negligible. The standard linear model appears to be adequate to examine these data; use of local correlation procedures is an unnecessary complication in this situation.

Two similar computations, using the fielding measure *Fielding Percentage* and the hitting measure *OPS*, respectively, as predictors of *Win Percentage* reveal essentially the same pattern. Namely, the local correlation function is flat, the local mean function is approximately linear, and the local slope and scedastic functions are small in magnitude. In the interests of brevity we do not reproduce these graphs here; they are similar to the graphs in Figure 10. Standard linear techniques are appropriate to answer the research question.

Using the traditional linear approach, we can compute the three (constant) Pearson correlation coefficients between the predictors and *Win Percentage*. The results are given in the table below:

Covariate	Dependent Variable	Pearson Correlation
<i>ERA</i> (pitching)	<i>Win Percentage</i>	-0.532
<i>FP</i> (fielding)	<i>Win Percentage</i>	0.263
<i>OPS</i> (hitting)	<i>Win Percentage</i>	0.499

Note from these three sample correlations that, taken alone, pitching performance (as measured by *ERA*) is the best single predictor of a team's success at winning ball games. The sample correlation between *Fielding Percentage* and *Win Percentage* is much lower, although this may reflect the fact that *Fielding Percentage* is a poor measure of a team's defensive ability. But it may reflect the reality that it is pitching, and not defense, that wins championships.

4 Conclusions

In this paper, we have introduced and illustrated the use of a fully nonparametric nonlinear regression model and its associated local correlation function. In addition to describing a method for estimating the functions that govern the nonlinear regression model, we stated a number of asymptotic results that facilitate the construction of confidence intervals and hypothesis testing.

We used this regression model and the local correlation function to analyze three situations in sabermetrics. First, we have characterized the relationship between *Command* and *ERA* as unlikely to be linear in nature. In particular, our nonparametric analysis suggests that the local correlation between *Command* and *ERA* is relatively large and negative for relatively small values of *Command* and declines to zero as higher percentiles of *Command* are considered. We used nonlinear regression to determine that the age of peak performance in baseball, as measured by standardized *OPS* (on-base percentage plus slugging percentage), is more likely to be 28.25 than 26 or 27. Finally, we used nonlinear regression to verify that various models of success in baseball—say, using team earned run average, team fielding percentage, or team *OPS* to explain team winning percentage—are perfectly well-managed by typical linear regression models. In this sense, the software described in Appendix A can be quickly used to check if a linear regression model is plausible for some set of candidate independent and dependent data.

While we have illustrated use of nonlinear regression models (and their attendant local correlation functions) in the context of several sabermetric examples,

future research efforts might focus on the use of nonlinear regression to explain phenomena in other sporting venues, such as tennis, football, and soccer, or economic matters related to a particular sporting enterprise.

A A Guide to the Software

The software used to execute the nonlinear regression in this paper was written in MATLAB. We briefly outline its use in this section of the paper. To obtain the software, contact the first author by emailing him at hamrickj@rhodes.edu.

The MATLAB functions necessary to generate $\hat{m}$, $\hat{\sigma}$, $\hat{\beta}$ and $\hat{\rho}$ are gathered in a directory called `NonlinearRegressionDir`. The user's data should be placed in either a MATLAB data file called, for example, `Sports.mat`, or in a CSV file called, for example `Sports.csv`. We will assume that the MATLAB data file contains a two-column MATLAB array called `Sports`. The first column of the array contains the covariate X (e.g., *Command*) and the second column contains the dependent variable Y (e.g., *ERA*). Each row then corresponds to one joint observation of X and Y . A description of how to deploy the software follows.

First, invoke MATLAB. Add the directory `NonlinearRegressionDir` directory to MATLAB's working path. For purposes of specificity, we will assume that the directory is located in the `MyHomePath` subdirectory:

```
addpath('C:\MyHomePath\NonlinearRegressionDir');
```

Then load the data—we assume the use of a MATLAB data file—into the MATLAB workspace:

```
load('C:\MyHomePath\Sports.mat');
```

The MATLAB workspace now contains the array `Sports`. Next, define the MATLAB vectors `X` and `Y` appropriately:

```
X = Sports(:,1); Y = Sports(:,2);
```

Next, let us define a set of target points for which estimates of the values of the local correlation function $\hat{\rho}$, the local mean function $\hat{m}$, the local slope function $\hat{\beta}$, and the local scedastic function $\hat{\sigma}$ are desired. Suppose that you want 101 equally-spaced estimates of the aforementioned functions from the first to the ninety-ninth percentile of the X data; that is, from $x_{\min} = \hat{F}_X^{-1}(0.01)$ to $x_{\max} = \hat{F}_X^{-1}(0.99)$, where $\hat{F}_X$ is the empirical cumulative distribution function of the data

in the column associated with the covariate X . To work with this set of target points, enter

```
num_targets = 101;
x_min = prctile(X, 1.0);
x_max = prctile(X, 99.0);
x_0 = linspace(x_min, x_max, num_targets);
```

The MATLAB variable x_0 is now a column vector with the target points. The following command estimates the local correlation at the targets points x_0 and plots the estimate of the local correlation function and the other functions (the local mean, local slope, and local scedastic functions) that define the local correlation function (see, for example, Figure 3):

```
plot_flag = 1;
[Rho, Beta, Sigma, StdRho] = CorrCurve(Y, X, x_0, plot_flag);
```

The function `CorrCurve` returns the following data.

- **Rho**, an array (of length `num_targets`) of local correlation estimates;
- **Beta**, an array (of dimension `num_targets` by 3) of local regression coefficients. The first column corresponds to the local mean $\hat{m}$ along the target points, the second column corresponds to the local slope estimates $\hat{\beta}$ along the target points, and the third column corresponds to $1/2!$ times the estimate of the second derivative of the regression function $\hat{m}^{(2)}$ along the target points.
- **Sigma**, an array (of length `num_targets`) of local residual standard deviation estimates; and
- **StdRho**, an array (of length `num_targets`) of local standard deviations of the estimator $\hat{\rho}(x)$, to be used in establishing confidence intervals around $\hat{\rho}(x)$.

To examine the data, type **Rho**, **Beta**, **Sigma**, or **StdRho** at the MATLAB command prompt. To access particular values of these functions in MATLAB, simply evaluate the function at the position in the array corresponding to value of interest. For example, if 3.00 is the 38th member of the array **Rho**, then executing `Rho(38)` in the command prompt will display the value of the estimated local correlation function at 3.00. It may be necessary to manipulate the code controlling x_0 to guarantee that the functions $\hat{\rho}$, $\hat{\beta}$, etc., can be evaluated at that point. Another option is to fit a spline to, for example, the pairs $\{x_0, \hat{\rho}(x_0)\}$, either using

MATLAB or some other computer algebra system. To save the estimation results to a MATLAB data file called, for example, `Results.mat`, type

```
save 'C:\MyHomePath\Results' Rho Beta Gamma StdRho;
```

To verify that the QQ and PP plots of the distributions of bootstrapped values of $\hat{\rho}$ at some particular point are, in fact, approximately normal even under when the number of data are finite, enter the following into the MATLAB prompt:

```
num_boot = 1000;  
BootstrapLocalCorr(Y, X, num_boot);
```

The QQ and PP plots will be displayed automatically. They should look similar to the examples in Figure 8 of Section 3.2.

References

- Albright, S. (1993): "A statistical analysis of hitting streaks in baseball," *Journal of the American Statistical Association*, 88, 1175–1183.
- Anderson, T. and G. Sharp (1997): "A new measure of baseball batters using DEA," *Annals of Operations Research*, 73, 141–155.
- Barry, D. and J. Hartigan (1993): "Choice models for predicting divisional winners in major league baseball," *Journal of the American Statistical Association*, 88, 766–774.
- Baumer, B. (2008): "Why on-base percentage is a better indicator of future performance than batting average: An algebraic proof," *Journal of Quantitative Analysis in Sports*, 4:2.
- Bennett, J. (1993): "Did Shoeless Joe Jackson throw the 1919 World Series?" *The American Statistician*, 47, 241–250.
- Bennett, J. and J. Flueck (1983): "An evaluation of major league baseball offensive performance models," *The American Statistician*, 37, 76–82.
- Bjerve, S. and K. Doksum (1993): "Correlation curves: Measures of association as functions of covariate values," *Annals of Statistics*, 21:2, 890–902.
- Bradley, B. and M. Taquq (2003): "Financial risk and heavy tails," in *Handbook of Heavy-Tailed Distributions in Finance*, Elsevier-Science.
- Bradley, B. and M. Taquq (2005): "How to estimate spatial contagion between financial markets," *Finance Letters*, 3:1, 64–76.

- Doksum, K., S. Blyth, E. Bradlow, X. Meng, and H. Zhao (1994): "Correlation curves as local measures of variance explained by regression," *Journal of the American Statistical Association*, 89:426, 571–582.
- Drouet-Mari, D. and S. Kotz (2001): *Correlation and Dependence*, Imperial College Press.
- Embrechts, P., A. McNeil, and D. Straumann (2002): "Correlation and dependence in risk management: Properties and pitfalls," in M. Dempster and H. Moffatt, eds., *Risk Management: Value at Risk and Beyond*, Cambridge University Press, 176–223.
- Fair, R. (2008): "Estimated age effects in baseball," *Journal of Quantitative Analysis in Sports*, 4:1.
- Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.
- Fan, J. and Q. Yao (2005): *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer Series in Statistics, New York: Springer.
- Freiman, M. (2010): "Using random forests and simulating annealing to predict probabilities of election to the baseball hall of fame," *Journal of Quantitative Analysis in Sports*, 6:2.
- Ghahramani, S. (2005): *Fundamentals of Probability with Stochastic Processes*, Pearson Prentice Hall, 3rd. edition.
- Gibbons, J., I. Olkin, and M. Sobel (1978): "Baseball competitions—Are enough games played?" *The American Statistician*, 32:3, 89–95.
- Gnandesikan, R. and M. Wilk (1968): "Probability plotting methods for the analysis of data," *Biometrika*, 55:1, 1–17.
- Hadley, L. and E. Gustafson (1991): "Major league baseball salaries: The impacts of arbitration and free agency," *Journal of Sport Management*, 6, 111–127.
- Hadley, L. and J. Ruggiero (2006): "Final-offer arbitration in major league baseball: A nonparametric approach," *Annals of Operations Research*, 145, 201–209.
- Hakes, J. and R. Sauer (2006): "An economic evaluation of the *Moneyball* hypothesis," *Journal of Economic Perspectives*, 20, 173–185.
- Hamrick, J. and M. Taqqu (2008a): "Contagion and confusion in credit default swap markets," Working paper, Boston University.
- Hamrick, J. and M. Taqqu (2008b): "Is there contagion or confusion in bond markets? Evidence from local correlation," Working paper, Boston University.
- Horowitz, I. and C. Zappe (1998): "Thanks for the memories: Baseball veterans' end-of-career salaries," *Managerial and Decision Economics*, 19, 377–382.
- James, B. (1984): *Bill James Baseball Abstract*, Ballantine Books.
- Joe, H. (1997): *Multivariate Models and Dependence Concepts*, Chapman & Hall.
- Kaplan, D. (2006): "A variance decomposition of individual offensive baseball performance," *Journal of Quantitative Analysis in Sports*, 2:3.

- Keller, J. (1994): "A characterization of the Poisson distribution and the probability of winning a game," *The American Statistician*, 48, 294–298.
- Koop, G. (2002): "Comparing the performance of baseball players: A multiple-output approach," *Journal of the American Statistical Association*, 97, 710–720.
- Koop, G. (2004): "Modelling the evolution of distributions: An application to major league baseball," *Journal of the Royal Statistical Society, Part A*, 167, 639–655.
- Kuper, S. and S. Szymanski (2009): *Soccernomics: Why England Loses, Why Germany And Brazil Win, And Why The U.S., Japan, Australia, Turkey-And Even Iraq-Are Destined To Become The Kings Of The World's Most Popular Sport*, Nation Books.
- Lacritz, J. (1990): "Salary evaluation for professional baseball players," *The American Statistician*, 44, 4–8.
- Lewis, M. (2003): *Moneyball*, W.W. Norton & Co.
- Lindsey, G. (1959): "Statistical data useful for the operation of a baseball team," *Operations Research*, 7, 197–207.
- Lindsey, G. (1961): "The progress of the score during a baseball game," *Journal of the American Statistical Association*, 56, 703–728.
- Lindsey, G. (1963): "An investigation of strategies in baseball," *Operations Research*, 11, 477–501.
- Markus, A. and J. Rasp (1994): "Logistic regression models for the baseball Hall of Fame," in *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 46–49.
- Mathur, A. (1998): *Partial Correlation Curves*, Ph.D. dissertation, University of California, Berkeley.
- Michael, J. (1983): "The stabilized probability plot," *Biometrika*, 70:1, 11–17.
- Rosner, B., F. Mosteller, and C. Youtz (1996): "Modeling pitcher performance and the distribution of runs per inning in major league baseball," *The American Statistician*, 50, 352–360.
- Ruppert, D., M. Wand, U. Holst, and O. Hössjer (1997): "Local polynomial variance function estimation," *Technometrics*, 39, 262–273.
- Schell, M. (1999): *Baseball's All-Time Best Hitters: How Statistics Can Level the Playing Field*, Princeton University Press.
- Schell, M. (2005): *Baseball's All-Time Best Sluggers: Adjusting Batting Performance from Strikeouts to Home Runs*, Princeton University Press.
- Schwartz, A. (2004): *The Numbers Game: Baseball's Lifelong Fascination with Statistics*, Thomas Dunne Books.
- Shandler, R. (2010): *Baseball Forecaster*, Triumph Books, 24th edition.
- Simon, G. and J. Simonoff (2006): "'Last licks': Do they really help?" *The American Statistician*, 60, 13–18.

- Smith, L. and J. Downey (2009): "Predicting baseball hall of fame membership using a radial basis function network," *Journal of Quantitative Analysis in Sports*, 5:1.
- Spanos, A. (1999): *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press.
- Thode, H. (2002): *Testing for normality*, Marcel Dekker.
- Winfrey, J. (2009): "Fan substitution and market definition in professional sports leagues," *Antitrust Bulletin*, 54, 801–822.
- Yang, T. and Y. Swartz (2004): "A two-stage Bayesian model for predicting winners in major league baseball," *Journal of Data Science*, 2, 61–73.
- Young, W., W. Holland, and G. Weckman (2008): "Determining hall of fame status for major league baseball using an artificial neural network," *Journal of Quantitative Analysis in Sports*, 4:4.