統計数理(2017)
第 65 巻 第 2 号 185–200

特集「スポーツ統計科学の新たな挑戦」
[原著論文]

ストレートに着目した空振りに影響を与える要因の定量的分析

永田 大貴[1]・南 美穂子[2]

(受付 2016 年 12 月 28 日;改訂 2017 年 5 月 17 日;採択 5 月 24 日)

要　　旨

PITCHf/x は投球の軌道を追尾することによってボールの座標や変化量などのデータを計測できるシステムである．本稿では，PITCHf/x データを用いてノビについて分析を行った．ノビとは空振りしやすいストレートに対して用いられる言葉であり，ノビのあるストレートは初速と終速の差が小さいという定説がある．しかし実際の PITCHf/x データを眺めると定説とは逆の関係が見て取れる．そこで打者のボールへのコンタクトを定義した上で，コンタクトを球速差で説明するロジスティック回帰モデルを適用した．それにより，球速差はコンタクトに対して負の関係性を有するという結果が得られた．また本稿では，ボールの変化量に着目し，変化量とコンタクトとの関係を評価するために多変量スプライン平滑法を用いた一般化加法モデルによる分析を行い，縦変化量の大きさが重要である事が分かった．さらに，ボールの質以外の各投手ごとの打ちにくさを変量効果として追加したモデルについても解析を行い，その予測値を比較する事により上原は MLB(メジャーリーグベースボール)2014 シーズンにおいて最も打ちづらい特徴を有した投手であるという結果を得た．

キーワード：PITCHf/x データ，ストレート，ノビ，ボールの変化量，一般化加法モデル，変量効果．

1. はじめに

近年，スポーツにおいてデータ活用による戦術分析や選手のパフォーマンス向上を図ろうという動きが日本でも活発化してきている．野球やサッカーを始め，国内においてはバレーボール・ラグビーなどを中心に戦術的または要因的分析を行うことにより，勝利に焦点を置いたデータ活用が行われている．野球において統計学的な見地から分析を行い，選手の評価や戦略を考える分析手法であるセイバーメトリクスはまさにその代表格と言えるであろう．野球においては様々な価値基準や選手の能力を示す指標が存在するが，セイバーメトリクスではこれらの重要性を数値から客観的に分析し，それによってプレー戦術に対し統計学的根拠を与えた．そしてスポーツアナリティクスの分野において，今最も注目を集めているのがトラッキングシステムによって得られたデータを用いた解析である．トラッキングシステムとは主に野球やサッカーなどの球技において，選手個別の動作やボールの軌跡を追跡・記録・分析するためのシステムであり，それによって取得されたデータのことをトラッキングデータと呼ぶ．本研究

[1] 慶應義塾大学大学院 理工学研究科：〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
[2] 慶應義塾大学 理工学部：〒 223–8522 神奈川県横浜市港北区日吉 3–14–1

では，野球における投球に対するトラッキングシステムである PITCHf/x により取得されるデータに着目し，ストレートのノビについて解析を行った．

1.1 PITCHf/x データ

PITCHf/x は米国 SPORTVISION 社によって開発されたシステムで，球場に設置した複数台のカメラの映像を基にして投球におけるボールの座標や軌道，速度や変化量など様々な情報を自動的に取得する．メジャーリーグベースボール(MLB)においてはこのシステムが全 30 スタジアムに設置されており，チーム内での分析やトレード，またファン向けのコンテンツとしても活用されている．PITCHf/x システムによって取得できるデータは以下のようにまとめられる．

- 座標に関するデータ：リリース点，プレート到達点
- 速度に関するデータ：初速，終速，加速度
- 変化に関するデータ：総回転数，変化量，球種

PITCHf/x データの特徴として，3 次元の座標軸と原点を定めボールの位置を計測することがあげられる．ホームプレートを原点とし，x 軸を水平方向(サード方向を負，ファースト方向を正)，y 軸を前後方向(投手方向であれば正)，z 軸を垂直方向と各軸をフィート単位で定めている．また，投球のリリース時におけるボールの速度(初速)と，プレート到達時における速度(終速)，さらにリリースからプレートまでの平均的な加速度がデータとして得られる．変化量は，ボールに回転がないという仮定のもとで到達する点と実際の(主に回転などによって引き起こされた変化による)到達点との偏差としている．ただし，ここでのプレート到達点は原点から 1.417 フィート離れた x-z 平面，リリース到達点は原点から 50 フィート離れた x-z 平面における座標点の近似値である．

今回分析に用いたデータベース内の変数 pfx は重力等の加速度を含めた軌道偏差と定義されている．変化量の計算を以下に示す(Kagan, 2009)．投手から投じられたボールの到達点は

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{v}_0 + \frac{t^2}{2}\boldsymbol{a} \tag{1.1}$$

を用いて計算される．ここではボールの動きに対して等加速(減速)度運動を仮定している．

ここで，

$$\boldsymbol{x}_t = \begin{pmatrix} \text{時刻 } t \text{ におけるボールの } x \text{ 座標} \\ 〃 \qquad\qquad y \text{ 座標} \\ 〃 \qquad\qquad z \text{ 座標} \end{pmatrix} \tag{1.2}$$

であり，$\boldsymbol{x}_0$ はリリース点の座標，$\boldsymbol{v}_0$ はリリース時における速度ベクトル，$\boldsymbol{a}$ はリリースからプレート到達時までの平均的な加速度ベクトルである．つまり，プレート到達時刻 t^* における回転のない場合の予測到達点 $\boldsymbol{x}_{t^*}$ は

$$\boldsymbol{x}_{t^*} = \boldsymbol{x}_0 + t^*\boldsymbol{v}_0 + \frac{t^{*2}}{2}\boldsymbol{a} \tag{1.3}$$

と表される．しかし，ボールに対して回転などの変化が加えられるため，予測到達点 $\boldsymbol{x}_{t^*}$ と実際のプレート到達点は異なる．したがって，変化量は

$$\begin{pmatrix} x \text{ 方向の変化量} \\ 0 \\ z \text{ 方向の変化量} \end{pmatrix} = \begin{pmatrix} \text{プレート到達時のボールの } x \text{ 座標} \\ 1.417 \\ \text{プレート到達時のボールの } z \text{ 座標} \end{pmatrix} - \boldsymbol{x}_{t^*} \tag{1.4}$$

と計算される．ここで，x 方向の変化量を横変化量，z 方向の変化量を縦変化量と呼ぶことにする．本稿では MLB の公式オンラインサイト Gameday から 2014 年レギュラーシーズンの PITCHf/x データを取得し，解析に用いた．

本稿では，ストレートの変化量や各投手の打ちづらさなどの要因が空振りに与えている影響を定量的に分析する．空振りしやすいストレートに対してはノビという言葉が用いられるが，ここではコンタクトに対して影響を与えている要因を探ることでノビについて議論を行った．まず第 2 節ではコンタクトを定義した上で，「初速と終速の差が小さいストレートがノビのあるボールである」という定説に着目し，球速差を説明変数としたロジスティック回帰分析を行い定説について議論する．第 3 節では球速差ではなく変化量に着目した分析の必要性について主張し，第 4 節では解析に用いたスプライン法による一般化加法モデリング手法について紹介した上で解析を行い，その結果に対する考察を行う．第 5 節は計測されるデータでは記述できない要因として各投手の打ちづらさを考え，変量効果としてモデルに取り入れた解析を行い各投手の打ちづらさを評価した．第 6 節ではまとめと今後の課題について述べる．

2. 球速差に着目したノビの定説の検証

本稿ではコンタクトに着目した解析を行う．日本におけるストレートは MLB ではフォーシームファストボールという名称であり，PITCHf/x データにおいては各データから球種が自動判別され記録されている．PITCHf/x データに基づいて判別された球種ラベルが FF(フォーシームファストボール)のみを対象に分析を行うこととする．

ここでコンタクトとは，打者が投球に対してバットを振りに行って当てられたかどうかを示すものである．ボールに対するコンタクトを表 1 のように定める．

打者が投球に対してバットを振りにいって空振りした時を非コンタクト，凡打・ファウル・ヒットなどボールをバットに当てることができた時をコンタクトとする．ストレートに対するコンタクトを考えることは，空振りを考える事と等しい．打者がバットに当てることが難しいストレートを投じることができることは投手にとって最大の強みともいえる．ここでは，コンタクトしにくいストレートとはどのような特徴を持つボールなのかを定量的に明らかにしたい．

2.1 ノビの定説と日本人投手の比較

ノビとは空振りしやすいストレートに対して用いられる言葉であり，初速と終速の差が小さいストレートがノビのあるボールであるという定説が存在する．しかし，PITCHf/x システムにより観測されるデータからは，定説とは逆の関係が見て取れる(金沢, 2015)．

各投球におけるボールの球速差を，球速差 = 初速 − 終速 と定める．ただし，単位はマイル/時である．また，球速差を初速で除したものを減速率とする．コンタクト率を以下のように定める．

$$(\text{コンタクト率}) = \frac{(\text{コンタクト数})}{(\text{コンタクト数}) + (\text{非コンタクト数})} \tag{2.1}$$

表 1. コンタクトの定義.

投球結果	変数の定義
空振りストライク	非コンタクト
凡打，ファウル，エラー出塁，ヒット	コンタクト

表 2. 日本人投手のストレートにおける速度に関する特徴量とコンタクト率.

投手名	初速	球速差	減速率	コンタクト率
田澤	93.9	7.86	8.4%	84%
ダルビッシュ	92.6	7.20	8.3%	86%
田中	91.3	7.15	7.8%	88%
黒田	90.9	7.52	8.3%	86%
藤川	90.6	7.78	8.6%	68%
岩隈	89.6	7.60	8.5%	81%
和田	88.9	7.08	8.0%	82%
上原	88.0	7.76	8.8%	67%

表 3. 球速差を説明変数とするロジスティック回帰モデルの推定結果.

パラメータ	推定値	標準誤差	z 値	p 値
α	3.3	0.076	43.07	$< 2 \times 10^{-16}$
β	-0.21	0.010	-21.19	$< 2 \times 10^{-16}$

PITCHf/x データから実際の投手の球速差とコンタクトとの関係を確認したい．表 2 は日本人投手の MLB2014 シーズンにおける初速，球速差の標本平均とそれを用いて計算した減速率，およびコンタクト率の表である．

これらの投手の中で，上原や藤川は一般的にノビのあると言われている部類の投手であり，実際にストレートのコンタクト率が他の投手に比べ極めて小さいことがわかる．しかし，両投手の球速差を見てみると他の投手や MLB 平均(7.40 マイル)と比べ決して小さいとは言えないどころか，むしろ大きい傾向にある．また，初速に対してどれほど減速したかを減速率として表しているが，減速率が最大となったのは上原であった．これは，これまで考えられてきた球速差の小さいストレートがノビのあるストレートであるという定説とは正反対の事実をデータが示していることになる．

2.2 球速差とコンタクトの解析

ここで，コンタクトと球速差の関係を解析するため以下のロジスティック回帰モデルを用いた解析を考える．目的変数 Y_i を i 番目の投球に対して打者がボールに対してコンタクトできたかを表す二値変数とする．MLB2014 年シーズンにおける投手が投じたストレートに対して投球結果がコンタクト・非コンタクトに該当するものを対象データとして解析を行った．モデル式と推定結果を以下に示す．

$$\log \frac{p_i}{1-p_i} = \alpha + \beta(\text{球速差})_i,\ Y_i \sim Bernoulli(p_i). \tag{2.2}$$

表 3 はモデル(2.2)における回帰係数パラメータの推定値，標準誤差，z 値，p 値をまとめたものである．

球速差の回帰係数推定値 $\hat{\beta}$ は負であり p 値が十分に小さく有意であるという結果が得られた．この結果をそのまま解釈すると，球速差が大きいストレートほど空振りが取りやすいということは否定できないという結論が与えられる．次に球速差ではなく，変化量に着目した解析を行う．

3. 変化量に着目した分析

3.1 変化量に着目した分析の必要性

先ほどのロジスティック回帰モデルの解析から，初速と終速の差が大きいストレートは空振りしやすいボールであるということを否定できない結果が得られた．以下の表 4 は対象データにおけるボールの減速率の平均値・中央値・第 1 四分位数・第 3 四分位数・標準誤差をまとめたものである．これによると減速率の平均値 8.0% ほどで，第 1 四分位数・第 3 四分位数は平均から 0.6% ほども差がないことがわかる．つまりストレートに関してはボールの減速率はそこまで大きな違いはなく，この差が打者のコンタクトに対して大きく影響しているとは考えづらい．

以上の理由から，ここでは球速差ではなく球速差とトレードオフの関係にあるボールの変化量に着目することでコンタクトを説明することを考える．変化量や球速差はボールの回転数と回転軸によって決まり互いに関係しあっているため，実際にコンタクトに対して影響を有している変数は変化量であると考え，ボールの変化量に着目しコンタクトとの関係性を分析する．

3.2 各投手の変化量の比較

日本人 6 投手のストレートの変化量を図 1 に示した．図 1 は縦・横の変化量を 2 平面にプロットした図である．各軸はインチ単位であり原点から離れた位置にある点は大きく変化している事になる．横変化量が負の値を取っている場合にはサード方向にボールが変化しており右投手であればシュートしていることになる． また，基本的にストレートの縦変化量は正の値

表 4. 減速率の平均値・中央値・第 1 四分位数・第 3 四分位数・標準誤差(パーセント単位).

第 1 四分位数	中央値	平均値	第 3 四分位数	標準誤差
7.486	8.051	8.058	8.621	0.851

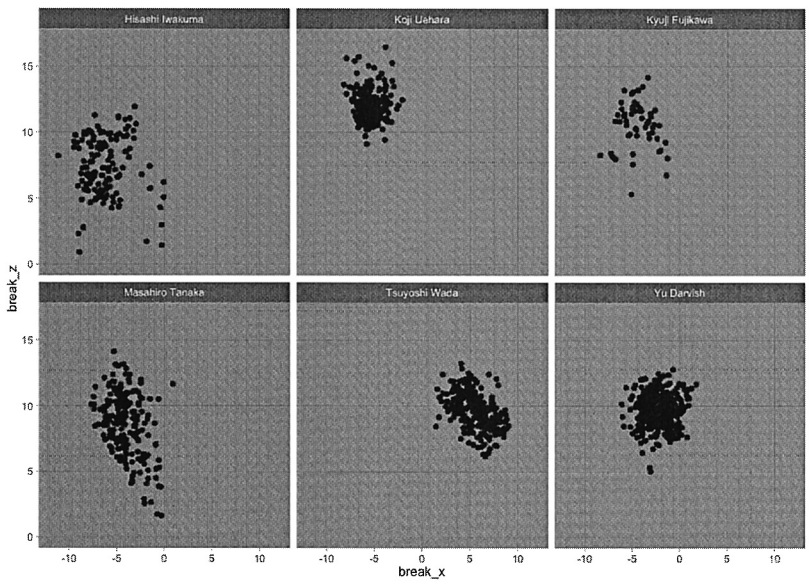

図 1. 日本人投手のストレートの変化量.

をとる．これはボールに対してバックスピンがかかる事によりボールに対して揚力が働くためである．

図 1 を見ると変化量は各投手それぞれに特徴を有することがわかる．上原や藤川は比較的縦変化量が大きく，一方で岩隈は横変化量が大きい．また，この中で唯一左投手である和田は，全く異なる変化量(特に横変化量)を有する．ダルビッシュや田中は比較的平均的な変化量である．

図 1 において上原・藤川などのノビのあると言われている投手の変化量をその他の投手と比較すると縦変化量が大きく，また実際のコンタクト率(表 2)も小さいことが分かった．そこで，PITCHf/x データにおける変化量に着目してコンタクトとの関係性を明らかにしたい．変化量は縦・横の二方向に対してデータが得られ，各変化量はコンタクトに対して単調な線形関係で影響するものではない(変化量の僅かな差がコンタクトに対して大きく影響を及ぼす可能性がある)と考えられる．また，これらとコンタクトの関係を適切に評価するには 2 変量間の交互作用を柔軟にモデリングを行う必要があるため，ここでは多変量間と目的変数との関係を柔軟にモデリングすることができる一般化加法モデルによる解析を行う．

4. スプライン平滑法を用いた解析と解析結果の考察

ロジスティック回帰モデルにおいて，変化量などの変数に対してスプライン関数 f を適用したモデルを考える．スプライン関数は局所的な特徴を捉えることを可能にする多数の基底関数の線形和で表されるなめらかな関数であり，変化量とコンタクトとの関係を柔軟にモデリングすることを可能にする．4.1 節では解析に用いたスプライン法による一般化加法モデリング手法について概要を示す．

4.1 平滑化関数を用いた一般化加法モデルによるモデリング

一般化加法モデルとは，一般化線形モデルの線形予測子に非線形関数を含むように拡張したものである(Hastie and Tibshirani, 1986)．ロジスティック回帰モデルは一般化線形モデルに含まれるモデルであり，ここでは，線形予測子に変化量などのスプライン関数を含むロジスティック回帰モデルを用いている．Thin plate regression spline 法は Wood (2006)によって提案された平滑化手法で，自然 3 次スプライン法，thin plate spline 法の柔軟性を保持しつつ計算量を抑えるように工夫されている．

ここでは，説明変数に対する非線形な関数として解析に用いた自然 3 次スプライン法とそれを多変量に拡張した thin plate spline 法についての表現とパラメータの推定方法について示す．ただし自然 3 次スプライン法は thin plate spline 法の単変量の場合を指す手法である．

まず，目的変数の平均構造を 1 次元の説明変数 x を関数 f で表すモデル

$$y_i = f(x_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2), \qquad f(x) = \sum_{j=1}^{q} \beta_j b_j(x) \tag{4.1}$$

を用いて自然 3 次スプライン法について説明する．ここで，ϵ_i は互いに独立な正規誤差であり，$b_j(x)$ はパラメータを含まない基底関数である．

関数 f は $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ の線形な関数として表されているので，目的変数ベクトルを $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ とした時に

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.2}$$

と表現できる．ここで，X の i 行 j 列成分を $X_{(ij)}$ とした時に，$X_{(ij)} = b_j(x_i)$ である．

3 次スプライン関数は，3 次多項式を 2 階微分までが連続であるようにつなぎ合わせたものであり，各区間において 3 次多項式のつなぎ目を節点(knot)という．節点の数を $q-2$ 個とし，節点を $x_1^* < x_2^* < \cdots < x_{q-2}^*$ とする．

3 次スプライン関数の基底関数の表現としては様々なものがあるが, 例えば Wood (2006)や Gu (2002)で詳細が示されているような次の表現がある．$b_1(x) = 1, b_2(x) = x, b_{j+2}(x) = R(x, x_j^*)$ であり，$R(x, z)$ は以下のように表される $(j = 1, 2, \ldots, q-2)$.

$$R(x,z) = \left[(z-1/2)^2 - 1/12\right]\left[(x-1/2)^2 - 1/12\right]/4 - \left[(|x-z|-1/2)^4 - 1/2(|x-z|-1/2)^2 + 7/240\right]/24. \tag{4.3}$$

3 次スプライン関数に対して以下の端点での 2 次微分がゼロという制約

$$f''(x_1^*) = 0, f''(x_{q-2}^*) = 0 \tag{4.4}$$

を付け加えたものが自然 3 次スプライン関数である．

関数の柔軟性は節点(基底)の数によって変化するため，節点を多くすると柔軟な関数を表現できる一方，最小二乗法による推定ではデータに当てはまりすぎて複雑な関数を選んでしまう．そこですべてのデータ点を節点とすることによって十分な柔軟性を保ちつつ，当てはまりすぎを抑えるために関数の複雑さに対して罰則を与えることで関数のなめらかさを制御することとする．つまり，罰則付き二乗誤差,

$$V(\boldsymbol{\beta}) = ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \lambda \int_\Omega f''(x)^2 dx \tag{4.5}$$

の最小化によってパラメータ $\boldsymbol{\beta}$ の推定を行う．ここで，$\lambda(> 0)$ は平滑化パラメータであり，Ω は関数を定義する空間とする．$\int_\Omega f''(x)^2 dx$ は関数の複雑さを表しており，λ は複雑さに対する罰則を調整するパラメータである．

関数 f は基底関数で $f(x) = \sum_j \beta_j b_j(x)$ と表されるので，罰則項は $\boldsymbol{\beta}$ の 2 次形式であり，罰則付き誤差二乗和はある半正定値行列 S を用いて,

$$V(\boldsymbol{\beta}) = ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \ \boldsymbol{\beta}^T S \boldsymbol{\beta} \tag{4.6}$$

と表すことができる．ここで先ほどの基底関数に対しては行列 S の各成分は $S_{(i+2,j+2)} = R(x_i^*, x_j^*)$ と表される $(i, j = 1, 2, \ldots, q-2)$．$V(\boldsymbol{\beta})$ を $\boldsymbol{\beta}$ について最小化を行うことで推定値

$$\hat{\boldsymbol{\beta}} = (X^T X + \lambda S)^{-1} X^T \boldsymbol{y} \tag{4.7}$$

を得る．

次に適切な平滑化パラメータ λ の値を選択することが必要である．λ の値を大きくとれば，推定における罰則を重くとるため比較的直線に近づき，λ の値を小さくとれば推定結果は複雑な曲線となる．平滑化パラメータの選択については一般化交差検証法(GCV; Wood, 2008)，制約付き最尤法(REML; Wood, 2011)を用いた選択などがある．本研究では一般化交差検証法を用いて選択を行った．

次に，自然 3 次スプラインと同様な考え方に基づいた多変量平滑法である thin plate spline について，ここでは簡単のため 2 変量の場合に限定して述べる．実際のモデリングでは，変化量など 2 次元の変数のコンタクトに与える影響を表すことを考えている．いま 2 次元の説明変数ベクトルを $\boldsymbol{x} = (x_1, x_2)^T$ とし，観測されたデータを $(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n$ とする．ここで，モデル

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i \tag{4.8}$$

を考える．f は x_1, x_2 について 2 階微分まで連続な関数とする．このとき，f の推定における罰則項 $J(f)$ を

$$J(f) = \int\int \left(\frac{\partial^2 f}{\partial x_1^2} + 2\frac{\partial^2 f}{\partial x_1 \partial x_2} + \frac{\partial^2 f}{\partial x_2^2} \right) dx_1 dx_2 \tag{4.9}$$

と定義すると，これを最小にする関数は，$\eta(r) = r^2 \log(r)/(8\pi)$ としたとき

$$f(\boldsymbol{x}) = \alpha_1 + \alpha_2 x_1 + \alpha_3 x_2 + \sum_{i=1}^{n} \delta_i \eta(||\boldsymbol{x} - \boldsymbol{x}_i||) \tag{4.10}$$

と表せる(Wood, 2006; Green and Silverman, 1994)．また，改めて半正定値行列 E の成分を $E_{(ij)} = \eta(||\boldsymbol{x}_i - \boldsymbol{x}_j||)$, $T_i = (1, x_{1i}, x_{2i})$, $T = (T_1, T_2, \ldots, T_n)^T$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^T$ と定める．制約 $T^T\boldsymbol{\delta} = \mathbf{0}$ を満たすとき，f を thin plate spline と呼び，当てはめは罰則付き誤差二乗和，

$$S(\boldsymbol{\alpha}, \boldsymbol{\delta}) = ||\boldsymbol{y} - E\boldsymbol{\delta} - T\boldsymbol{\alpha}||^2 + \lambda \boldsymbol{\delta}^T E \boldsymbol{\delta} \tag{4.11}$$

の $T^T\boldsymbol{\delta} = \mathbf{0}$ という条件のもとでの最小化問題となる．ここで罰則項 $J(f)$ は f の二階の微分(偏微分)により求められるため罰則は $\boldsymbol{\delta}$ にのみ依存する．

Thin plate spline は，2 階微分までが連続な関数の中で

$$||\boldsymbol{y} - g(\boldsymbol{x})||^2 + \lambda J(g) \tag{4.12}$$

を最小にするという点において最良の平滑法であり，また節点や基底関数の選択が不要であるという利点があるが，計算負荷が高くデータ点が多くなると計算時間が大きな問題となる．Thin plate regression spline (Wood, 2003)は式(4.11)における行列 E を，固有値分解により求めた固有値の大きい成分のみで構成されるランク k の行列 E_k に置き換えることにより細かい変動を除去し，基底の次元を低くして計算量を抑えている．

一般化加法モデルは，一般化線形モデルの線形予測子にスプライン項を含めるように拡張したものであり，平均構造はリンク関数 g を用いて，

$$g(\mu_i) = X_i\boldsymbol{\theta} + f_1(x_{1i}) + f_{23}(x_{2i}, x_{3i}) + \cdots \tag{4.13}$$

のように表される．スプライン項は，基底関数の線形和として表すことができるので，対数尤度関数は一般化線形モデルと同様に表せ，これに罰則項を加えた罰則付き対数尤度関数を最小化することによってモデルの当てはめを行う．

統計解析ソフト R の mgcv パッケージは thin plate regression spline を含む様々な平滑化関数を用いた一般化加法モデルによる解析を行うためのものであり，本研究ではこのパッケージを用いて thin plate regression spline を用いたロジスティック回帰モデルの当てはめを行った．

4.2 変化量に着目した解析結果の考察

本節では，変化量とコンタクトの関係に着目し，変化量に対して柔軟なモデリングを行うために thin plate regression spline を適用した解析を行う．

コンタクトを目的変数とするロジスティック回帰モデルにおいて，説明変数においては変化量に加えてコンタクトと関係があると思われるボール・ストライクカウントや球速，またプレート到達点やリリース点などの変数に対してはスプライン関数を用いることにする．まず，

変化量とコンタクトとの関係を明らかにしたい．しかし，投手の利き手によって横変化方向が異なってくるため，ここでは右投手を基準として左投手の投じた投球の横変化量の正負を反転させ，解析の対象データとした．また，コンタクトに対しては当然のようにプレート到達点が関係しているため，データの均一性を保つ目的でストライクゾーンに到達したボールのみを対象に分析を行った．ここでストライクゾーンの横幅はホームプレートの幅であり，縦幅としてはPITCHf/xによって取得される各投球ごとのストライクゾーンの上限と下限のデータの平均値を用いた．

MLB2014シーズンにおける全投手が投じたストレートで対象となったものの標本サイズは90774であった．目的変数はコンタクト(2値変数)，説明変数としては，カウント(ボール・ストライクカウント，12水準のカテゴリカル変量)，打者の対角フラグ(投手と利き手が異なる場合を1，同じ場合を0とする2値変数)，球速(リリース時の速度で連続変数)，変化量(縦・横の2次元連続変数)，プレート到達点(縦・横の2次元連続変数)，リリース点(縦・横の2次元連続変数)を用いる．球速には自然3次スプライン法，変化量，プレート到達点，リリース点には2次元thin plate regression splineを用いたスプライン項としてモデルに含めた．

スプライン関数を用いたロジスティック回帰モデルはコンタクト確率p_iが以下のように表せるモデルである．

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + f_1(z_{i1}) + f_{23}(z_{i2}, z_{i3}) + \cdots \tag{4.14}$$

ここで，$x_{i1}, x_{i2}, \ldots, z_{i1}, z_{i2}, z_{i3}, \ldots$は説明変数であり，$f_1, f_{23}, \ldots$はスプライン関数である．

表5は線形項の係数推定値とそのt値，p値である．カウントの各水準に対する推定値は，カウント0-0を基準としたものである．2ストライクであったときのカウントは有意であり推定値は正であった．これは，追い込まれたカウントにおいて打者は三振したくないという意識により，できる限りボールに対してコンタクトしにいくようなバッティング傾向になることの表れであると解釈できる．また，対角打者フラグの値は正であり，有意である．これは，野球の一般論として投手は対角の打者に対して不利であることと整合性が取れており，コンタクトにおいても，右投手に対しては左打者の方が一定量有利であると解釈できる．

表6はスプライン項に対する有効自由度とカイ二乗値，そのp値である．コンタクトに対して非線形な関係を有していると思われる変数に対してはスプライン関数(球速に対しては単変

表5．スプライン項を含むロジスティック回帰モデルによる推定結果(線形項)．

変数	推定値	t値	p値
(Intercept)	1.99	72.54	$< 2.0 \times 10^{-16}$
カウント0-1	-0.020	-0.54	0.58
カウント0-2	0.166	3.45	5.4×10^{-4}
カウント1-0	-0.014	-0.36	0.71
カウント1-1	-0.024	-0.64	0.51
カウント1-2	0.171	4.16	3.1×10^{-5}
カウント2-0	0.108	1.92	0.05
カウント2-1	0.063	1.43	0.15
カウント2-2	0.296	7.15	8.2×10^{-13}
カウント3-0	0.412	2.07	0.03
カウント3-1	0.340	5.43	5.5×10^{-8}
カウント3-2	0.494	10.76	$< 2.0 \times 10^{-16}$
対角打者フラグ	0.188	8.923	2×10^{-16}

表 6. スプライン項を含むロジスティック回帰モデルによる推定結果(スプライン項).

変数	有効自由度	カイ二乗値	p 値
球速	6.0	353.0	$< 2 \times 10^{-16}$
リリース点(横)・リリース点(縦)	17.4	463.4	$< 2 \times 10^{-16}$
横変化量・縦変化量	14.6	650.9	$< 2 \times 10^{-16}$
プレート到達点(横)・プレート到達点(縦)	32.2	4529.2	$< 2 \times 10^{-16}$

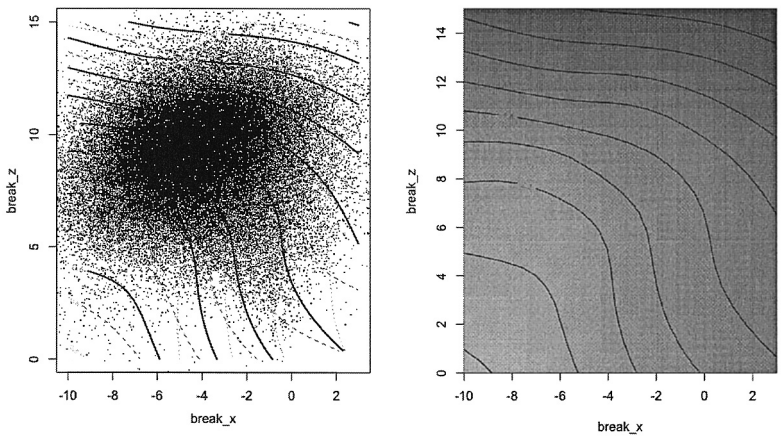

図 2. 変化量のスプライン関数の等高線とサンプル点(左), モノクロ等高線(右).

量の自然 3 次スプライン, リリース点・プレート到達点・変化量に対しては thin plate regression spline)を適用しモデルを構築した. また, スプライン項の各関数における回帰係数に対して検定統計量を構成しカイ二乗検定を行っている(Wood, 2006). 変化量に対する p 値は十分に小さく有意である. つまり, 変化量がコンタクトに対して影響を有していると言える. 他の変数についても p 値は十分に小さく, 有意水準 5% で棄却できるという結果が得られた. また, モデル全体の null deviance と deviance の差(カイ二乗値)は 3636.1 であった. この値はモデルの有効自由度 76.38 のカイ二乗分布に従う. カイ二乗値は自由度に対して十分大きく, このことからもモデル全体でも有意であるという結果が得られた.

次にモデルの AIC の比較を行う. ここでは, さきほどの変化量に対してスプライン関数を適用し推定を行ったモデルの AIC と, それぞれの変化量の変数をそのままロジスティック回帰モデルの線形項に当てはめたモデルの AIC を比較した(スプライン項を含む場合の AIC は自由度として有効自由度を用いている). その結果, スプライン項を適用したモデルは, 変化量に対して線形性を仮定したモデルに比べ AIC が 65009.8 から 64943.7 へと減少するという結果が得られた. この結果から変化量に対しては線形なモデルよりも柔軟なスプライン関数を用いたモデリングの方が, AIC の観点からは適切であると判断できる.

図 2 における左図は横変化量と縦変化量の散布図に推定した 2 次元スプライン関数の等高線図を描いたもので, 右図はスプライン関数のモノクロの等高線図である. モノクロ等高線図は, 色が濃いほど対数オッズ/コンタクト率が低い. 図 2 を見ると, 縦変化量が大きい領域(y 軸の値が 10〜15)ではコンタクト確率(対数オッズ)が比較的低く縦変化量の変化に対して対数オッズも大きく変化することがわかる. その一方で縦変化量が小さく(y 軸の値が 5〜10)かつシュート方向に変化(x 軸の値が −5〜−10)するボールはコンタクト確率(対数オッズ)が比較的

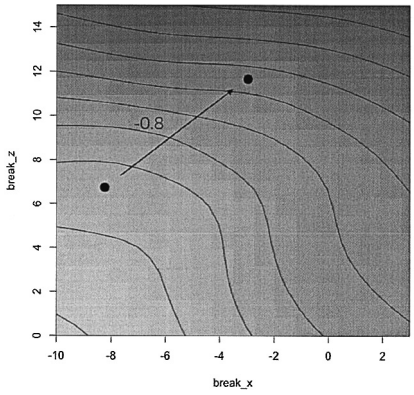

図 3．変化量の変化による対数オッズの変化．

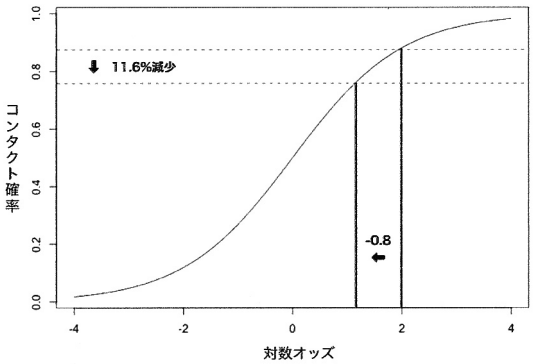

図 4．ロジスティック関数における変化．

高い．つまり，縦変化量の大きいストレートはコンタクトしにくいボールであり，縦変化量が小さくかつシュートするストレートはコンタクトしやすいボールであると解釈できる．推定結果から，縦変化量が増加するにつれてコンタクト確率は減少するという関係が見受けられ，空振りを考えた時，特に縦変化量が非常に重要であることがわかる．また単に縦変化量だけでなく，縦変化量が比較的小さい時に限って横変化量(シュート変化が小さいこと)が重要であることも分かった．これは変数に対して柔軟な関数を仮定し推定・視覚化することにより得られた結果であり，単純なロジスティック回帰モデルからはこのような解釈を得ることは難しい．

次に，変化量の変化に対する対数オッズの変化を確認することによって，コンタクトしにくさがどれほど変化するか考察を行いたい．図 3 では変化量のスプライン関数の値が大きい領域から小さい領域に対数オッズの値が変化した場合の例を示している．

図 3 のようにボールの変化量が変化した場合，対数オッズは 0.8 ほど減少すると推定された．仮に対数オッズ推定値の平均値から対数オッズが 0.8 減少したとすると，コンタクト確率は 0.875 から 0.758 へ減少し 11.6% ほど減少するという結果が得られた(図 4)．以上より，変化量はコンタクトに対して影響を有しており，変化量の変化に対してコンタクト確率が大きく変化することもわかった．

図 5 は対象データにおいて 100 球以上のストレートを投じた投手のコンタクト率をヒストグラムにしたものである．これによるとストレートの平均のコンタクト率で多いのは 86% 近辺

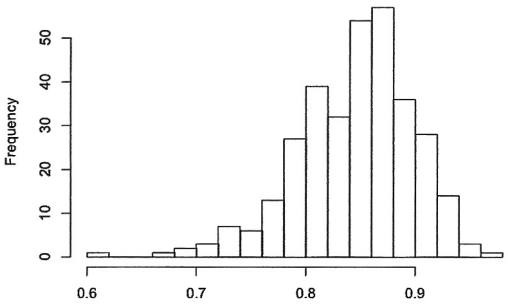

図 5. 投手ごとのコンタクト率のヒストグラム.

であり，80% を下回る投手は限られていることがわかる．推定した対数オッズの平均値から対数オッズが 0.8 減少するという仮定はあまり現実的ではないが，変化量の変化によってコンタクト確率が大きく変化するということは明らかであろう．

ここでは，変化量がコンタクトという現象に影響を有しており，変化量の変化に対するコンタクト確率の変化を定量的な観点から解釈することができた．

5. 投手の打ちづらさの評価

5.1 変量効果を用いた解析の必要性

前節では，コンタクトに関係のあると思われる変数を用いてロジスティック回帰モデルを当てはめ，それらの関係を評価した．ここで一般的にトラッキングシステムにより計測可能な変数を用いたが，投手と打者との対戦を考えた時，それらの他に考えるべき要素が存在すると考えられる．例えば，同じボールを打者に対して投球したとしても，投手のフォームやその他の持ち玉(投手が有している変化球)によってコンタクトのしづらさは異なる．

図 6 は先ほどのモデル(式(4.14))から算出された当てはめ値と実際のコンタクト率を二人の投手について示したものである．右図のダルビッシュに関しては当てはめ値がコンタクト率付近に分布しているが，左図の上原に関してはコンタクト率が当てはめ値から大きく乖離している．このことから上原においては各投球のコンタクト確率を低下させている要因があると考え

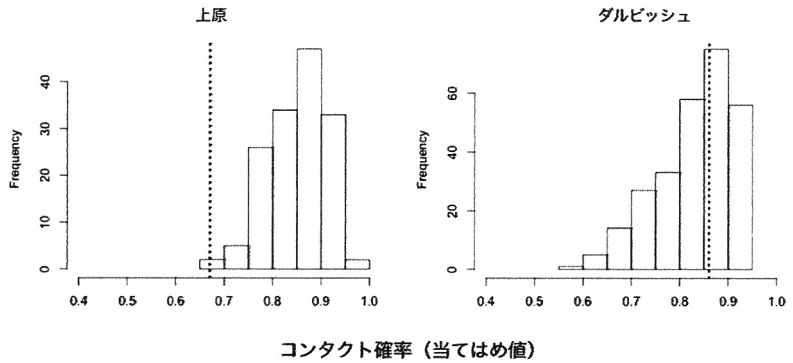

図 6. 上原(左図)とダルビッシュ(右図)におけるコンタクト確率の当てはめ値のヒストグラムとデータにおけるコンタクト率(破線).

られる．

上原投手を例にとってみると，ストレートの他にスプリット・フィンガー・ファストボール(split-finger fastball)のような縦に急激に落ちるボールを有している．また，上原のフォームの特徴に，同じフォームから異なる球種のボールを投げられること，投球時のテイクバックが小さくボールの出所がわかりにくいこと，テイクバックしてからボールが手から離れるまでの時間が短いことなどがあげられる(http://www.tokyo-sports.co.jp/sports/baseball/485297/)．これらの要因は打者に対してはストレートをコンタクトしづらくさせるため，同じ球質のボールであったとしてもそれらがコンタクトに与える影響は大きいと思われる．

5.2 変量効果を加えたモデリングと予測値についての考察

本節では，各投手が有する打ちづらさを個体差のように捉えそれらの能力を変量効果としてモデルに組み込み推定を行うことにより各投手の打者対戦における優位性を評価することを試みる．以下のような，線形予測子の中に変量効果を線形で加えたモデルを考える．

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + f_1(z_{i1}) + f_{23}(z_{i2}, z_{i3}) + \cdots + W_i\boldsymbol{\gamma} \tag{5.1}$$

このモデルは式(4.14)の線形予測子に投手の変量効果 $\boldsymbol{\gamma}$ を加えたものであり，$\boldsymbol{\gamma}$ は 2014 年シーズンにおいて投球を行った全投手の変量効果のパラメータベクトルである．W_i はその投球に対応する投手を表す変数であある．i 番目の投球が k 番目の投手によるものであれば，k 番目の要素のみが 1 であるような $W_i = [0, \ldots, 1, \ldots, 0]$ であり，$W_i\boldsymbol{\gamma}$ は変量効果 γ_k を表すものとする．また，それぞれの変量効果パラメータは $\gamma_k \sim N(0, \sigma_\gamma^2)$ に従うとする．mgcv パッケージの gam 関数ではスプライン関数内の引数を指定することにより単純な変量効果をモデルに組み込むことができる．モデルの推定に関しては再パラメータ化を行うことにより，単純な変量効果を罰則付き回帰モデルとして表すことで変量効果の予測を行うことができる．

ここで投手ごとの変量効果パラメータをモデルに加えることの妥当性を示すため，変量効果をモデルに加えなかったモデル(式(4.14))と変量効果を加えたモデル(式(5.1))とで AIC を比較した．変量効果を加えたモデルは加えなかったモデルに比べ AIC が 64943.7 から 64470.0 に大きく減少するという結果が得られた．この結果からコンタクトに対しては個々の投手の打ちづらさによる要因が大きく影響しており，変量効果を加えることによるモデリングが妥当であることが言える．

次に各投手に対しての変量効果をモデルに組み込むことで投手の打ちづらさを定量的に評価したい．特に投手の打ちづらさに興味があるため，各投手に対する変量効果を予測し，それらの予測値を投手ごとに比較することで各投手の打ちづらさの考察を行う．表 7 は予測した投手ごとの変量効果の値に関して標本サイズが 100 球以上の投手を対象に最も値が小さい 6 名について示したものである．ストレートの平均球速，一試合当たりの平均投球数(全球種)，そして主に用いている変化球(持ち球)についてまとめた．

表 7 に示したように上原は対象シーズンの MLB においてコンタクトへの投手固有の影響を示す変量効果の予測値が最も低く，つまりモデルに含めた変数以外にコンタクト率を最も大きく下げる特徴を持っているという解析結果を得た．その他の投手に関しても，スプリット・フィンガー・ファストボールやチェンジアップなど，ストレートと逆の縦変化を起こすボールを有する投手はストレートを打ちづらくさせる傾向にあることがわかった．また全体的に，一試合当たりの平均投球数の少ない中継ぎや抑えピッチャーは打ちづらいと言う結果も得られた．

鶴岡(2016)にも，上原のストレートは鉛直方向に対して極めて大きく変化しており，それが打ちづらさの要因であると記述されている．しかし，先ほども述べたように上原においてはス

表 7. 各投手の変量効果の予測値(下位 6 名).

投手名	予測値	平均球速	投球数/試合	主な持ち球（投球数が多い順）
上原浩治	-0.650	88.02	14.86	ストレート，SFF，ツーシーム
Bradley Boxberger	-0.588	93.24	16.76	ストレート，チェンジアップ，カットボール
Charlie Furbush	-0.580	91.92	9.98	ストレート，スライダー，ツーシーム
Jose Valverde	-0.498	93.07	17.80	ストレート，SFF
Nick Vincent	-0.472	90.07	13.53	スライダー，ストレート
Brandon McCarthy	-0.459	93.62	94.74	ツーシーム，カットボール，カーブ，ストレート

* SFF はスプリット・フィンガー・ファストボール

トレートの球質自体はもちろん容易に打てるものではないが，ストレートが有している特徴以外にもコンタクト確率を低下させている要素があり，それは並の投手とは比べものにならないほどのものであると考えられるため極めて低いコンタクト率が実現している．

他にも特徴的な予測値を有する投手をいくつかあげたい．日本人投手の中で予測値が低かったのは岩隈(−0.327)であった．岩隈に関しても縦に落ちるスプリット・フィンガー・ファストボールを有しており，またサイドハンドから投げる投手としては非常に大きな縦変化を有するストレートを投じることができることが要因であると思われる．また，ダルビッシュ(0.410)は平均的な投手よりも大きな値であった．

6. まとめ

本稿では PITCHf/x データを用いて，ストレートの各特徴量と空振りとの関係性を分析した．このデータの特徴は，座標や速度・変化量など多次元の変量が各球ごとに得られることであり，これらのデータを分析する上で，目的変数との多次元変量の関係を柔軟に分析できるモデルを用いたモデリングが求められる．その点において多変量スプライン平滑法を用いた解析は，有用な手法であると考えられる．今回の分析においては，ストレートの特に変化量に着目をし，コンタクトとの関係の解析を行った．空振りを考えた時，球速やコース・高さが重要であることは直感的に理解できると思うが，ストレートの変化量の違いによってコンタクトのしやすさが変わることはこれまであまり考えられていなかったように思える．そもそも，ストレートとは変化しないボールであり，変化を起こすボールは変化球であるといった認識が日本においては一般的である．しかし，トラッキングシステムが導入され普及するにつれてストレートの変化に着目することでより適切な投手の評価を行うことが期待される．また，ストレートの一つの表現としてノビがあるといった言葉が使われてきたが，その定義は曖昧であり明確な議論はされてこなかった．今回の分析において，一般的にノビのある投手と言われている上原や藤川といった類の投手は他の投手に比べストレートの縦変化量が大きく，実際のコンタクト率も低いことがわかった．また，ノビの感覚的な理解として，バットがボールの下を通過するといった表現が用いられる．もしノビが縦変化量であると仮定すれば，ボールが回転によって物理的に到達する点よりも上にあればこの現象との対応関係は取れており，そういった感覚をバッターボックスで感じることは縦変化量によるところが大きいのかもしれない．

では，はたして縦変化量の大きいストレートは良いものなのだろうか．その答えについてはさらに詳細な分析を行う必要があると考えられる．特に，飛翔する(ホームランになりやすい)ストレートやゴロアウトの取れるストレートは，おそらくボールの変化量との関係性を有しており，単に空振りといった観点からのみでなく飛翔やゴロアウトといった観点から変化量との関係性を分析する必要があるため，それはこれからの課題としたい．

さらに本稿では，各投手の打ちづらさを変量効果を用いてモデリングすることを試みた．予

測値の比較を行うことで上原や岩隈はストレートの質以外にも打ちづらい要素を持っておりそれを定量的に評価することができた．また，その他の投手においてもストレートと逆の変化をするチェンジアップやスプリット・フィンガー・ファストボールのような縦に落ちる球種を有している投手は予測値の値が小さくなる傾向にあることがわかった．

今回は PITCHf/x データに着目し，統計的な手法を用いて分析を行った．トラッキングデータは日本においてはあまり馴染みのあるものとは言えないが，徐々に様々なスポーツ・分野で導入が進んでおりこれから活用が行われていくものと思われる．また，データのみから解釈を行うだけでなく，統計的な解析により定量的な評価や，戦術的な分析を行うことが求められるであろう．

謝　辞

本論文を執筆するにあたり，原稿を注意深くお読み頂き多くの重要な指摘をして下さった 2 名の査読者の方に感謝を申し上げます．また本論文は第 5 回スポーツデータ解析コンペティションにおける発表結果をもとに作成されたものである．コンペティションの主催者である日本統計学会スポーツ分科会とデータ提供者であるデータスタジアム株式会社様にも重ねて感謝申し上げます．なお本研究の一部は，先端研究拠点事業（日本学術振興会：JSPS Core-to-Core Program）の助成を受けたものである．

参 考 文 献

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, New York.

Gu, C. (2013). *Smoothing Spline ANOVA Models*, Springer, New York.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models, *Statistical Science*, **1**, 297–318.

Kagan, D. (2009). The anatomy of a pitch: Doing physics with PITCHf/x data, *The Physics Teacher*, **42**, 412–416.

金沢慧 (2015).『「初速」と「終速」の差が小さければ良いストレートなのか？』，http://www.baseball-lab.jp/column/entry/194/（閲覧日：2016 年 11 月 30 日）．

鶴岡弘之 (2016).『上原のストレートはなぜ打たれない？ ICT で明らかに最先端テクノロジーがスポーツ市場を活性化する』，http://jbpress.ismedia.jp/articles/-/48463（閲覧日：2016 年 11 月 30 日）．

Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society*, *Series B*, **65**, 95–114.

Wood, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*, Chapman and Hall, New York.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models, *Journal of the Royal Statistical Society*, *Series B*, **70**, 495–518.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation semi-parametric generalized linear models, *Journal of the Royal Statistical Society*, *Series B*, **73**, 3–36.

Factors Affecting Batters' Contact with a Four-seam Fastball

Daiki Nagata[1] and Mihoko Minami[2]

[1]Graduate School of Science and Technology, Keio University
[2]Department of Mathematics, Keio University

In baseball, "nobi" is a four-seam fastball in which a batter has trouble making contact. Our research aims to understand the origin of nobi. It has been speculated that the velocity a four-seam fastball with nobi does not change much from the time it leaves the pitcher's hand to when it crosses the plate. Our previous analysis of nobi using PITCHf/x, which is a system that measures data such as the coordinates and break of a pitch by tracking the ball's trajectory, revealed the opposite relation. Consequently, we applied a logistic regression model to explain bat contact by the difference in the ball speed after defining the batter's contact with a pitch. A negative relation was obtained.

This study focuses on the break of a pitch. We analyzed the relationship between the break of a pitch and contact quantitatively. Additionally, we investigated the break of the ball by a generalized additive model using a multivariate spline smoothing method to evaluate the relationship between the break of the ball and bat contact. Vertical breaks are important. Moreover, adjusting the model to replace pitch quality as a random effect with hitting difficulty by pitcher revealed that in the 2014 MLB (Major League Baseball) season, Uehara was the most difficult pitcher for batters to face.

Key words: PITCHf/x data, four-seam fastball, nobi, break of the pitch, generalized additive model, random effect.