The Extinction of 0.400 Hitter in MLB With Extreme Value Distribution

Seho Park

The Department of Applied Statistics
Yonsei University

sehop1990@gmail.com

April 11, 2018

Contents

Introduction

Gould's Theory [Gould, 1996]

- The last seasonal batting average over 0.400 in the Major League Baseball (MLB) was achieved in 1941, by Ted Williams.
- In 1996, from **"Full House"** by Stephen Jay Gould, renowned evolutionary biologist,
  he argued that extinction of 0.400 batting evidenced improvement of the entire system of baseball.
- Since 1920s, variance of the MLB BA(batting average) has declined and mean remains still.
- And, BA Became stable.(Both Top-end & Bottom-end extinct)

Extreme Value Theory

- Most of Studies focused on changes of mean and variance based on all hitter data.
  ([Ahn et al, 2012], [Chatterjee and Hawkes, 1995], [Leonard, 1995], etc)
- However, Not as usual data, 0.400 hitters can be considered as extreme Values(Top-end).
- Fitting Extreme Value Distribution(GEV, GPD) could be possible. Especially, **GPD(Generalized Pareto Distribution)** is applied to this study.

Extreme Value Theory(Changes through time)

- To Evaluate Gould' Theory, we need to find out changing(especially decreasing) trend in top performance in BA.
- Detection of changes in mean[Hawkins, 1977], changes in variance[Chen and Gupta, 1997] has been studied.
- To detect smooth change pattern in parameters, **ISpline**[Ramsay, 1988] will be used.
- Likelihood ratio test could be used to detect changes in extreme values [Dierckx and Teugels, 2010]
- Parametric bootstrap has used to obtain approximate null distribution to test time-varying GEV parameters.[Chiou et al, 2015]

Extreme Value Distribution

Generalized Extreme Value Distribution(GEV)

- **Block Maxima Approach**
- Consider $X - 1, ..., X_n$ iid from $F(x)$
- The distribution of $Z_n = max(X_1, ..., X_n)$ converges to

GEV

$$G(z) = exp\left[- \left\{ 1 + \xi \left(\frac{z - \mu}{\beta} \right) \right\}_{+}^{-1/\xi} \right]$$

$$where - \infty < \mu < \infty, -\infty < \xi < \infty, \beta > 0$$

- ξ : shape Parameter
  μ : location parameter
  β : scale parameter

Generalized Pareto Distribution(GPD)

- **Peaks-over-Threshold Approach**
- Let $X_{t_1}, ..., X_{t_n}$ denotes the exceedances over a high threshold u with corresponding excesses $Y_{t_i} = X_{t_i} - u, i \in \{1, ..., n\}$
- $Y_{t_1}, ..., Y_{t_{N_t}}$ which are over threshold u follow **Generalized Pareto Distribution(GPD)**[Davison and Smith, 1990]

GPD

$$H(x) = 1 - \left[1 + \xi\left(\frac{x-u}{\beta_u}\right)\right]_+^{-1/\xi}$$

$$where -\infty < \xi < \infty, \beta_u > 0$$

- ξ : shape Parameter
  u : threshold
  β : scale parameter

Generalized Pareto Distribution(GPD)

- Approximately, the number of exceedances N_t follows Poisson distribution with λ
- $N_t \sim Poisson(\lambda(t))$ with integrated rate function $\Lambda(t) = \lambda t$
- To obtain MLE of GPD (ξ, β), asymptotic independence beteen Poisson exceedance times and GPD excesses yields

$$L(\lambda, \xi, \beta : Y) = \frac{(\lambda T)^n}{n!} exp(-\lambda T) \prod_{i=1}^{n} g_{\xi,\beta}(Y_{t_i})$$

where $Y = (Y_{t_1}, ..., Y_{t_{N_t}})$ and $g_{\xi,\beta}$ is the density of $G_{\xi,\beta}$

Generalized Pareto Distribution(GPD)

- Likelihood function can be divided by two parts.

$$\ell(\lambda, \xi, \beta; Y) = \ell(\lambda; Y) + \ell(\xi, \beta, ; Y)$$

$\ell(\lambda; Y) = \lambda T + nlog(\lambda) + log(\frac{T^n}{n!})$ and $\ell(\xi, \beta; Y) = \sum_{i=1}^{n} \ell(\xi, \beta, Y_{t_i})$ with

$$\ell(\xi, \beta; y) = \begin{cases} -log(\beta) - (1 + 1/\xi)log(1 + \xi y/\beta), & if \;\; \xi = 0, \\ -log(\beta) - y/\beta, & if \;\; \xi = 0 \end{cases}$$

- And we can maximize likelihood separately.

Method

Non-stationary GPD with GAM
[V. Chaves and A. C. Davison, 2005]

- For Non-stationary model, We can assume model parameters depend on time t

$$\theta_i = g_i\{x^T\eta_i + h_i(t)\}, \quad i = 1, ..., r$$

  g_i : link function
  $\eta_i \in \mathbb{R}^p$
  h_i : smooth nonparametric function

- With this form, $\hat{\theta} \in \mathbb{R}^r$ can be estimated by using **penalized log-likelihood**

$$\ell(\theta; y) - \sum_{i=1}^{r} \left[\gamma_i \int_{\dashv} h_i''(t)^2 dt \right]$$

Method

Non-stationary GPD with GAM
[V. Chaves and A. C. Davison, 2005]

- Because of computational difficulty, ξ and β can be reparameterized as
$$(\xi, \beta) \to (\xi, \nu(\xi, \beta))$$
with $\nu(\xi, \beta) = log((1 + \xi)\beta)$ which is orthogonal to ξ
- Therefore, reparameterized log-likelihood is
$$\ell^r(\xi, nu; Y) = \ell(\xi, exp(\nu)/(1 + \xi); Y)$$
- And we can assume
$$\xi = \xi(x, t) = x^T \eta_\xi + h_\xi(t)$$
$$\nu = \nu(x, t) = x^T \eta_\nu + h_\nu(t)$$

Method

Non-stationary GPD with GAM

[V. Chaves and A. C. Davison, 2005]

- To fit smooth functions h_ξ, h_ν, penalized log-likelihood

$$\ell^p(\eta_\xi, h_\xi, \eta_\nu, h_\nu; z_1, ..., z_n) =$$

$$\ell^r(\xi, \nu; y) - \gamma_\xi \int_0^T h_\xi''(t)^2 dt - \gamma_\nu \int_0^T h_\nu''(t)^2 dt$$

  where $\gamma_\xi, \gamma_\nu \geq 0$

- If we assume $0 = s_0 \leq s_1 \leq ... \leq s_m < rs_{m+1} = T$ are knots for smooth spline, then

$$\int_0^T h''(t)^2 dt = h^T K h$$

  where $h = (h(s_1), ...h(s_m))$ and K is a symmetric matrix of rank $m - 2$

Method

Non-stationary GPD with GAM
[V. Chaves and A. C. Davison, 2005]

- then penalized log-likelihood can be rewritten as

$$\ell^p(\eta_\xi, h_\xi, \eta_\nu, h_\nu; z_1, ...z_n) = \ell^p(\xi, \nu; y) - \gamma_\xi h_\xi^T K h_\xi - \gamma_\nu h_\nu^T K h_\nu$$

  with $h_\xi(s_1) = (h_\xi(s_1), ..., h_\xi(s_m))$ and $h_\nu = (h_\nu(s_1), ..., h_\nu(s_m))$
- Using back-fitting algorithm to estimate parameters(ξ, ν) simultaneously.
- In this procedure, instead of natural cubic spline, **monotone spline(ISpline)** will be used to detect decline trend.

Method

I-Spline

[Ramsay, 1988]

- Basis Spline function in I-Spline is M-Spline.

$$M_i(x|1,t) = \frac{1}{t_{i+1} - t_i}, \quad t_i \leq x < t_{i+1}$$

and 0 otherwise.

$$M_i(x|k,t) = \frac{k[(x - t_i)M)i(x|k-1,t) + (t_{i+k} - x)M_{i+1}(x|k-1,t)]}{(k-1)(t_{i+k} - t_i)},$$

$$k > 1$$

- I-Spline is integration of M-Spline with formula

$$I_i(x|k,t) = \int_L^x (M_i(u|k,t)du)$$

where L is the lower limit of the domain of the splines.

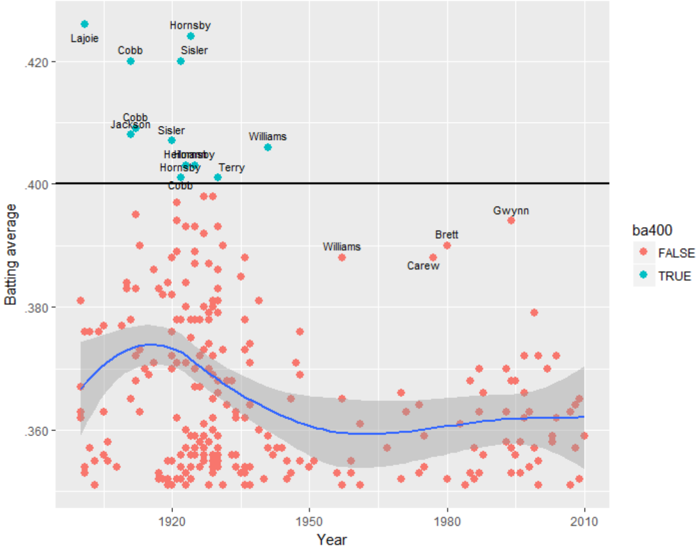
- Because M-spline is non-negative spline, I-Spline has monotonicity.

MLB Data Analysis

- From "Lahman Database", 1871-2017 baseball datasets are available.
- Before 1900, game rules are a lot different from present.
- And to filter eligible hitter for study, Plate Appearance should be over 450.
- $1900 \leq year \leq 2017$ and $PA \geq 450$

MLB Data Analysis

- Plot of eligible hitter over 0.350 BA

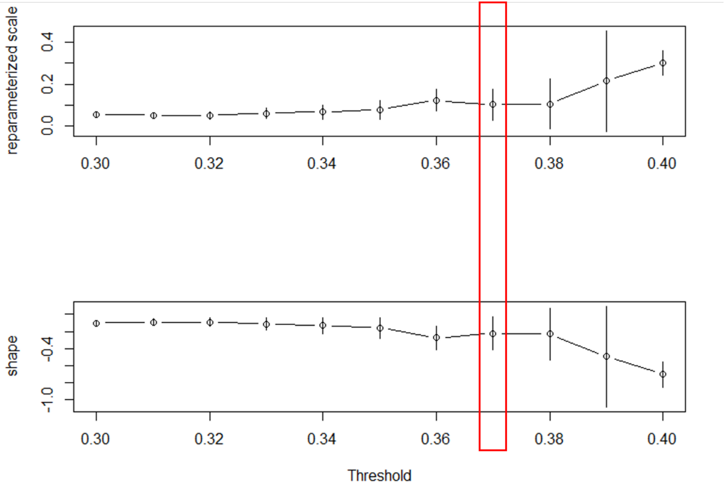


GPD Fitting

Choosing Threshold

- Before fitting GPD to the data, we need to set proper threshold which can determine extreme values.
- Sufficiently **high threshold** in torder taht the theoretical justification applies reducing bias.
- However, **low threshold** enough in order to reduce the variance of the estimates.
- Use threshold range plot to select low variance

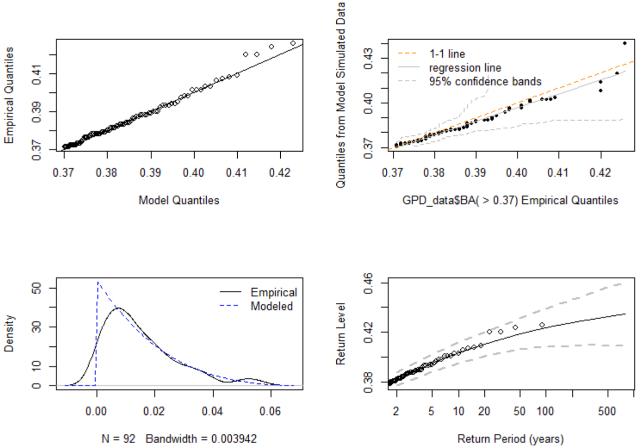
Threshold Range



- Choose 0.37 as Threshold for GPD model.

Fitting GPD

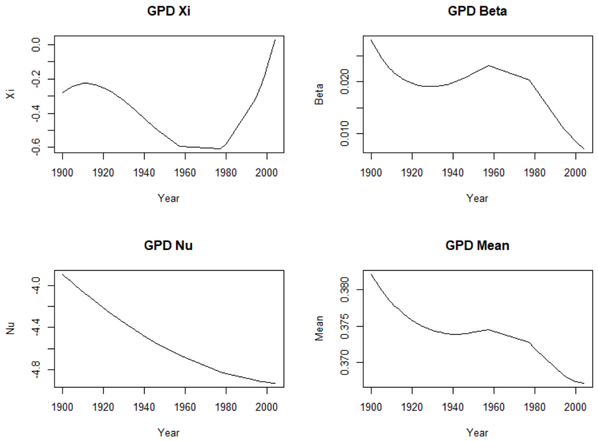
- $\beta = 0.01860762, \xi = -0.22220840$
- with Log-likelihood of 294.9881



GPD Fitting

Fitting GPD with Non-stationary GAM parameters

- I-Spline with degree $= 2$, on ξ and β, time varying parameters and means plots are shown below.
- log-likelihood : 298.7039



Further Study

Model Comparison

- Constant vs Time-Varying
- Natural Cubic Spline vs I-Spline(monotonicity)

Likelihood Ratio Test with Bootstrap

- Because null model and alternative model are not nested, we can use general parametric bootstrap
- The likelihood ratio statistics : $T_n = -2log[L_0(X)/L_1(X)]$
- By Bootsrapping, we can obtain approximate null distribution of T_n
- then approximate p-value of T_n can be estimated.

Apply to other dataset

- There're 2 leagues in MLB.(National League American League)
- KBO Data

References

Gould S.J. (1997)
Full House: The Spread of Excellence from Plato to Darwin.
Three Rivers Press

Ahn, j. Y., G.J. Byun, Y. Chekal, G. Cheon, B. H. Cheon, S.H. Cho, S. Choi, K. H. Chung, B. Hong, S. Jang, W. C. Jang, S. M. Jeon, J. Jeong, K. Jun, Y. Jung, M. Kang, K. M. Kim, K. Kim, D. J. Kim, D. S. Kim, L. Y. Kim, S. M. Kim, S. W. Kim, Y. J. Kim, Y. N. Kim, Y. K. Kim, J. H. Kim, T. H. Kim, H. Kim, H. I. Kim, J. H. Kwon, K. Lee, S. H. Lee, S. K. Lee, J. S. Lee, J. S. Lee, C. H. Lee, H. K. Lee, S. Lim, S. Lim, S. Nam, S. W. Nam, N. H. Noh, J. m. Noh, W. Oh, M. Pak, S. H. Park, S. Park, S. Park, J. H. Park, C. E. Park, H. J. Park, B. G. Shin, E. J. Song, M. Yi, and S. Y. Yoon (2012)
Why 0.400 hitters have disappeard: 30-year evolution of the Korean Professional Baseball League. Baek In-Cheon Project

Chatterjee S, Hawkes J (1995)
Apparent decline as sign of improvement? Or, can less be more?
Teaching Statistics 17, 82-83

References

LeonardII W. M. (1995)
The decline of the .400 hitter: an explanation and a test.
Journal of Sport Behavior 18, 226-236

Hawkins, D. M (1977)
Testing a sequence of observations for a shift in location.
Journal of the American Statistical Association 72(357), 180-186

Chen, J. and A. Gupta (1997)
Testing and locating variance changepoints with application to stock prices.
Journal of the American Statistical Association 92(438), 739-747

Ramsay, J. O (1988)
Monotone regression splines in action.
Statistical Science 3(4), 425-441

Dierckx, G. and J. L.Teugels (2010)
Change point analysis of extreme values.
Environmetrics 21(7-8), 661-686

References

S. H. Chiou, S. W. Kang, J. Yan (2015)
Change point analysis of top batting Average
Extreme Value Modeling and Risk Analysis: Methods and Applications

A. C. Davison, R. L. Smith (1990)
Models for exceedances over high threshold
Journal of the Royal Statistical Society 52(3), 393-442

V. Chavez-Demoulin, A. C. Davison (2005)
Generalized Additive Modelling of Sample Extremes.
Journal of the Royal Statistical Society 54(1), 207-222

V. Chavez-Demoulin, A. C. Davison (2005)
Generalized Additive Modelling of Sample Extremes.
Journal of the Royal Statistical Society 54(1), 207-222

The End