Survival Analysis Project Report

생존분석을 통한 메이저리그 타자들의 홈런을 기록할 때까지의 시간 분석

2019년 12월 10일

서울대학교

통계학과 2014-16604 고우진

통계학과 2014-12512 김민국

Abstract

본 연구의 목적은 2019시즌 메이저리그 타자들이 5, 10, 20, 30번째 홈런을 칠 때까지 걸리는 타석수를 토대로 생존 분석을 진행함으로써 “에이스는 우투좌타”라는 야구계의 통상적인 말이 홈런의 경우에도 적용이 되는지 확인해 보는 것이다.

본 연구에서 사용한 데이터는 Rotowire에서 제공하는 “2019 MLB PLAYER STATS DATA”를 기반으로 직접 2019시즌 시작일부터 5, 10, 20, 30개 홈런을 칠 때까지의 타석수와 키, 몸무게, 나이 등의 정보를 조사한 후 추가해서 완성하였다.

용이한 분석을 위해 Failure time(5, 10, 20, 30번째 홈런까지의 타석수)과 Censoring time(2019시즌이 끝날 때까지의 타석수)의 독립성을 가정하였다. 또한, 데이터간의 독립성 가정을 위해 지구별 이적이 있었던 선수들의 데이터는 삭제하였다.

본 보고서에서 진행한 연구는 5, 10, 20, 30개 홈런을 칠 때까지의 타석수에 대한 생존함수를 좌타자, 우타자, 양손타자 별로 구분하여 모수적 방법과 비모수적 방법을 이용하여 비교하였고, Log-rank test를 통하여 각각의 경우에 좌·우·양손 타자별로 생존함수의 차이가 있는지 검정하였다. 또한, 5, 10, 20, 30개 홈런까지의 타석수에 대한 분포 가정을 하여 parametric regression model인 AFT model을 적합하여 그에 대한 가정이 맞는지 확인하였고, 이를 보완하여 semi-parametric regression model인 Cox-PH model을 적합하여 추가적으로 어떤 요인(나이, 키, 몸무게, 리그)이 유의한지 확인하였다.

Log-rank test, AFT model, Cox-PH mode의 결과에서 모두 좌·우·양손 타자에 따른 5, 20, 30번째 홈런까지의 생존 함수에 유의한 차이가 없었지만 10번째 홈런까지의 생존 함수에서는 유의한 차이가 있다는 것을 확인할 수 있었다. 특히나 AFT model과 Cox-PH model의 경우 coefficients의 값과 유의성 여부를 통해 좌·우·양손 타자의 생존함수 간의 차이가 어느 정도인지를 알 수 있었고, 추가적 변수(나이, 키, 몸무게, 리그)의 영향에 대해서도 확인할 수 있었다.

Index

1. Introduction

1.1 Motivation & Goal of Analysis

야구에서는 “에이스는 우투좌타”라는 말이 있다. 우투좌타란 타격 때는 왼손으로 공을 치고, 수비할 때는 오른손으로 공을 던지는 야구선수를 의미한다.

일반적으로 타자의 경우 우타석보다 좌타석에서 타격하는 것이 유리하다. 다음에는 2가지 이유가 있다. 첫 번째로, 왼쪽 타석이 오른쪽 타석보다 1루까지의 거리가 한 발짝가량 가깝고, 스윙 후에도 우타자는 몸이 3루 쪽으로 향하지만, 좌타자는 몸의 방향이 자연스럽게 1루를 향하게 되므로, 접전상황에서 출루 시에 한 발짝 이상 이익을 본다. 두 번째로, 오른손잡이가 왼손잡이보다 훨씬 많기 때문에, 상대적으로 우투수의 수가 많다. 좌타자는 우투수를 상대로 유리한 점을 갖는다. 좌타석에 들어설 때 우투수가 던지는 공을 조금이라도 더 오래 볼 수 있다. 따라서 타격 시 유리하다. 우타자보다 좌타자가 가지는 유리한 점들로 인해 좌타자의 타율이 우타자보다 높게 기록된다.

2017년 11월, ‘뉴잉글랜드 저널 오브 메디신’에 네덜란드 암스테르담자유대 인간운동과학연구소 다비드 만 교수팀은 우투좌타의 성공을 보여주는 논문을 발표하였다. 논문에 따르면 우투좌타의 선수들이 다른 유형의 타자들보다 메이저리거가 될 확률과 높은 타율을 유지한 채 은퇴를 할 확률이 높다고 하였다. 하지만 타율을 기준만으로 한 이러한 결과에 대해 미국 일리노이주립대의 앨런 네이선 교수는 ‘홈런의 지표를 기준으로 삼았다면 다른 결과가 나왔을지도 모른다.’라고 인터뷰를 하기도 했다.

본 프로젝트에서는 홈런의 경우 좌·우 타자 간의 차이가 있는지를 분석해 보고자 한다. 2019시즌 메이저리그 타자들의 5개, 10개, 20개, 30개 홈런까지의 기간에 대한 생존함수를 비교·분석하고 추가적으로 데이터 내에 기재된 구체적인 범주(지구, 나이, 키, 몸무게)가 좌·우 타자의 홈런까지 기간에 영향을 미치는지 확인하고자 한다.

1.2 Data Description

본 프로젝트에서는 https://www.rotowire.com/baseball/stats.php “2019 MLB PLAYER STATS” 자료를 사용하여 데이터를 만들었다. 해당 자료에는 메이저

리그 타자들의 2019시즌 중의 소속(팀, 지구), 타율, 홈런수, 타석수, 경기수 등이 기본적으로 제공되었다. 2019시즌의 시작일부터 5개, 10개, 20개, 30개 홈런을 칠 때까지의 타석수를 직접 계산하여 입력하였으며 이외에도 좌·우타 여부, 키, 무게, 나이와 같은 정보를 직접 조사해서 데이터에 추가하였다. 해당 조건에 맞는 개수의 홈런을 치지 못한 선수들은 right censored data로 처리하였다.

Player	Age	A group	Bat	B group	Height(inc	H group	Weight	W group	5thHR	5thC	10thHR	10hC	20thHR	20thC	30thHR	30thC	Team	League	HR
A.J. Polloc	31	2	R	0	73	1	193	1	134	1	245	1	342	0	342	0	LAD	0	15
Aaron Altl	28	1	R	0	77	2	220	2	66	0	66	0	66	0	35	0	NYM	0	1
Aaron Hic	30	2	B	2	73	1	205	1	131	1	209	1	255	0	255	0	NYY	1	12
Aaron Jud	27	1	R	0	79	2	230	2	89	1	165	1	369	1	446	0	NYY	1	27
Abiatal Av	24	0	R	0	71	1	186	1	8	0	8	0	8	0	8	0	SF	0	0
Abraham	30	2	B	2	69	0	170	0	38	0	38	0	38	0	38	0	ARI	0	1
Abraham	22	0	B	2	72	1	210	1	89	0	89	0	89	0	89	0	HOU	1	2
Adalberto	24	0	B	2	73	1	165	0	156	1	443	0	443	0	443	0	KC	1	9
Adam Du	31	2	R	0	73	1	205	1	26	1	130	1	130	0	130	0	ATL	0	10
Adam Eat	30	2	L	1	68	0	184	1	224	1	518	1	656	0	656	0	WAS	0	15
Adam Enc	27	1	R	0	74	1	215	1	210	1	248	0	248	0	248	0	CWS	1	6
Adam Fra	27	1	L	1	71	1	170	0	387	1	570	1	608	0	608	0	PIT	0	10
Adam Ha	23	0	L	1	73	1	195	1	179	1	242	0	242	0	242	0	PHI	0	5
Adam Jon	34	2	R	0	74	1	215	1	80	1	206	1	527	0	527	0	ARI	0	16
Addison F	25	1	R	0	72	1	200	1	103	1	241	0	241	0	241	0	CHC	0	9
Adeiny He	30	2	R	0	71	1	180	1	100	1	221	0	221	0	151	0	NYM	0	5
Adrian Sa	29	1	R	0	72	1	160	0	32	0	32	0	32	0	32	0	WAS	0	0
AJ Reed	26	1	L	1	76	2	240	2	49	0	49	0	49	0	49	0	CWS	1	1
Albert Aln	25	1	R	0	74	1	180	1	146	1	302	1	363	0	363	0	CHC	0	12
Albert Puj	39	2	R	0	75	2	240	2	120	1	199	1	450	1	545	0	LAA	1	23

위의 그림은 분석에 필요한 데이터 일부이다. 아래의 표는 데이터 변수에 해당하는 변수명과 변수가 가지는 값을 설명한다.

Variable name	Detail
Player	이름
A.group	0 : ~ 25세 1 : 25 ~ 30세 2 : 30세 ~
B.group	0 : 우타자 1 : 좌타자 2 : 양손타자
H.group	0 : ~ 70inch 1 : 70 ~ 75inch 2 : 75inch ~
W.group	0 : ~ 80kg 1 : 80 ~ 100kg 2 : 100kg ~
X5thC	0 : 5개 홈런 X 1 : 5개 홈런 O
X10thC	0 : 10개 홈런 X

Variable name	Detail
Age	나이
Bat	R : 우타자 L : 좌타자 B : 양손타자
Height.inch	키(inch)
Weight	몸무게(파운드)
X5thHR	5번째 홈런까지 타석 수
X10thHR	10번째 홈런까지 타석 수
X20thHR	20번째 홈런까지 타석 수
X30thHR	30번째 홈런까지 타석 수

	1 : 10개 홈런 O
X20thC	0 : 20개 홈런 X 1 : 20개 홈런 O
X30thC	0 : 30개 홈런 X 1 : 30개 홈런 O
League	0 : 아메리칸 리그 1. : 내셔널리그

Team	소속 팀
HR	홈런 수

아래의 그림은 우타자, 좌타자, 양손타자의 5, 10, 20, 30개의 홈런을 기준으로 right censored data 수와 failed data의 개수를 보여준다.

X5thC

B.group	0	1	Total
0	145	196	341
1	77	119	196
2	30	41	71
Total	252	356	608

X10thC

B.group	0	1	Total
0	193	148	341
1	110	86	196
2	43	28	71
Total	346	262	608

X20thC

B.group	0	1	Total
0	270	71	341
1	158	38	196
2	56	15	71
Total	484	124	608

X30thC

B.group	0	1	Total
0	308	33	341
1	177	19	196
2	66	5	71
Total	551	57	608

데이터의 독립 가정을 위반하는 데이터를 제거하고 남은 선수의 수는 608명이다. 608명 중 우타자는 341명, 좌타자는 196명, 양손타자는 71명이다. 341명의 우타자 중 196명의 선수가 5개 이상의 홈런을 쳤으며, 148명의 선수가 10개 이상의 홈런을, 71명의 선수가 20개 이상의 홈런을, 33명의 선수가 30개 이상의 홈런을 쳤다. 196명의 좌타자 중 119명의 선수가 5개 이상의 홈런을 쳤으며, 86명의 선수가 10개 이상의 홈런을, 38명의 선수가 20개 이상의 홈런을, 19명의 선수가 30개 이상의 홈런을 쳤다. 71명의 양손타자 중 41명의 선수가 5개 이상의 홈런을 쳤으며, 28명의 선수가 10개 이상의 홈런을, 15명의 선수가 20개 이상의 홈런을, 5명의 선수가 30개 이상의 홈런을 쳤다.

1.3 Assumptions

데이터의 간의 독립을 위해 지구별 이적이 있었던 선수들의 데이터는 삭제하였다. 또한, Failure time(5개, 10개, 20개, 30개 홈런까지의 타석수)와 Censoring time은 독립을 가정하였다.

2. Results & Interpretation

2.1 Non-parametric method

본 연구에서의 data set은 right censored data이므로 $((T_i,\delta_i)_{i=1,2,\ldots,n})$이 관측되고 Y(t)는 number of risks at time t, dN(t)는 the number of deaths at time t로 해석할 수 있다. 따라서 다음의 notation이 성립한다.

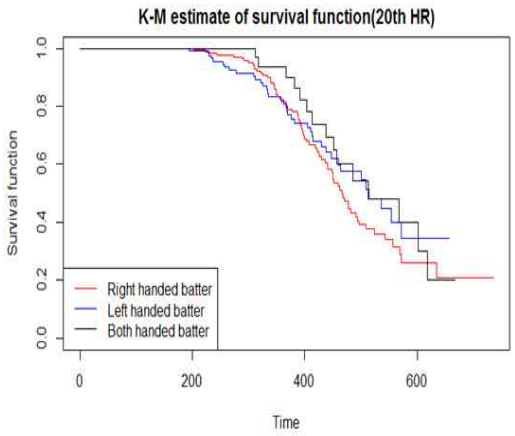
$$Y(t) := Y_i := \sum_{i=1}^{n} I(T_i \geq t) \text{ 이고, } dN(t) := d_i := \sum_{i=1}^{n} I(T_i = t, \delta_i = 1) \text{ 이다.}$$

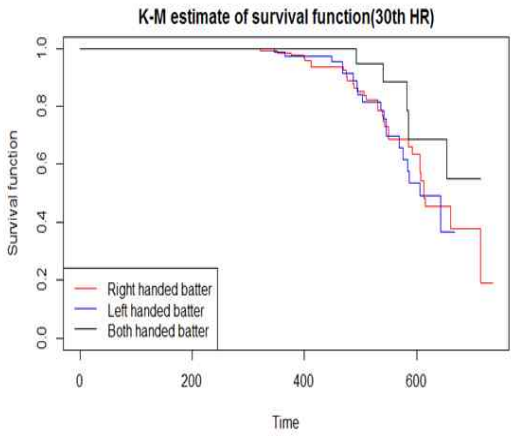
2.1.1 Kaplan Meier estimation

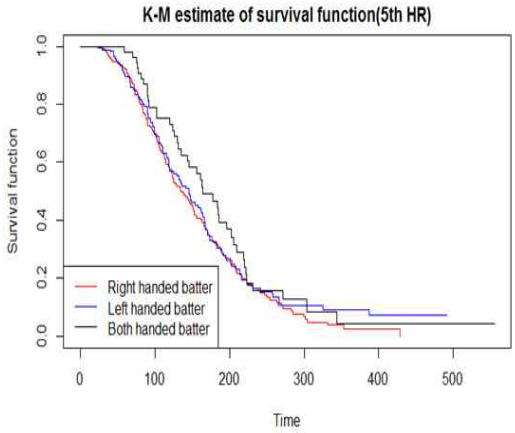
Kaplan Meier estimation은 위의 notation을 이용하여 나타내면 다음과 같다.

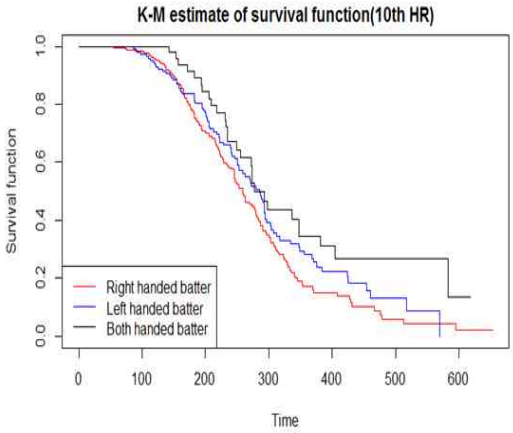
$$S^{KM}(x) = \prod_{u \leq x} [1 - \frac{dN(u)}{Y(u)}]$$

아래 4개의 그래프는 survival package의 surv 함수를 이용하여 각각 5개, 10개, 20개, 30개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 Kaplan Meier estimation으로 추정해 한 번에 도식화한 것이다. 빨간선은 추정된 우타자의 생존함수이고, 파란선은 추정된 좌타자의 생존함수이며 검정색은 추정된 양손타자의 생존함수를 나타낸다.







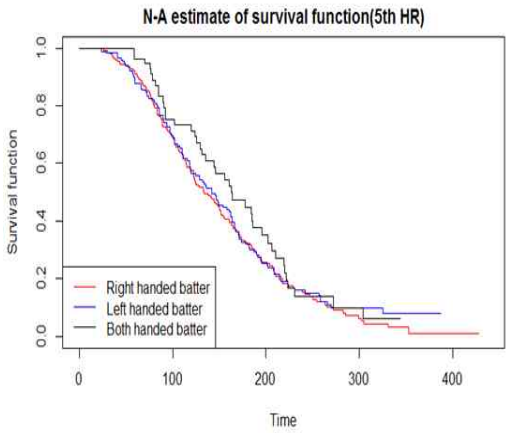


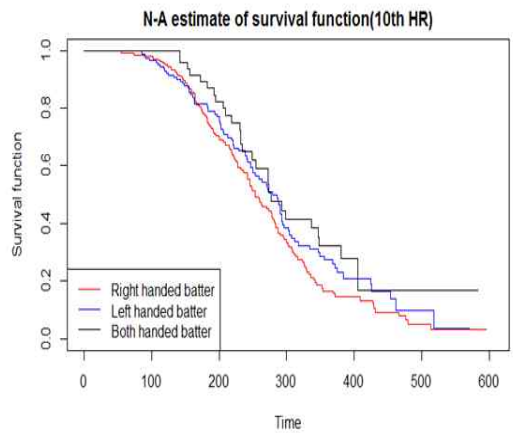
2.1.2 Nelson Aalen estimation

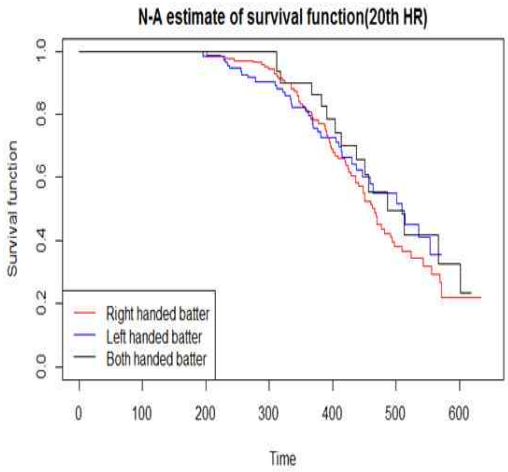
Nelson Aalen estimation 역시 위의 notation을 이용하여 나타내면 다음과 같다.

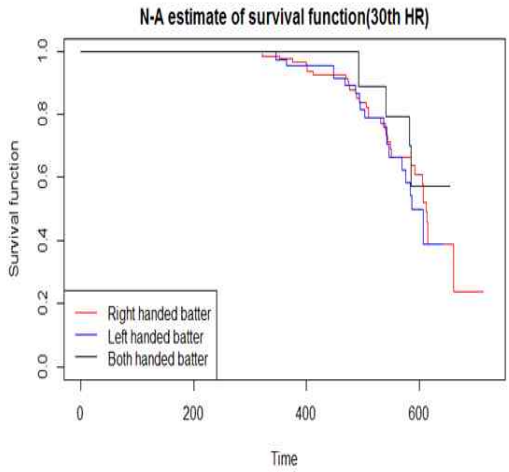
$$H^{NA}(x) = \sum_{u \le x} \frac{dN(u)}{Y(u)}$$

아래 4개의 그래프는 각각 5개, 10개, 20개, 30개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 Nelson Aalen estimation으로 추정해 한 번에 도식화한 것이다. 빨간선은 추정된 우타자의 생존함수이고, 파란선은 추정된 좌타자의 생존함수이며 검정색은 추정된 양손타자의 생존함수를 나타낸다. Kaplan Meier estimation으로 추정한 각각의 생존함수와 거의 유사한 형태를 보이는 것을 확인할 수 있다.









2.2 Parametric method

Right censoring data에서 Time에 대한 분포 가정을 한다면 다음과 같은 Likelihood function을 유도할 수 있다.

$$L(\theta;x,\delta) \propto \prod_{i=1}^{n} h(x;\theta)^{\delta_i} S(x;\theta)^{1-\delta_i}$$

이 절에서는 Data가 모수 λ인 지수 분포를 따르는 경우와 모수 λ와 α을 갖는 Weibull 분포 가정을 하여 2가지 경우에 대해 생존함수를 추정, 분석하였다.

2.2.1 Exponential distribution

Data가 모수 λ인 Exponential 분포($pdf(x;\lambda) = \lambda e^{-\lambda x}$)을 따른다면 hazard function은 $h(x;\lambda) = \lambda$, 가 되고 생존함수는 $S(x;\lambda) = e^{-\lambda x}$가 된다. Exponential 분포를 가정하면 Likelihood function은 $L(\theta;x,\delta) = \lambda^{\sum_{i=1}^{n}\delta_i} \exp(-\lambda \sum_{i=1}^{n} t_i)$로 계산된다. 이 경우 모수 λ의 추정값은

$$\hat{\lambda} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} T_i}$$

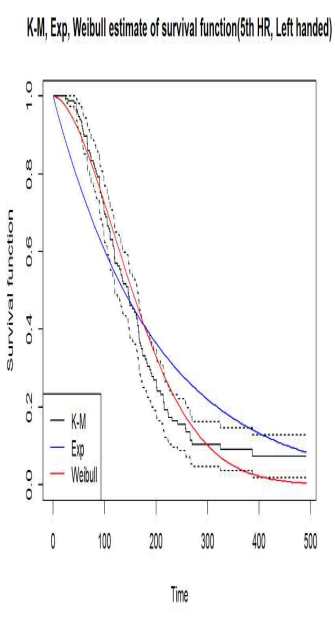
로 얻어지며, 이를 이용하여 생존함수를 추정할 수 있다.

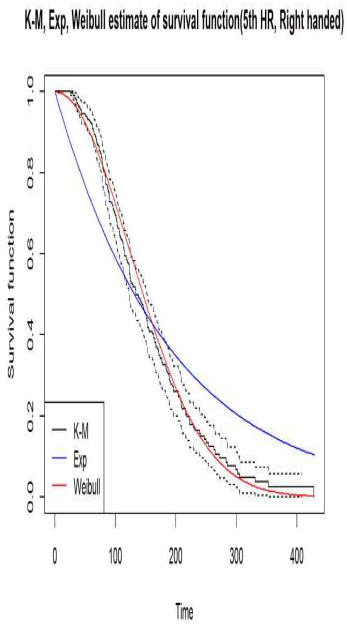
2.2.2 Weibull distribution

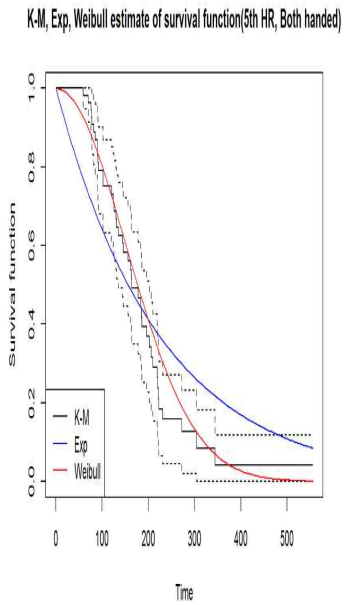
Data가 모수 λ와 α을 갖는 Weibull 분포($pdf(x;\lambda,\alpha)=\alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}$)을 따른다면, hazard function은 $h(x;\lambda,\alpha)=\alpha\lambda x^{\alpha-1}$ 가 되고, 이에 해당하는 생존함수는 $S(x;\lambda,\alpha)=e^{-\lambda x^{\alpha}}$이다. 하지만 모수가 2개인 Weibull 분포의 경우 Likelihood function이 closed form으로 주어지지 않기 때문에 모수에 대한 추정을 근사적으로 해결해야 한다. 본 프로젝트에서는 R 패키지 중 하나인 parmsurvfit을 통해서 Weibull 분포의 모수 α의 추정값을 얻었다. 이렇게 얻은 $\hat{\alpha}$을 이용하여 모수 λ의 추정값을 다음과 같은 식을 통해 추정하였다.

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}T_i^{\hat{\alpha}}}$$

아래 그림들은 순서대로 5개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 비모수적 방법인 Kaplan Meier estimation과 모수적 방법인 Exponentail 분포를 가정했을 때와 Weibull 분포를 가정했을 때 추정된 생존함수들을 그린 것이다.



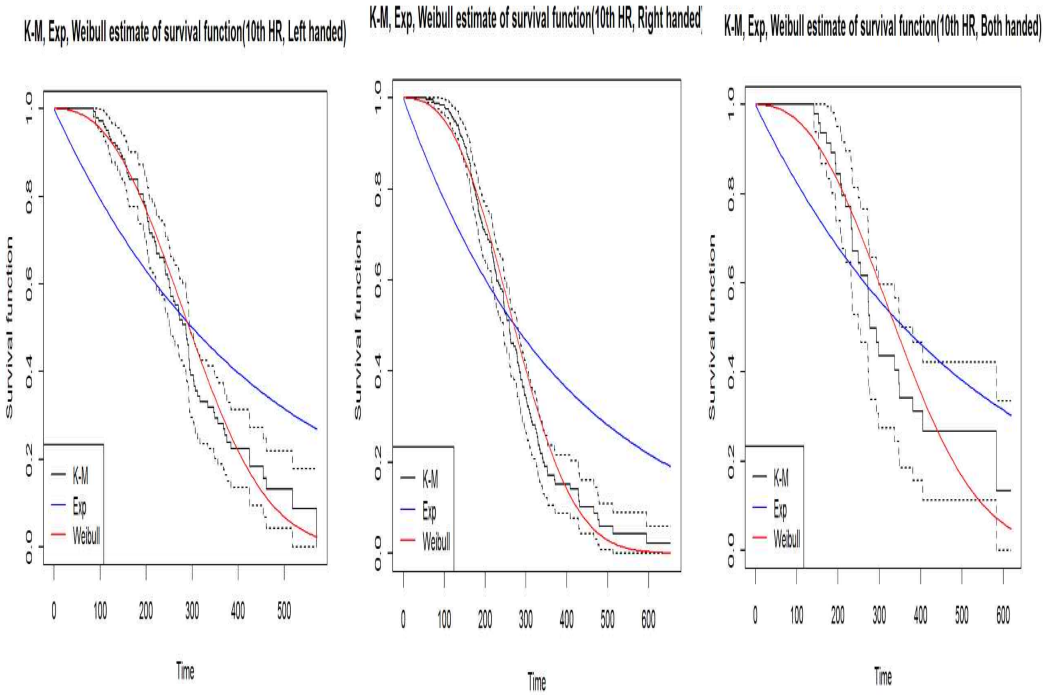




검은색 선은 5번째 홈런까지의 타석수를 Kaplan Meier estimation로 추정한

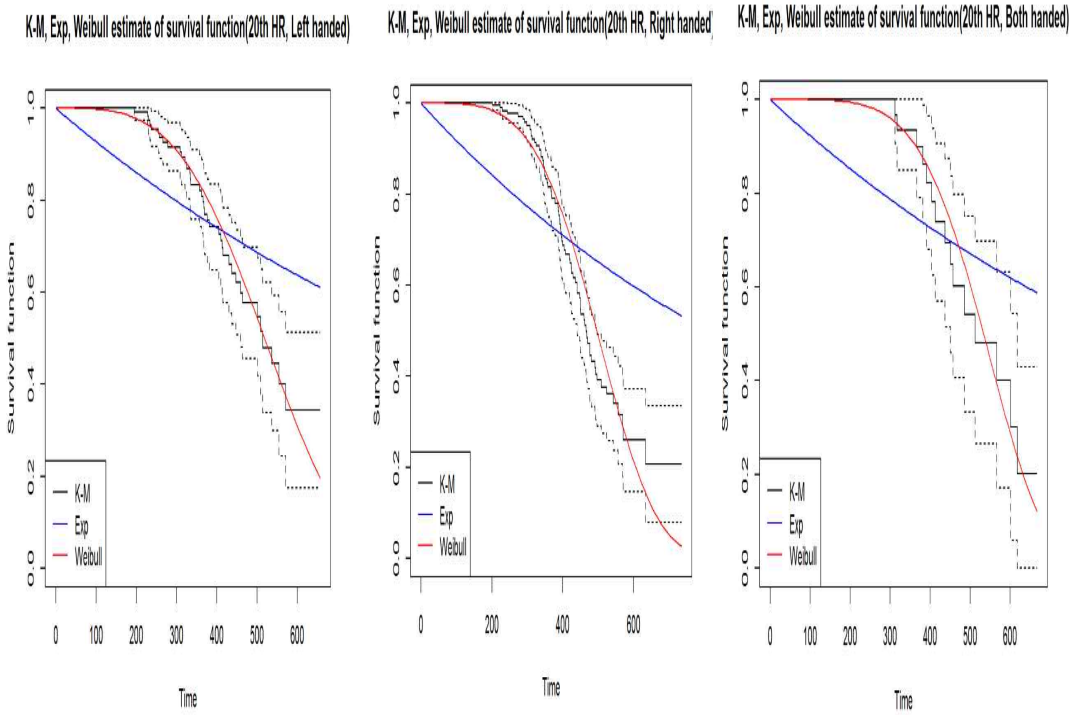
생존함수이며 파란색 선은 Expoenetial 분포를 가정했을 때 추정되는 생존함수, 그리고 마지막으로 빨간색 선은 Weibull 분포를 가정했을 때 추정되는 생존함수를 나타낸다. 위의 그림을 통해 5번째 홈런까지의 타석수를 Weibull 분포를 가정했을 때 추정되는 생존함수가 Kaplan Meier estimation 추정된 생존함수와 매우 유사한 것을 확인할 수 있다.

아래 그림들은 순서대로 10개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 비모수적 방법인 Kaplan Meier estimation과 모수적 방법인 Exponentail 분포를 가정했을 때와 Weibull 분포를 가정했을 때 추정된 생존함수들을 그린 것이다.



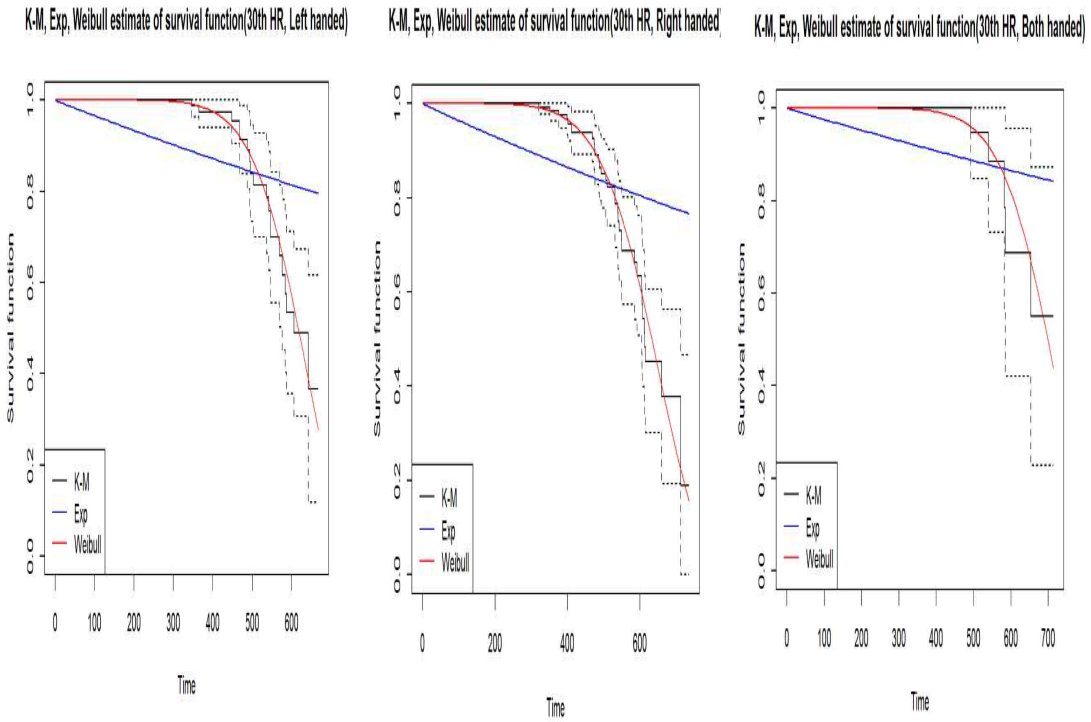
5번째와 마찬가지로 검은색 선은 10번째 홈런까지의 타석수를 Kaplan Meier estimation로 추정한 생존함수이며 파란색 선은 Expoenetial 분포를 가정했을 때 추정되는 생존함수, 그리고 마지막으로 빨간색 선은 Weibull 분포를 가정했을 때 추정되는 생존함수를 나타낸다. 위의 그림을 통해 10번째 홈런까지의 타석수 역시 5번째 홈런까지의 타석수와 마찬가지로 Weibull 분포를 가정했을 때 추정되는 생존함수가 Kaplan Meier estimation 추정된 생존함수와 매우 유사한 것을 확인할 수 있다.

아래 그림들은 순서대로 20개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 비모수적 방법인 Kaplan Meier estimation과 모수적 방법인 Exponentail 분포를 가정했을 때와 Weibull 분포를 가정했을 때 추정된 생존함수들을 그린 것이다.



20번째 홈런까지의 타석수도 마찬가지로 검은색 선은 Kaplan Meier estimation로 추정한 생존함수이며 파란색 선은 Expoenetial 분포를 가정했을 때 추정되는 생존함수, 그리고 마지막으로 빨간색 선은 Weibull 분포를 가정했을 때 추정되는 생존함수를 나타낸다. 위의 그림을 통해 20번째 홈런까지의 타석수 역시 5번째, 10번째 홈런까지의 타석의 경우와 비슷하게 Weibull 분포를 가정했을 때 추정되는 생존함수가 Kaplan Meier estimation 추정된 생존함수와 매우 유사한 것을 확인할 수 있다.

아래 그림들은 순서대로 30개 홈런을 기준으로 좌타자, 우타자, 양손타자의 생존함수를 비모수적 방법인 Kaplan Meier estimation과 모수적 방법인 Exponentail 분포를 가정했을 때와 Weibull 분포를 가정했을 때 추정된 생존함수들을 그린 것이다.



30번째 홈런까지의 타석수도 검은색 선은 마찬가지로 Kaplan Meier estimation로 추정한 생존함수이며 파란색 선은 Expoenetial 분포를 가정했을 때 추정되는 생존함수, 그리고 마지막으로 빨간색 선은 Weibull 분포를 가정했을 때 추정되는 생존함수를 나타낸다. 위의 그림을 통해 30번째 홈런까지의 타석수 역시 위의 모든 경우와 비슷하게 Weibull 분포를 가정했을 때 추정되는 생존함수가 Kaplan Meier estimation 추정된 생존함수와 매우 유사한 것을 확인할 수 있다.

2.3 Hypothesis testing

주어진 데이터를 좌타자, 우타자, 양손타자로 분류하여 각 범주별로 각각 5개, 10개, 20개, 30개 홈런까지의 생존함수 간의 유의한 차이가 있는가를 검정하였다. 이에 통계량은 다음과 같이 주어지며 카이제곱 분포를 따르게 된다.

	1	⋯	K	total
# of death	$dN_1(x)$	⋯	$dN_K(x)$	$dN(x)$
# alive	$Y_1(x)-dN_1(x)$	⋯	$Y_K(x)-dN_K(x)$	$Y(x)-dN(x)$
# at risk	$Y_1(x)$	⋯	$Y_K(x)$	$Y(x)$

$$Z=\begin{pmatrix} \sum_x [dN_1(x) - Y_1(x)\frac{dN(x)}{Y(x)}] \\ \cdots \\ \sum_x [dN_{K-1}(x) - Y_{K-1}(x)\frac{dN(x)}{Y(x)}] \end{pmatrix}$$

$$V_{jj} = \frac{Y_j(x)}{Y(x)}[1-\frac{Y_j(x)}{Y(x)}]\frac{Y(x)-dN(x)}{Y(x)-1}dN(x)$$

$$V_{jl} = -\frac{Y_j(x)Y_l(x)}{Y(x)^2}\frac{Y(x)-dN(x)}{Y(x)-1}dN(x)$$

$$Z^T V^{-1} Z \backsim \chi^2(K-1)$$

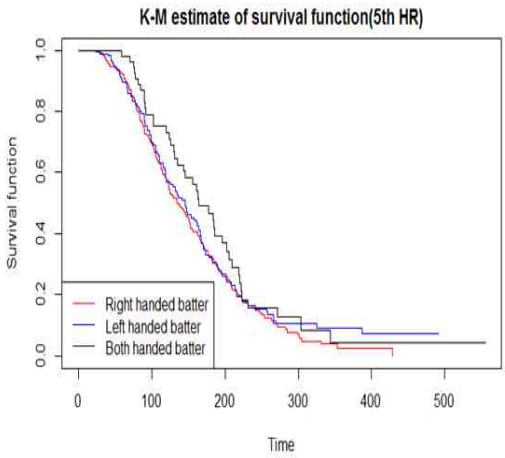
2.3.1 Log-rank test via bating hand

5번째 홈런까지의 생존함수의 경우, 좌타자, 우타자, 양손타자 간의 유의한 차이가 없다는 가정을 세웠고, 유의수준 0.05에서 p-value 값으로 0.2를 얻어 이 가설을 기각할 수 없다는 결론을 얻었다. 즉, 5번째 홈런까지의 타석수의 경우 좌·우·양손타자 간의 차이가 없었다.

```
Call:
survdiff(formula = surv_data_B_5th ~ data$B.group, rho = 0)

                N Observed Expected (O-E)^2/E (O-E)^2/V
data$B.group=0 341      196    183.3    0.8869    1.8578
data$B.group=1 196      119    121.6    0.0575    0.0886
data$B.group=2  71       41     51.1    1.9979    2.3623

 Chisq= 3  on 2 degrees of freedom, p= 0.2
Call:
```

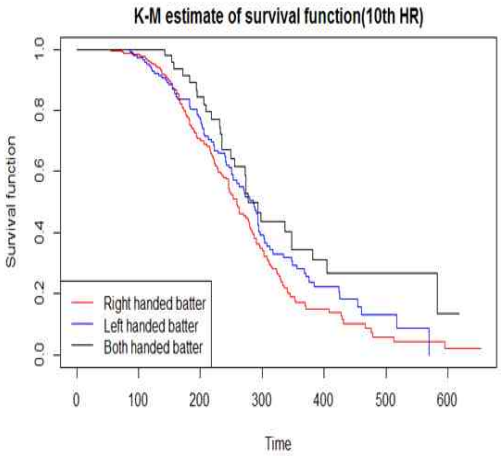


10번째 홈런까지의 생존함수의 경우, 좌타자, 우타자, 양손타자 간의 유의한 차이가 없다는 가정을 세웠고, 유의수준 0.05에서 p-value 값으로 0.02를 얻어 이 가설을 기각할 수 있다는 결론을 얻었다. 즉, 10번째 홈런까지의 타석수의 경우 좌·우·양손타자 간의 차이가 있었다.

```
Call:
survdiff(formula = surv_data_B_10th ~ data$B.group, rho = 0)

                N Observed Expected (O-E)^2/E (O-E)^2/V
data$B.group=0 341      148    128.8     2.847     5.704
data$B.group=1 196       86     92.3     0.427     0.669
data$B.group=2  71       28     40.9     4.055     4.902

 Chisq= 7.5  on 2 degrees of freedom, p= 0.02
```

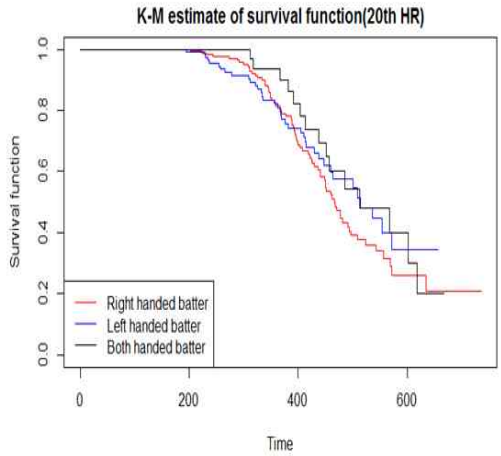


20번째 홈런까지의 생존함수의 경우, 좌타자, 우타자, 양손타자 간의 유의한 차이가 없다는 가정을 세웠고, 유의수준 0.05에서 p-value 값으로 0.5를 얻어 이 가설을 기각할 수 없다는 결론을 얻었다. 즉, 20번째 홈런까지의 타석수의 경우 좌·우·양손타자 간의 차이가 없었다.

```
Call:
survdiff(formula = surv_data_B_20th ~ data$B.group, rho = 0)

                N Observed Expected (O-E)^2/E (O-E)^2/V
data$B.group=0 341       71     64.8     0.599     1.261
data$B.group=1 196       38     40.8     0.191     0.286
data$B.group=2  71       15     18.4     0.641     0.759

 Chisq= 1.4  on 2 degrees of freedom, p= 0.5
```



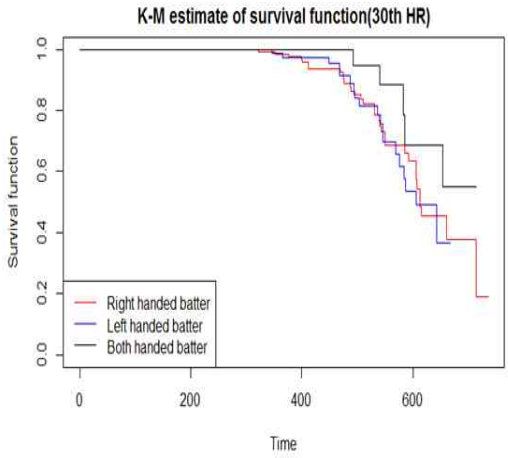
30번째 홈런까지의 생존함수의 경우, 좌타자, 우타자, 양손타자 간의 유의한 차이가 없다는 가정을 세웠고, 유의수준 0.05에서 p-value 값으로 0.3를 얻어 이 가설을 기각할 수 없다는 결론을 얻었다. 즉, 30번째 홈런까지의 타석수의 경우 좌·우·양손타자 간의 차이가 없었다.

2.4 Survival regression model

```
Call:
survdiff(formula = surv_data_B_30th ~ data$B.group, rho = 0)

                N Observed Expected (O-E)^2/E (O-E)^2/V
data$B.group=0 341       33    30.25      0.25     0.536
data$B.group=1 196       19    17.11      0.21     0.305
data$B.group=2  71        5     9.64      2.24     2.737

 Chisq= 2.7  on 2 degrees of freedom, p= 0.3
```

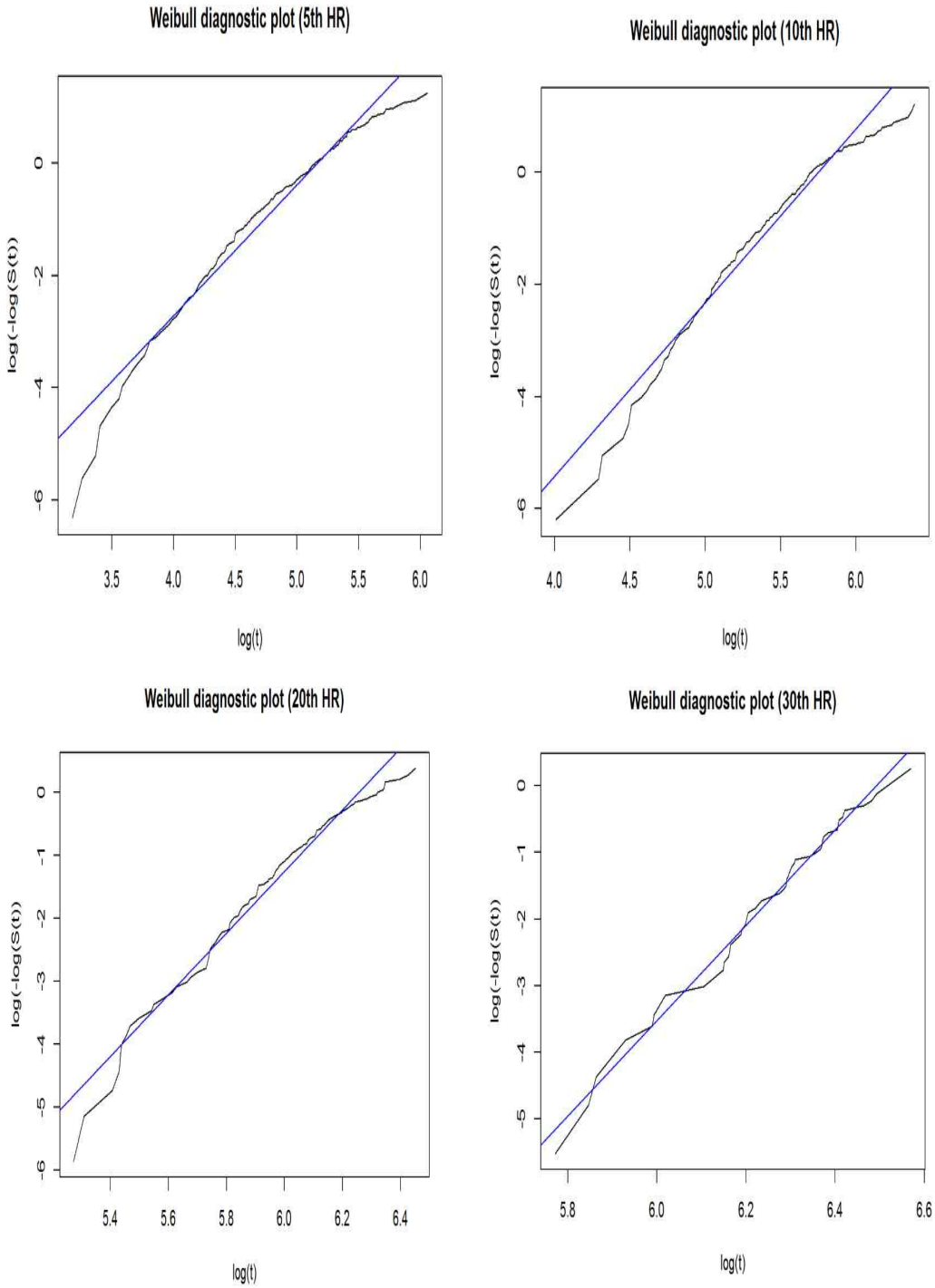


2.4.1 Weibull distribution diagnostic

Data를 Exponential 분포를 가정했을 때와 Weibull 분포를 가정했을 때의 추정된 생존함수 비교를 통하여 Weibull 분포를 가정했을 때와 Kaplan Meier estimation을 통해 추정된 생존함수가 유사하다는 것을 알 수 있었다. 이를 통해 Data가 정말로 Weibull 분포를 따르는지 Diagnostic을 하고자 한다.

Weibull 분포를 가정했을 때 Hazard function은 $h(x;\lambda,\alpha) = \alpha\lambda x^{\alpha-1}$이고, 이를 통해 Cumulative hazard function $H(x;\lambda,\alpha) = \lambda x^{\alpha}$을 얻을 수 있다. ln을 취하게 되면 $\ln(H(x)) = \ln(\lambda) + \alpha\ln(x)$ 관계식을 얻게 된다. $H(x) = -\ln(S(x))$을 관계식에 대입해주면 $\ln(-\ln(S(x))) = \ln(\lambda) + \alpha\ln(x)$ 새로운 관계식을 얻을 수 있다. 만일 데이터가 Weibull 분포를 따른다면 $\ln(-\ln(S(x)))$과 $\ln(x)$가 선형관계가 있어야 한다. 이를 위해 $\ln(-\ln(\hat{S}(x)))\ vs\ \ln(x)$의 plot을 그려본다. $\hat{S}(x)$는 Kaplan Meier estimation을 통해 추정된 생존함수이다.

아래의 그림은 순서대로 5개, 10개, 20개, 30개 홈런까지의 Data에 대해 Weibull distribution diagnostic을 진행한 결과다.



검은색 선은 X축으로 $\ln(x)$을, Y축으로 $\ln(-\ln(\hat{S}(x)))$ 두었을 때 Data들의 plot을 보여준다. 파란색 선은 Data를 통해 $\ln(-\ln(\hat{S}(x)))$을 $\ln(x)$을 변수로 Linear regression을 한 결과이다. 5개, 10개, 20개, 30개 홈런까지의 Data 모두 $\ln(-\ln(\hat{S}(x)))\ vs\ \ln(x)$ plot이 Linear regression 직선과 거의 일치하는 것을

확인할 수 있다. 따라서 Data들이 Weibull 분포를 따른다고 가정하는 것이 타당하다고 생각할 수 있다.

2.4.2 AFT model

Data의 Time이 Weibull 분포를 가정할 때 문제가 없다는 것을 알 수 있었다. 따라서 이번 절에서는 Weibull 분포를 가정하는 Parametric AFT model을 통해 Time과 Time에 영향을 주는 요인들을 분석하고자 한다. AFT model은 Time에 log 값을 취한 후 다른 요인들과의 Linear 관계를 생각한다. 따라서 다음과 같은 관계식을 얻을 수 있다. $logX = \mu + \gamma^T Z + \sigma W$ 이러한 관계식을 통해서 Time에 영향을 주는 요인들의 계수를 구할 수 있으며 계수의 p-value를 확인할 수 있다.

$S_0(x)$를 Z=0일 때의 X의 Survival function이라고 하자. AFT model에 따르면 $\Pr(X > x|Z) = S_X(x|Z) = S_0(xexp(-\gamma^T Z))$을 얻을 수 있다. 이를 통해 Hazard function $h(x|Z) = h_0(xexp(-\gamma^T Z))\exp(-\gamma^T Z)$을 얻게 된다. Weibull 분포를 가정했을 때는 $\lambda = \exp(-\frac{\mu}{\sigma})$ and $\alpha = \frac{1}{\sigma}$ 로 두면, $h(x|Z) = \alpha\lambda x^{\alpha-1}\exp(-\beta^T Z)$, where $\beta = \frac{\gamma}{\sigma}$ 이 된다. Hazard function을 통해서 Proportional hazard $\frac{h(x|Z)}{h(x|Z_0)} = \exp(-\beta^T(Z - Z_0))$ 관계식을 얻게 된다. 즉, AFT model 분석을 통해서 얻은 계수들을 Scale로 나눠준 후 얻은 값을 Exponential 함수를 취한 후 역수를 해주면 Proportional hazard 값을 얻을 수 있다.

본 프로젝트에서는 AFT model을 통해 우타자, 좌타자, 양손타자가 5개, 10개, 20개, 30개 홈런까지의 타석수에 영향을 주는지 분석을 해본 후 차이가 있는 홈런의 개수에 대해서 다른 변수들을 넣은 후 추가 분석을 진행하였다.

아래의 결과들은 5개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부만 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_5 ~ factor(B.group), data = data, dist = "weibull")
                  Value Std. Error      z      p
(Intercept)      5.1607     0.0370 139.45 <2e-16
factor(B.group)1 0.0858     0.0602   1.43  0.154
factor(B.group)2 0.1916     0.0889   2.15  0.031
Log(scale)      -0.6583     0.0375 -17.54 <2e-16

Scale= 0.518

Weibull distribution
Loglik(model)= -2127.1   Loglik(intercept only)= -2129.9
 Chisq= 5.61 on 2 degrees of freedom, p= 0.061
Number of Newton-Raphson Iterations: 8
n= 608
```

```
Call:
survreg(formula = s_5 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data, dist = "weibull")
                                    Value Std. Error      z      p
(Intercept)                        5.2466     0.0476 110.16 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0858     0.0602  -1.43   0.15
factor(B.group, levels = c(1, 0, 2))2  0.1057     0.0938   1.13   0.26
Log(scale)                        -0.6583     0.0375 -17.54 <2e-16

Scale= 0.518

Weibull distribution
Loglik(model)= -2127.1   Loglik(intercept only)= -2129.9
 Chisq= 5.61 on 2 degrees of freedom, p= 0.061
Number of Newton-Raphson Iterations: 8
n= 608
```

해당 모델의 $\chi^2 Test$ 의 p-value 0.061을 얻을 수 있다. Log rank test와 마찬가지로 5번째 홈런까지의 타석수가 우타자, 좌타자, 양손타자 여부에 따라 차이가 없다는 귀무가설을 기각할 수 없다. 5개의 홈런은 우타자, 좌타자, 양손타자 모든 그룹에서 절반 이상의 선수들이 쳤다. 즉, 5개라는 기준의 홈런은 대다수의 선수가 충분히 칠 수 있을 정도이므로 큰 차이가 없다는 생각을 할 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부만 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_10 ~ factor(B.group), data = data, dist = "weibull")
                  Value Std. Error      z      p
(Intercept)      5.7379     0.0317 181.27 <2e-16
factor(B.group)1 0.0884     0.0522   1.69  0.091
factor(B.group)2 0.2359     0.0794   2.97  0.003
Log(scale)      -0.9543     0.0431 -22.15 <2e-16

Scale= 0.385

Weibull distribution
Loglik(model)= -1692.3   Loglik(intercept only)= -1697.6
 Chisq= 10.56 on 2 degrees of freedom, p= 0.0051
Number of Newton-Raphson Iterations: 9
n= 608
```

```
Call:
survreg(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data, dist = "weibull")
                                    Value Std. Error      z      p
(Intercept)                        5.8263     0.0415 140.30 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0884     0.0522  -1.69  0.091
factor(B.group, levels = c(1, 0, 2))2  0.1475     0.0838   1.76  0.078
Log(scale)                        -0.9543     0.0431 -22.15 <2e-16

Scale= 0.385

Weibull distribution
Loglik(model)= -1692.3   Loglik(intercept only)= -1697.6
 Chisq= 10.56 on 2 degrees of freedom, p= 0.0051
Number of Newton-Raphson Iterations: 9
n= 608
```

해당 모델의 $\chi^2 Test$ 의 p-value 0.0051을 얻을 수 있다. Log rank test와 마찬가지로 10번째 홈런까지의 타석수가 우타자, 좌타자, 양손타자 여부에 따라 차이가 없다는 귀무가설을 기각할 수 있다. 변수들의 p-value를 확인해 보면 우타자와 양손타자 사이에 차이가 있지만, 우타자와 좌타자, 좌타자와 양손타자 사이에는 차이가 없음을 확인할 수 있다. 우타자와 좌타자의 경우 절반의 선수들이 10개의 홈런을 쳤지만 양손타자는 1/3 정도의 선수들만 10개의 홈런을 쳤다. 즉, 10개의 기준이 5개 기준과는 다르게 선수들에게도 달성하기 충분히 어려운 조건이므로 결과에 차이를 줬다고 생각을 해보았다. 따라서 10개의 경우 다른 변수들을 추가했을 경우 어떠한 결과를 나오는지 분석하기로 하였다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 나이를 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_10 ~ factor(B.group) + factor(A.group), data = data,
    dist = "weibull")
                  Value Std. Error     z      p
(Intercept)      5.7003     0.0722 78.91 <2e-16
factor(B.group)1 0.0900     0.0522  1.72 0.0851
factor(B.group)2 0.2415     0.0796  3.03 0.0024
factor(A.group)1 0.0338     0.0748  0.45 0.6510
factor(A.group)2 0.0531     0.0789  0.67 0.5009
Log(scale)      -0.9571     0.0433 -22.11 <2e-16

Scale= 0.384

Weibull distribution
Loglik(model)= -1692.1   Loglik(intercept only)= -1697.6
 Chisq= 11.02 on 4 degrees of freedom, p= 0.026
Number of Newton-Raphson Iterations: 11
n= 608
```

```
Call:
survreg(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(A.group), data = data, dist = "weibull")
                                    Value Std. Error     z      p
(Intercept)                        5.7903     0.0738 78.48 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0900     0.0522 -1.72  0.085
factor(B.group, levels = c(1, 0, 2))2 0.1515     0.0839  1.81  0.071
factor(A.group)1                   0.0338     0.0748  0.45  0.651
factor(A.group)2                   0.0531     0.0789  0.67  0.501
Log(scale)                        -0.9571     0.0433 -22.11 <2e-16

Scale= 0.384

Weibull distribution
Loglik(model)= -1692.1   Loglik(intercept only)= -1697.6
 Chisq= 11.02 on 4 degrees of freedom, p= 0.026
Number of Newton-Raphson Iterations: 11
n= 608
```

```
factor(B.group)1 factor(B.group)2 factor(A.group)1 factor(A.group)2
       0.7911360        0.5331948        0.9156653        0.8707791
```

```
factor(B.group, levels = c(1, 0, 2))2
                            0.6739610
```

해당 모델의 $\chi^2 Test$ 의 p-value 0.026을 얻을 수 있다. 우타자, 좌타자, 양손타자 여부에 선수들의 나이까지 고려했을 때, 우타자와 양손타자의 차이가 있다는 것을 알 수 있었다. 두 그룹의 Proportional hazard 값은 0.5331948로 같은 연령대에서 우타자가 양손타자에 비해 2배 적은 타석수에서 10개의 홈런을 친

다는 것을 알 수 있었다. 나머지 경우에는 차이가 없다는 것을 확인할 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 키를 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_10 ~ factor(B.group) + factor(H.group), data = data,
    dist = "weibull")
                  Value Std. Error     z      p
(Intercept)      6.1014     0.1264 48.27 <2e-16
factor(B.group)1 0.0591     0.0496  1.19 0.2334
factor(B.group)2 0.1989     0.0753  2.64 0.0083
factor(H.group)1 -0.3125    0.1253 -2.49 0.0126
factor(H.group)2 -0.5789    0.1296 -4.47  8e-06
Log(scale)       -1.0124    0.0435 -23.29 <2e-16

Scale= 0.363

Weibull distribution
Loglik(model)= -1674.9   Loglik(intercept only)= -1697.6
 Chisq= 45.47 on 4 degrees of freedom, p= 3.2e-09
Number of Newton-Raphson Iterations: 11
n= 608

    factor(B.group)1 factor(B.group)2 factor(H.group)1 factor(H.group)2
           0.8498132        0.5784389        2.3634412        4.9196633
```

```
Call:
survreg(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(H.group), data = data, dist = "weibull")
                                      Value Std. Error     z      p
(Intercept)                          6.1606     0.1238 49.77 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0591    0.0496 -1.19  0.233
factor(B.group, levels = c(1, 0, 2))2 0.1398     0.0791  1.77  0.077
factor(H.group)1                     -0.3125     0.1253 -2.49  0.013
factor(H.group)2                     -0.5789     0.1296 -4.47  8e-06
Log(scale)                           -1.0124     0.0435 -23.29 <2e-16

Scale= 0.363

Weibull distribution
Loglik(model)= -1674.9   Loglik(intercept only)= -1697.6
 Chisq= 45.47 on 4 degrees of freedom, p= 3.2e-09
Number of Newton-Raphson Iterations: 11
n= 608

factor(B.group, levels = c(1, 0, 2))2
                            0.680666
```

해당 모델의 $\chi^2 Test$ 의 p-value는 매우 작은 값을 가지고 있다. 우타자, 좌타자, 양손타자 여부에 선수들의 키까지 고려했을 때, 우타자와 양손타자의 차이가 있다는 것을 알 수 있었다. 이에 더불어 선수들의 키와 관련된 계수들이 모두 유의하다는 것을 알 수 있었다. 이는 선수들의 키가 10개까지 홈런을 치는 타석수에 영향을 미친다는 것을 알 수 있다. Proportional hazard를 계산해보면 키를 고려했을 경우 우타자가 양손타자에 비해 0.6배 적은 타석수만으로 10개 홈런을 칠 수 있었다. 10개까지 홈런에서 가장 작은 키 그룹에 속한 선수들에 비해 중간 키 그룹에 속한 선수들이 2.4배 빠른 타석수에서 10개의 홈런을 칠 수 있으며 가장 큰 키 그룹에 속한 선수들은 5배 빠른 타석수에서 10개의 홈런을 칠 수 있다. 키가 홈런을 잘 치는 것에 직접적인 영향을 주기보다는 키가 몸무게에 영향을 줄 수 있으므로 결과가 이렇게 나왔다고 생각해보았다. 키가 클

수록 몸무게가 늘어나게 되고 그로 인해 선수들의 힘이 세지므로 홈런을 더욱 잘 칠 수 있다는 것을 확인할 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 몸무게를 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_10 ~ factor(B.group) + factor(W.group), data = data,
    dist = "weibull")
                  Value Std. Error     z       p
(Intercept)      6.0535     0.0709 85.38 < 2e-16
factor(B.group)1 0.0591     0.0493  1.20   0.231
factor(B.group)2 0.1366     0.0760  1.80   0.072
factor(W.group)1 -0.3113    0.0715 -4.36 1.3e-05
factor(W.group)2 -0.5305    0.0810 -6.55 5.7e-11
Log(scale)       -1.0154    0.0433 -23.47 < 2e-16

Scale= 0.362

Weibull distribution
Loglik(model)= -1669.4   Loglik(intercept only)= -1697.6
 Chisq= 56.47 on 4 degrees of freedom, p= 1.6e-11
Number of Newton-Raphson Iterations: 11
n= 608

factor(B.group)1 factor(B.group)2 factor(W.group)1 factor(W.group)2
       0.8494243        0.6858198        2.3617576        4.3257474
```

```
Call:
survreg(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(W.group), data = data, dist = "weibull")
                                    Value Std. Error     z       p
(Intercept)                        6.1126     0.0720 84.94 < 2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0591     0.0493 -1.20    0.23
factor(B.group, levels = c(1, 0, 2))2 0.0775     0.0797  0.97    0.33
factor(W.group)1                   -0.3113     0.0715 -4.36 1.3e-05
factor(W.group)2                   -0.5305     0.0810 -6.55 5.7e-11
Log(scale)                         -1.0154     0.0433 -23.47 < 2e-16

Scale= 0.362

Weibull distribution
Loglik(model)= -1669.4   Loglik(intercept only)= -1697.6
 Chisq= 56.47 on 4 degrees of freedom, p= 1.6e-11
Number of Newton-Raphson Iterations: 11
n= 608

factor(B.group, levels = c(1, 0, 2))2
                           0.8073937
```

해당 모델의 $\chi^2 Test$ 의 p-value는 매우 작은 값을 가지고 있다. 우타자, 좌타자, 양손타자 여부에 선수들의 몸무게까지 고려했을 때, 우타자, 좌타자, 양손타자 여부는 10개 홈런까지 타석수에 영향을 미치지 않는 것을 확인할 수 있다. 하지만 선수들의 몸무게 관련된 계수들이 모두 유의한 것을 알 수 있었다. 이는 선수들의 몸무게가 10개까지 홈런을 치는 타석수에 영향을 미친다는 것을 알 수 있다. Proportional hazard를 계산해보면 10개까지 홈런에서 가장 가벼운 몸무게 그룹에 속한 선수들에 비해 중간 몸무게 그룹에 속한 선수들이 2.4배 빠른 타석수에서 10개의 홈런을 칠 수 있으며 가장 무거운 그룹에 속한 선수들은 4.3배 빠른 타석수에서 10개의 홈런을 칠 수 있다. 선수들이 무거울수록 10개까지 홈런을 빨리 친다는 것을 알 수 있었다. 이러한 결과는 키에서 추론한 결과와 일맥상통한다. 선수들의 몸무게가 커질수록 선수들의 힘이 세지므로 홈런을 더욱 잘 칠 수 있다고 생각을 할 수 있다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 리그를 고려했을 때의 AFT model을 통해 분석한 결과이다

```
Call:
survreg(formula = s_10 ~ factor(B.group) + factor(League), data = data,
    dist = "weibull")
                  Value Std. Error     z      p
(Intercept)      5.7734     0.0400 144.22 <2e-16
factor(B.group)1 0.0851     0.0521  1.63 0.1023
factor(B.group)2 0.2483     0.0795  3.12 0.0018
factor(League)1 -0.0721     0.0478 -1.51 0.1316
Log(scale)      -0.9578     0.0430 -22.27 <2e-16

Scale= 0.384

Weibull distribution
Loglik(model)= -1691.2   Loglik(intercept only)= -1697.6
 Chisq= 12.83 on 3 degrees of freedom, p= 0.005
Number of Newton-Raphson Iterations: 11
n= 608

factor(B.group)1 factor(B.group)2  factor(League)1
       0.8011143        0.5235085        1.2066777
```

```
Call:
survreg(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(League), data = data, dist = "weibull")
                                    Value Std. Error     z      p
(Intercept)                        5.8585     0.0471 124.37 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.0851     0.0521 -1.63  0.102
factor(B.group, levels = c(1, 0, 2))2  0.1633     0.0841  1.94  0.052
factor(League)1                   -0.0721     0.0478 -1.51  0.132
Log(scale)                        -0.9578     0.0430 -22.27 <2e-16

Scale= 0.384

Weibull distribution
Loglik(model)= -1691.2   Loglik(intercept only)= -1697.6
 Chisq= 12.83 on 3 degrees of freedom, p= 0.005
Number of Newton-Raphson Iterations: 11
n= 608

factor(B.group, levels = c(1, 0, 2))2
                            0.6534754
```

해당 모델의 $\chi^2 Test$ 의 p-value는 0.005를 가지고 있다. 우타자, 좌타자, 양손타자 여부에 선수들의 몸무게까지 고려했을 때, 우타자와 양손타자, 좌타자와 양손타자가 10개 홈런까지 타석수에 차이를 확인할 수 있다. 하지만 리그 차이는 10개 홈런까지 타석수에 영향을 미치지 않았다. 우타자, 좌타자, 양손타자 여부와 선수들의 리그를 고려하였을 때 같은 리그에서 우타자가 양손타자에 비해 2배 적은 타석수에서 10개의 홈런을 칠 수 있다는 알 수 있었다. 좌타자는 양손타자에 비해 1.6배 적은 타석수에서 10개의 홈런을 칠 수 있다는 사실을 알 수 있었다.

아래의 결과들은 20개, 30개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부만 고려했을 때의 AFT model을 통해 분석한 결과이다.

```
Call:
survreg(formula = s_20 ~ factor(B.group) + factor(League), data = data,
    dist = "weibull")
                  Value Std. Error       z       p
(Intercept)     5.75307    0.00273 2107.55 < 2e-16
factor(B.group)1 0.07562    0.00499   15.17 < 2e-16
factor(B.group)2 0.02577    0.00555    4.64 3.5e-06
factor(League)1  0.15833    0.00483   32.78 < 2e-16
Log(scale)      -2.02591    0.00000    -Inf < 2e-16

Scale= 0.132

Weibull distribution
Loglik(model)= -4825.9   Loglik(intercept only)= -872.4
 Chisq= -7906.99 on 3 degrees of freedom, p= 1
Number of Newton-Raphson Iterations: 30
n= 608
```

```
Call:
survreg(formula = s_30 ~ factor(B.group), data = data, dist = "weibull")
                  Value Std. Error      z      p
(Intercept)     5.354343   0.000995 5378.8 <2e-16
factor(B.group)1 0.102046  0.001616   63.2 <2e-16
factor(B.group)2 0.090515  0.002346   13.0 <2e-16
Log(scale)      -2.088761  0.000000   -Inf <2e-16

Scale= 0.124

Weibull distribution
Loglik(model)= -52738.4   Loglik(intercept only)= -408.7
 Chisq= -104659.3 on 2 degrees of freedom, p= 1
Number of Newton-Raphson Iterations: 30
n= 608
```

```
Call:
survreg(formula = s_20 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(League), data = data, dist = "weibull")
                                      Value Std. Error       z       p
(Intercept)                         5.82869    0.00439 1328.51 < 2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.07562    0.00499  -15.17 < 2e-16
factor(B.group, levels = c(1, 0, 2))2 -0.04985    0.00657   -7.59 3.3e-14
factor(League)1                     0.15833    0.00483   32.78 < 2e-16
Log(scale)                         -2.02591    0.00000    -Inf < 2e-16

Scale= 0.132

Weibull distribution
Loglik(model)= -4825.9   Loglik(intercept only)= -872.4
 Chisq= -7906.99 on 3 degrees of freedom, p= 1
Number of Newton-Raphson Iterations: 30
n= 608
```

```
Call:
survreg(formula = s_30 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data, dist = "weibull")
                                      Value Std. Error      z      p
(Intercept)                         5.45639    0.00127 4287.1 <2e-16
factor(B.group, levels = c(1, 0, 2))0 -0.10205    0.00162  -63.2 <2e-16
factor(B.group, levels = c(1, 0, 2))2 -0.07153    0.00248  -28.9 <2e-16
Log(scale)                         -2.08876    0.00000   -Inf <2e-16

Scale= 0.124

Weibull distribution
Loglik(model)= -52738.4   Loglik(intercept only)= -408.7
 Chisq= -104659.3 on 2 degrees of freedom, p= 1
Number of Newton-Raphson Iterations: 30
n= 608
```

두 모델의 $\chi^2 Test$ 의 p-value는 1를 가지고 있다. 이는 우타자, 좌타자, 양손타자의 20개, 30개까지 홈런의 타석수에 차이가 없는 것을 의미한다. 즉, log rank test와 같은 결과를 보여준다. 20개의 홈런부터는 모든 선수에게 달성하기 어려운 기록이므로 타석수까지의 차이가 없는 것을 생각할 수 있었다.

2.5.3 Cox-PH model

Cox-PH model 역시 AFT model과 마찬가지로 Data의 Time에 분포 가정을 한 후 분석을 진행할 수 있다. AFT model에서 가정한 Weibull 분포를 가정한 후 Cox-PH model 분석을 진행하였다. Cox-PH model에서는 Time에 영향을 주는 변수 Z에 대해 Hazard function의 모양을 가정한다고 생각할 수 있다.

$$h(x|Z) = h_0(x)exp(\beta^T Z)$$

Weibull 분포를 가정하였을 때 Cox-PH model을 통해서 Time에 영향을 주는 변수 Z의 변화에 따른 Proportional hazard를 구할 수 있다.

$$\frac{h(x|Z)}{h(x|Z_0)} = \frac{h_0(x)\exp(\beta^T Z)}{h_0(x)\exp(\beta^T Z_0)} = \exp[\beta^T(Z - Z_0)]$$

AFT model에서는 해당 변수의 계수에 Exponential을 취한 후 역수를 생각하였다면 Cox-PH model에서는 해당 변수의 계수에 Exponential만 취해주면 Proportional hazard를 구할 수 있다.

Cox-PH model에서 계수를 추정할 때 Partial Likelihood를 이용한다. Data의 Tie를 처리해주는 방법이 3가지가 있는데 방법마다 Partial Likelihood가 다르다. 본 프로젝트에서는 Tie에 크게 영향을 받지 않고 R에서 기본으로 지정되어 있는 Efron의 방법을 이용하여 분석을 진행하였다.

아래의 결과들은 5개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_5 ~ factor(B.group), data = data)

  n= 608, number of events= 356

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.09006   0.91388  0.11664 -0.772   0.4401
factor(B.group)2 -0.28910   0.74894  0.17206 -1.680   0.0929 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.9139      1.094    0.7271     1.149
factor(B.group)2    0.7489      1.335    0.5345     1.049

Concordance= 0.527  (se = 0.015 )
Likelihood ratio test= 3.09  on 2 df,   p=0.2
Wald test            = 2.95  on 2 df,   p=0.2
Score (logrank) test = 2.96  on 2 df,   p=0.2
```

```
Call:
coxph(formula = s_5 ~ factor(B.group, levels = c(1, 0, 2)), data = data)

  n= 608, number of events= 356

                                      coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0 0.09006   1.09424  0.11664  0.772
factor(B.group, levels = c(1, 0, 2))2 -0.19904   0.81951  0.18134 -1.098
                                     Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.440
factor(B.group, levels = c(1, 0, 2))2    0.272

                                     exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    1.0942     0.9139    0.8706
factor(B.group, levels = c(1, 0, 2))2    0.8195     1.2202    0.5744
                                     upper .95
factor(B.group, levels = c(1, 0, 2))0     1.375
factor(B.group, levels = c(1, 0, 2))2     1.169

Concordance= 0.527  (se = 0.015 )
Likelihood ratio test= 3.09  on 2 df,   p=0.2
Wald test            = 2.95  on 2 df,   p=0.2
Score (logrank) test = 2.96  on 2 df,   p=0.2
```

해당 모델의 모든 Test p-value는 0.2를 가지고 있다. 이는 우타자, 좌타자, 양손타자의 5개까지 홈런의 타석수에 차이가 없는 것을 의미한다. 즉, log rank test와 AFT model을 이용한 분석과 같은 결과를 보여준다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_10 ~ factor(B.group), data = data)

  n= 608, number of events= 262

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.2123    0.8087   0.1363 -1.558   0.1193
factor(B.group)2 -0.5257    0.5912   0.2075 -2.533   0.0113 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8087      1.237    0.6191    1.0564
factor(B.group)2    0.5912      1.692    0.3936    0.8878

Concordance= 0.541  (se = 0.018 )
Likelihood ratio test= 7.86  on 2 df,   p=0.02
Wald test            = 7.36  on 2 df,   p=0.03
Score (logrank) test = 7.47  on 2 df,   p=0.02
```

```
Call:
coxph(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data)

  n= 608, number of events= 262

                                      coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0 0.2123    1.2365   0.1363  1.558
factor(B.group, levels = c(1, 0, 2))2 -0.3134    0.7310   0.2187 -1.433
                                     Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.119
factor(B.group, levels = c(1, 0, 2))2    0.152

                                     exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0     1.237     0.8087    0.9466
factor(B.group, levels = c(1, 0, 2))2     0.731     1.3680    0.4762
                                     upper .95
factor(B.group, levels = c(1, 0, 2))0     1.615
factor(B.group, levels = c(1, 0, 2))2     1.122

Concordance= 0.541  (se = 0.018 )
Likelihood ratio test= 7.86  on 2 df,   p=0.02
Wald test            = 7.36  on 2 df,   p=0.03
Score (logrank) test = 7.47  on 2 df,   p=0.02
```

해당 모델의 LRT Test p-value는 0.02, Wald Test p-value는 0.03 Score Test의 p-value는 0.02를 가지고 있다. 해당 결과에서 우타자와 양손타자의 10개 홈런까지의 타석수까지는 차이가 있으나 좌타자와 우타자, 좌타자와 양손타자의 차이는 없다는 결과를 확인할 수 있었다. 이러한 결과를 바탕으로 10개의 홈런까지 타석수에 대해 다른 변수들을 추가하여 추가 분석을 진행해보았다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 나이를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_10 ~ factor(B.group) + factor(A.group), data = data)

  n= 608, number of events= 262

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.20912   0.81129  0.13684 -1.528   0.1265
factor(B.group)2 -0.52921   0.58907  0.20868 -2.536   0.0112 *
factor(A.group)1  0.02853   1.02895  0.19542  0.146   0.8839
factor(A.group)2 -0.02767   0.97271  0.20606 -0.134   0.8932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8113     1.2326    0.6204    1.0609
factor(B.group)2    0.5891     1.6976    0.3913    0.8867
factor(A.group)1    1.0289     0.9719    0.7015    1.5092
factor(A.group)2    0.9727     1.0281    0.6495    1.4567

Concordance= 0.546  (se = 0.019 )
Likelihood ratio test= 8.03  on 4 df,   p=0.09
Wald test            = 7.52  on 4 df,   p=0.1
Score (logrank) test = 7.64  on 4 df,   p=0.1
```

```
Call:
coxph(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(A.group), data = data)

  n= 608, number of events= 262

                                      coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0  0.20912   1.23260  0.13684  1.528
factor(B.group, levels = c(1, 0, 2))2 -0.32008   0.72609  0.21954 -1.458
factor(A.group)1                       0.02853   1.02895  0.19542  0.146
factor(A.group)2                      -0.02767   0.97271  0.20606 -0.134
                                      Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.126
factor(B.group, levels = c(1, 0, 2))2    0.145
factor(A.group)1                         0.884
factor(A.group)2                         0.893

                                      exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    1.2326     0.8113    0.9426
factor(B.group, levels = c(1, 0, 2))2    0.7261     1.3772    0.4722
factor(A.group)1                         1.0289     0.9719    0.7015
factor(A.group)2                         0.9727     1.0281    0.6495
                                      upper .95
factor(B.group, levels = c(1, 0, 2))0     1.612
factor(B.group, levels = c(1, 0, 2))2     1.117
factor(A.group)1                          1.509
factor(A.group)2                          1.457

Concordance= 0.546  (se = 0.019 )
Likelihood ratio test= 8.03  on 4 df,   p=0.09
Wald test            = 7.52  on 4 df,   p=0.1
Score (logrank) test = 7.64  on 4 df,   p=0.1
```

해당 결과에서 우타자와 양손타자의 10개 홈런까지의 타석수까지는 차이가 있으나 좌타자와 우타자, 좌타자와 양손타자의 차이는 없다는 결과를 확인할 수 있었다. 우타자의 경우 양손타자보다 10개 홈런까지의 타석수가 약 2배 적게 걸린다는 것을 알 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 키를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_10 ~ factor(B.group) + factor(H.group), data = data)

  n= 608, number of events= 262

                   coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.1735    0.8407   0.1368 -1.269 0.204587
factor(B.group)2 -0.4965    0.6087   0.2084 -2.382 0.017222 *
factor(H.group)1  0.6849    1.9837   0.3459  1.980 0.047688 *
factor(H.group)2  1.3322    3.7895   0.3607  3.694 0.000221 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8407     1.1894    0.6431    1.0992
factor(B.group)2    0.6087     1.6429    0.4046    0.9158
factor(H.group)1    1.9837     0.5041    1.0070    3.9076
factor(H.group)2    3.7895     0.2639    1.8689    7.6839

Concordance= 0.596  (se = 0.018 )
Likelihood ratio test= 33.85  on 4 df,   p=8e-07
Wald test            = 33.39  on 4 df,   p=1e-06
Score (logrank) test = 34.93  on 4 df,   p=5e-07
```

```
Call:
coxph(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(H.group), data = data)

  n= 608, number of events= 262

                                    coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0  0.1735    1.1894   0.1368  1.269
factor(B.group, levels = c(1, 0, 2))2 -0.3230    0.7240   0.2189 -1.476
factor(H.group)1                       0.6849    1.9837   0.3459  1.980
factor(H.group)2                       1.3322    3.7895   0.3607  3.694
                                   Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0 0.204587
factor(B.group, levels = c(1, 0, 2))2 0.140013
factor(H.group)1                      0.047688 *
factor(H.group)2                      0.000221 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                    exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0     1.189     0.8407    0.9098
factor(B.group, levels = c(1, 0, 2))2     0.724     1.3812    0.4715
factor(H.group)1                          1.984     0.5041    1.0070
factor(H.group)2                          3.790     0.2639    1.8689
                                   upper .95
factor(B.group, levels = c(1, 0, 2))0     1.555
factor(B.group, levels = c(1, 0, 2))2     1.112
factor(H.group)1                          3.908
factor(H.group)2                          7.684

Concordance= 0.596  (se = 0.018 )
Likelihood ratio test= 33.85  on 4 df,   p=8e-07
Wald test            = 33.39  on 4 df,   p=1e-06
Score (logrank) test = 34.93  on 4 df,   p=5e-07
```

해당 결과에서 키를 추가 고려하였을 때 우타자와 양손타자의 10개 홈런까지의 타석수가 우타자가 양손타자의 0.6배 즉, 약 1.6배 적은 타석수로 10개 홈런을 칠 수 있다는 결과가 나왔다. 키가 가장 작은 그룹에 비해 중간 그룹은 약 2배 적은 타석수로 10개의 홈런을 칠 수 있었으며 키가 가장 큰 그룹은 약 3.7배 정도 적은 타석수로 10개의 홈런을 칠 수 있다는 결과를 확인할 수 있었다. 이는 AFT model과 유사한 결과이며 키가 직접적인 영향을 줬다기 보다는 키가 클수록 선수들의 체중도 늘어 힘이 세지게 되며, 그로 인해 적은 타석수로 홈런을 칠 수 있었다는 결과를 생각해볼 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 몸무게를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_10 ~ factor(B.group) + factor(W.group), data = data)

  n= 608, number of events= 262

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.1556    0.8559   0.1366 -1.139 0.254673
factor(B.group)2 -0.3246    0.7228   0.2103 -1.544 0.122603
factor(W.group)1  0.6914    1.9966   0.2014  3.434 0.000595 ***
factor(W.group)2  1.3284    3.7750   0.2284  5.815 6.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8559     1.1684    0.6548     1.119
factor(B.group)2    0.7228     1.3835    0.4787     1.091
factor(W.group)1    1.9966     0.5009    1.3455     2.963
factor(W.group)2    3.7750     0.2649    2.4125     5.907

Concordance= 0.614  (se = 0.019 )
Likelihood ratio test= 44.8  on 4 df,   p=4e-09
Wald test            = 43.74  on 4 df,   p=7e-09
Score (logrank) test = 46.84  on 4 df,   p=2e-09
```

```
Call:
coxph(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(W.group), data = data)

  n= 608, number of events= 262

                                      coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0 0.1556    1.1684   0.1366  1.139
factor(B.group, levels = c(1, 0, 2))2 -0.1690    0.8445   0.2206 -0.766
factor(W.group)1                       0.6914    1.9966   0.2014  3.434
factor(W.group)2                       1.3284    3.7750   0.2284  5.815
                                      Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0 0.254673
factor(B.group, levels = c(1, 0, 2))2 0.443532
factor(W.group)1                      0.000595 ***
factor(W.group)2                      6.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                      exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    1.1684     0.8559    0.8939
factor(B.group, levels = c(1, 0, 2))2    0.8445     1.1841    0.5481
factor(W.group)1                         1.9966     0.5009    1.3455
factor(W.group)2                         3.7750     0.2649    2.4125
                                      upper .95
factor(B.group, levels = c(1, 0, 2))0     1.527
factor(B.group, levels = c(1, 0, 2))2     1.301
factor(W.group)1                          2.963
factor(W.group)2                          5.907

Concordance= 0.614  (se = 0.019 )
Likelihood ratio test= 44.8  on 4 df,   p=4e-09
Wald test            = 43.74  on 4 df,   p=7e-09
Score (logrank) test = 46.84  on 4 df,   p=2e-09
```

해당 결과에서 몸무게를 추가 고려하였을 때 몸무게의 그룹 차이에 대한 계수들만 유의한 결과를 얻을 수 있었다. 몸무게가 가장 가벼운 그룹에 비해 중간 그룹은 약 2배 정도 적은 타석수로 10개의 홈런을 칠 수 있었으며 가장 무거운 그룹은 약 3.7배 정도 적은 타석수로 10개의 홈런을 칠 수 있었다. 이는 선수들의 체중이 선수들의 힘에 영향을 미치게 되고 그로 인해 홈런을 칠 수 있는 확률을 높인다는 것을 생각해볼 수 있었다.

아래의 결과들은 10개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부와 리그를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_10 ~ factor(B.group) + factor(League), data = data)

  n= 608, number of events= 262

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.2087    0.8116   0.1363 -1.531  0.12577
factor(B.group)2 -0.5586    0.5720   0.2087 -2.677  0.00743 **
factor(League)1   0.1902    1.2095   0.1248  1.525  0.12731
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8116     1.2321    0.6213    1.0602
factor(B.group)2    0.5720     1.7483    0.3800    0.8611
factor(League)1     1.2095     0.8268    0.9472    1.5446

Concordance= 0.555  (se = 0.019 )
Likelihood ratio test= 10.18  on 3 df,   p=0.02
Wald test            = 9.68  on 3 df,   p=0.02
Score (logrank) test = 9.79  on 3 df,   p=0.02
```

```
Call:
coxph(formula = s_10 ~ factor(B.group, levels = c(1, 0, 2)) +
    factor(League), data = data)

  n= 608, number of events= 262

                                        coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0  0.2087    1.2321   0.1363  1.531
factor(B.group, levels = c(1, 0, 2))2 -0.3499    0.7047   0.2200 -1.590
factor(League)1                        0.1902    1.2095   0.1248  1.525
                                      Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.126
factor(B.group, levels = c(1, 0, 2))2    0.112
factor(League)1                          0.127

                                      exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    1.2321     0.8116    0.9492
factor(B.group, levels = c(1, 0, 2))2    0.7047     1.4190    0.4578
factor(League)1                          1.2095     0.8268    0.9472
                                      upper .95
factor(B.group, levels = c(1, 0, 2))0     1.609
factor(B.group, levels = c(1, 0, 2))2     1.085
factor(League)1                           1.545

Concordance= 0.555  (se = 0.019 )
Likelihood ratio test= 10.18  on 3 df,   p=0.02
Wald test            = 9.68  on 3 df,   p=0.02
Score (logrank) test = 9.79  on 3 df,   p=0.02
```

해당 결과에서 리그를 추가 고려하였을 때 우타자와 양손타자 여부만이 10개 홈런까지 타석수에 영향을 미친다는 것을 확인할 수 있었다. 우타자는 양손타자에 비해 약 2배 적은 타석수로 10개까지 홈런을 달성할 수 있었다. 선수들의 리그는 10개의 홈런까지 타석수에 미치는 영향이 없었다.

아래의 결과들은 20개, 30개 홈런까지의 타석수에 대해 우타자, 좌타자, 양손타자 여부를 고려했을 때의 Cox-PH model을 통해 분석한 결과이다.

```
Call:
coxph(formula = s_20 ~ factor(B.group), data = data)

  n= 608, number of events= 124

                    coef exp(coef) se(coef)      z Pr(>|z|)
factor(B.group)1 -0.1635    0.8492   0.2013 -0.812    0.417
factor(B.group)2 -0.3001    0.7408   0.2848 -1.054    0.292

                exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1    0.8492     1.178    0.5724     1.260
factor(B.group)2    0.7408     1.350    0.4239     1.294

Concordance= 0.524  (se = 0.025 )
Likelihood ratio test= 1.47  on 2 df,   p=0.5
Wald test            = 1.44  on 2 df,   p=0.5
Score (logrank) test = 1.44  on 2 df,   p=0.5
```

```
Call:
coxph(formula = s_20 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data)

  n= 608, number of events= 124

                                        coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0  0.1635    1.1776   0.2013  0.812
factor(B.group, levels = c(1, 0, 2))2 -0.1366    0.8724   0.3056 -0.447
                                      Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.417
factor(B.group, levels = c(1, 0, 2))2    0.655

                                      exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    1.1776     0.8492    0.7937
factor(B.group, levels = c(1, 0, 2))2    0.8724     1.1463    0.4793
                                      upper .95
factor(B.group, levels = c(1, 0, 2))0     1.747
factor(B.group, levels = c(1, 0, 2))2     1.588

Concordance= 0.524  (se = 0.025 )
Likelihood ratio test= 1.47  on 2 df,   p=0.5
Wald test            = 1.44  on 2 df,   p=0.5
Score (logrank) test = 1.44  on 2 df,   p=0.5
```

```
Call:
coxph(formula = s_30 ~ factor(B.group), data = data)

  n= 608, number of events= 57

                    coef exp(coef) se(coef)     z Pr(>|z|)
factor(B.group)1  0.02313   1.02340  0.28976  0.08    0.936
factor(B.group)2 -0.75043   0.47217  0.48116 -1.56    0.119

                  exp(coef) exp(-coef) lower .95 upper .95
factor(B.group)1     1.0234     0.9771    0.5800     1.806
factor(B.group)2     0.4722     2.1179    0.1839     1.212

Concordance= 0.541  (se = 0.036 )
Likelihood ratio test= 3.2  on 2 df,   p=0.2
Wald test            = 2.61  on 2 df,   p=0.3
Score (logrank) test = 2.73  on 2 df,   p=0.3
```

```
Call:
coxph(formula = s_30 ~ factor(B.group, levels = c(1, 0, 2)),
    data = data)

  n= 608, number of events= 57

                                        coef exp(coef) se(coef)      z
factor(B.group, levels = c(1, 0, 2))0 -0.02313   0.97713  0.28976 -0.080
factor(B.group, levels = c(1, 0, 2))2 -0.77356   0.46137  0.50619 -1.528
                                      Pr(>|z|)
factor(B.group, levels = c(1, 0, 2))0    0.936
factor(B.group, levels = c(1, 0, 2))2    0.126

                                      exp(coef) exp(-coef) lower .95
factor(B.group, levels = c(1, 0, 2))0    0.9771      1.023    0.5537
factor(B.group, levels = c(1, 0, 2))2    0.4614      2.167    0.1711
                                      upper .95
factor(B.group, levels = c(1, 0, 2))0     1.724
factor(B.group, levels = c(1, 0, 2))2     1.244

Concordance= 0.541  (se = 0.036 )
Likelihood ratio test= 3.2  on 2 df,   p=0.2
Wald test            = 2.61  on 2 df,   p=0.3
Score (logrank) test = 2.73  on 2 df,   p=0.3
```

해당 결과에서 20개, 30개의 홈런까지의 타석수에서 우타자, 좌타자, 양손타자의 차이가 없다는 것을 확인할 수 있었다.

3. Conclusion & Discussion

Log-rank test, AFT model, Cox-PH mode의 결과에서 모두 좌·우·양손 타자에 따른 5, 20, 30번째 홈런까지의 타석수에 대한 생존 함수에 유의한 차이가 없었다. 하지만 좌·우·양손 타자에 따른 10번째 홈런까지의 생존 함수에 유의한 차이가 있다는 것을 확인할 수 있었다. 특히나 AFT model과 Cox-PH model의 경우 coefficients의 값과 유의성 여부를 통해 좌·우·양손 타자의 생존함수 간의 차이가 어느 정도인지를 알 수 있었고, 추가적 변수(나이, 키, 몸무게, 리그)의 영향에 대해서도 확인할 수 있었다. 아래 표에 분석결과를 정리해 놓았다.

	AFT Model	Cox-PH Model
Bat + Age	1. 우타자- 양손타자 차이O -> 우타자가 2배 빠르게 10 홈런 달성 2. 나이 그룹별 차이 X	1. 우타자- 양손타자 차이O -> 우타자가 2배 빠르게 10 홈런 달성 2. 나이 그룹별 차이 X
Bat + Height	1. 우타자- 양손타자 차이O -> 우타자가 1.6배 빠르게 10홈런 달성 2. 키 그룹별 차이 O ->작은 그룹에 비해 중간그룹이 2.4배, 큰 그룹이 5배 빠르게 10홈런 달성	. 우타자- 양손타자 차이O -> 우타자가 1.6배 빠르게 10홈런 달성 2. 키 그룹별 차이 O ->작은 그룹에 비해 중간 그룹이 2배, 큰 그룹이 3.7배 빠르게 10홈런 달성
Bat + Weight	1.우타자- 양손타자 차이X 2. 몸무게 그룹별 차이 O ->가벼운 그룹에 비해 중간 그룹이 2.4배, 무거운 그룹이 4.3배 빠르게 10홈런 달성	1.우타자- 양손타자 차이X 2. 몸무게 그룹별 차이 O ->가벼운 그룹에 비해 중간 그룹이 2배, 무거운 그룹이 3.7배 빠르게 10홈런 달성
Bat + League	1. 우타자- 양손타자 차이O -> 우타자가 2배 빠르게 10 홈런 달성 2. 리그별 차이X	우타자- 양손타자 차이O -> 우타자가 2배 빠르게 10 홈런 달성 2. 리그별 차이X

Reference

[1] Myunghee Cho Paik. (2017). Survival Analysis Lecture Notes.

[2] John P. Klein. (2006). Survival Analysis: Techniques for Censored and Truncated Data, Second Edition.