A High Scoring Introduction to Survival Analysis Models

Rod Sturdivant and Mike Huber

BASUG, June 2015

THE OHIO STATE UNIVERSITY
COLLEGE OF PUBLIC HEALTH

The Lineup

- Why Study 20+ Runs?
- History
- Frequency Distribution
- An Exponential Model
  - Breakdown by Baseball Eras
- A Survival Analysis (Regression Model) Approach
- Some model predictions
- Questions

Yankees surrender 14 in second…

Indians 22, Yankees 4

Recap | Box Score | Play-By-Play | Conversation

	1	2	3	4	5	6	7	8	9	R	H	E	
CLE (4-8)	0	14	1	1	4	0	0	1	1	**22**	**25**	**1**	**Final**
NYY (6-6)	2	0	0	0	0	2	0	0	0	**4**	**7**	**1**	

Yankees surrender 14 in second, suffer lopsided loss to Indians

Associated Press

NEW YORK -- The jokes started flying as the Cleveland Indians piled up runs in the second inning Saturday. Ben Francisco was glad he wasn't in the field. Ryan Garko was hoping to avoid making another out.

It was one fun day for the Tribe at the New York Yankees' swanky new home.

Asdrubal Cabrera hit a grand slam and an RBI single in Cleveland's 14-run second -- the biggest inning ever against New York -- and the Indians set the bar for Yankee Stadium's new record book, coasting to a 22-4 victory.

"It was just

Cleveland Rocked

Cleveland teams have been especially unkind to the Yankees throughout history. In fact, the four highest single-game run totals against the franchise have all been posted by teams from the Rock 'n' Roll city. Take a look:

Most Runs Allowed, Yankees Franchise History

Runs	Opponent	Date
24	at Cleveland Indians	7/29/28
23	at Cleveland Indians	9/2/02
22	vs. Cleveland Indians	4/18/09
22	vs. Cleveland Indians	7/19/87

ALSO SEE

- New Yankee Stadium a home run nightmare
- Indians record biggest inning ever vs. Yankees
- Yankees' Girardi being cautious with C Posada
- Indians' Dellucci to begin rehab stint on Monday

WERE YOU THERE?

PASSPORT

Did you attend this game? If so, start chronicling your sports memories today with ESPN's Sports Passport. Enter the games you attend, upload your photos and share your memories!
I was there »

REGULAR SEASON SERIES

New York leads 5-3 (as of Sat 4/18)

Thu 4/16	CLE 10, @NYY 2	Recap
Fri 4/17	@NYY 6, CLE 5	Recap
>Sat 4/18	**CLE 22, @NYY 4**	Box Score
Sun 4/19	@NYY 7, CLE 3	Recap
Fri 5/29	NYY 3, @CLE 1	Recap
Sat 5/30	NYY 10, @CLE 5	Recap
Sun 5/31	@CLE 5, NYY 4	Recap
Mon 6/1	NYY 5, @CLE 2	Recap

· Complete Schedule: Yankees | Indians

SCORING SUMMARY

			CLE	NYY
	1st	M Teixeira homered to right, J Damon scored.	0	2
	2nd	S Choo homered to left center, T Hafner and J Peralta scored.	3	2
	2nd	A Cabrera singled to center, B Francisco scored.	4	2
	2nd	M DeRosa doubled to deep right, A Cabrera	6	2

Some Other Noteworthy Games

Thursday, May 17, 1979,, Wrigley Field
Attendance: 14,952, Time of Game: 4:03

Phillies		Cubs
23		22
24-10	at	16-16
⇐ Won 2 ⇒		⇐ Lost 2 ⇒
1st		4th, 7 GB
Danny Ozark		Herman Franks

W: Rawly Eastwick (1-(
L: Bruce Sutter (1-1)
Like 290 people like thi
see what your frier

	1	2	3	4	5	6	7	8	9	10	R	H	E
Phillies	7	0	8	2	4	0	1	0	0	1	23	24	2
Cubs	6	0	0	3	7	3	0	3	0	0	22	26	2

- Kingman (Cubs) 3 homeruns
- Schmidt (Phillies) 2 homeruns
- Schmidt game winning HR in 10th off of Bruce Sutter

One of 2 times both teams 20+

Wednesday, June 7, 1950,, Fenway Park
Attendance: 6,659, Time of Game: 2:28

Browns		Red Sox
4		20

Thursday, June 8, 1950,, Fenway Park
Attendance: 5,105, Time of Game: 2:42

Browns		Red Sox
4		29
13-28	at	30-19

Thursday, June 29, 1950,, Shibe Park
Attendance: 2,808, Time of Game: 2:50

Red Sox		Athletics
22		14
39-30	at	22-44

One of 2 consecutive games

Second time three in a season by one team

RED SOX June 1950

Who Holds the Records?

Rangers 30, Orioles 3

Recap | Box Score | Play-By-Play | Conversation

	1	2	3	4	5	6	7	8	9	R	H	E
TEX (55-70)	0	0	0	5	0	9	0	10	6	30	29	1
BAL (58-66)	1	0	2	0	0	0	0	0	0	3	9	1

Final

W: K. Gabbard (6-1)
L: D. Cabrera (9-13)
SV: W. Littleton (1)

Texas Rangers

Hitters	AB	R	H	RBI	BB	SO	LOB	AVG
F Catalanotto 1B	6	2	3	2	2	1	0	.267
J Botts DH	7	2	3	2	0	4	1	.234
N Cruz RF	7	2	2	0	0	2	5	.220
D Murphy LF	7	5	5	2	0	1	1	.550
J Saltalamacchia C	6	5	4	7	1	1	2	.219
R Vazquez 3B-SS	6	4	4	7	1	1	0	.240
Totals	57	30	29	30	8	11	19	

BATTING
2B: N Cruz (11, P Shuey); J Botts (4, P Shuey)
HR: R Vazquez 2 (7, 4th inning off D Cabrera 2 on, 1 Out; 9th inning off P Shuey 2 on, 2 Out); J Saltalamacchia 2 (4, 6th inning off D Cabrera 0 on, 0 Out; 8th inning off P Shuey 2 on, 1 Out); M Byrd (6, 6th inning off B Burres 3 on, 1 Out); T Metcalf (2, 8th inning off R Bell 3 on, 0 Out)
RBI: J Saltalamacchia 7 (12), R Vazquez 7 (24), M Byrd 4 (51), F Catalanotto 2 (37), I Kinsler 2 (45), T Metcalf 4 (10), D Murphy 2 (3), J Botts 2 (8)
2-out RBI: J Saltalamacchia, R Vazquez 4, F Catalanotto, I Kinsler
Runners left in scoring position, 2 out: M Young 3
Team LOB: 8

FIELDING
E: N Cruz (3, throw)
DP: 1 (M Young-I Kinsler-F Catalanotto).

Texas Rangers

Pitchers	IP	H	R	ER	BB	SO	HR	PC-ST	ERA
K Gabbard (W, 6-1)	6.0	7	3	3	1	3	0	89-56	3.64
W Littleton (S, 2)	3.0	2	0	0	1	1	0	43-29	3.86
Totals	9.0	9	3	3	2	4	0	132-85	

PITCHING
WP: K Gabbard
Batters faced: K Gabbard 25; W Littleton 12
Ground Balls-Fly Balls: K Gabbard 9-6; W Littleton 4-4
Game Scores: K Gabbard 48

Baltimore Orioles

Hitters	AB	R	H	RBI	BB	SO	LOB	AVG
B Roberts 2B	5	2	3	0	0	0	0	.314
atterson CF	3	0	1	0	0	0	2	.264
Redman CF	1	0	1	0	0	0	0	.364
arkakis RF	3	1	1	1	0	0	2	.295
Bynum LF-SS	1	0	0	0	0	0	1	.271
M Tejada SS	3	0	1	1	1	0	2	.301
P Shuey P	0	0	0	0	0	0	0	.000
K Millar 1B-LF	4	0	1	0	0	1	3	.266
M Mora 3B	3	0	1	0	1	1	1	.266
J House DH-C	4	0	0	0	0	1	5	.300
R Hernandez C-1B	4	0	0	0	0	1	1	.242
J Payton LF-RF	4	0	0	0	0	0	0	.255
Totals	35	3	9	2	2	4	17	

BATTING
2B: B Roberts (37, K Gabbard); N Markakis (35, K Gabbard)
RBI: N Markakis (76), M Tejada (62)
2-out RBI: N Markakis, M Tejada
GIDP: R Hernandez
Runners left in scoring position, 2 out: J House 3, K Millar 2
Team LOB: 7

FIELDING
E: M Mora (8, ground ball)

Baltimore Orioles

Pitchers	IP	H	R	ER	BB	SO	HR	PC-ST	ERA
D Cabrera (L, 9-18)	5.0	9	6	6	1	4	2	96-62	5.10
B Burres	0.2	8	8	8	1	1	1	34-22	5.24
R Bell	1.1	5	7	7	3	1	1	54-32	6.14
P Shuey	2.0	7	9	9	3	5	2	68-41	9.49
Totals	9.0	29	30	30	8	11	6	252-157	

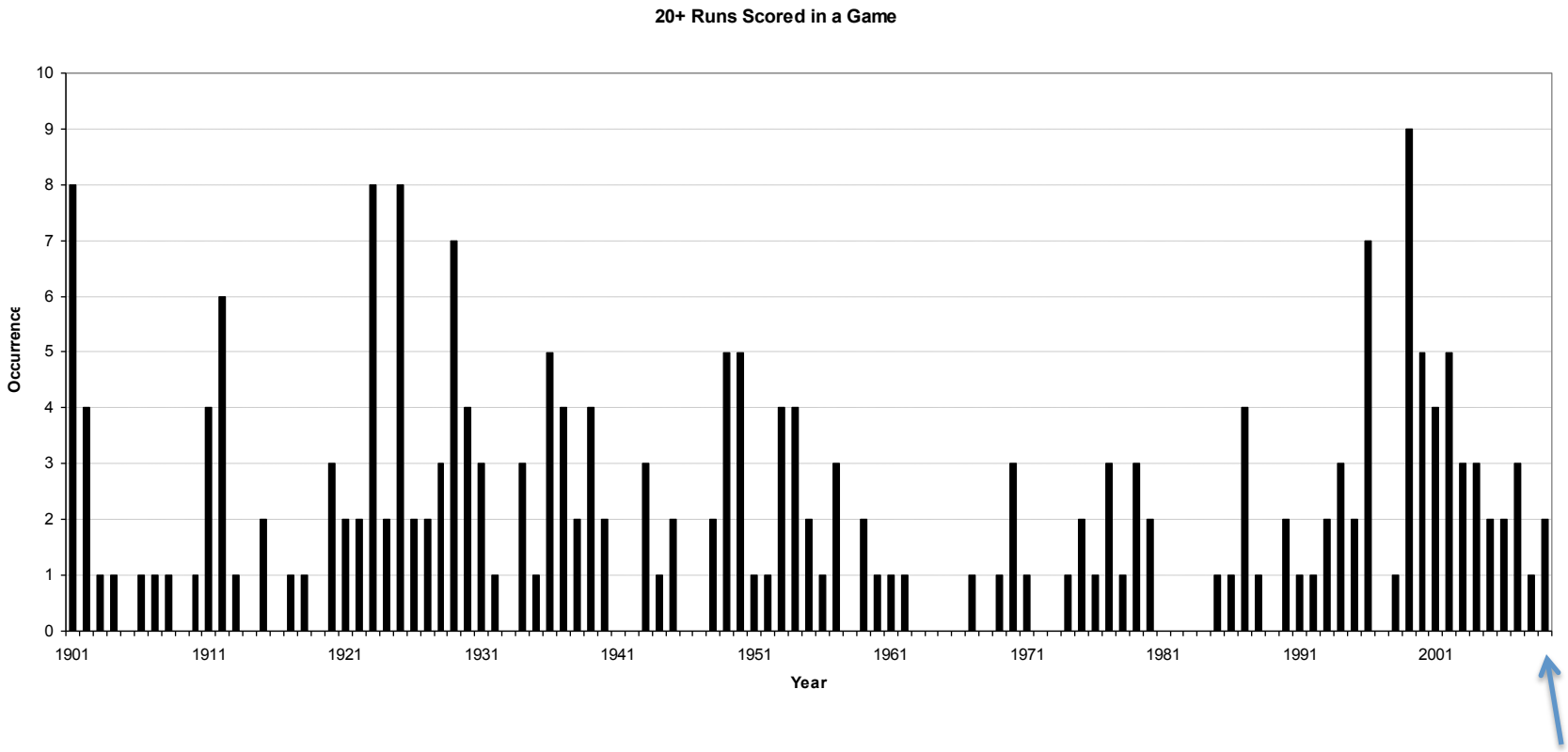
PITCHING
WP: B Burres
Batters faced: D Cabrera 26; B Burres 11; R Bell 12; P Shuey 16
Ground Balls-Fly Balls: D Cabrera 6-5; B Burres 1-0; R Bell 1-2; P Shuey 0-1
Game Scores: D Cabrera 28

Most runs in a game

A save when up 11 to 27 runs?

Rk	Tm	Year	G	W	L	W-L%	RS	RA	pythW-L%
1	NYY	1939	3	3	0	1.000	66	4	0.994
2	BOS	1950	3	3	0	1.000	71	22	0.895
3	BRO	1901	2	2	0	1.000	46	9	0.952
4	NYG	1912	2	2	0	1.000	43	22	0.773
5	CLE	1923	2	2	0	1.000	49	5	0.985
6	PIT	1925	2	2	0	1.000	45	11	0.929
7	PHA	1929	2	2	0	1.000	45	9	0.950
8	NYY	1931	2	2	0	1.000	42	13	0.895
9	CHC	1945	2	2	0	1.000	44	8	0.958
10	NYY	1949	2	2	0	1.000	40	7	0.960
11	DET	1993	2	2	0	1.000	40	7	0.960
12	MIN	1994	2	2	0	1.000	42	9	0.944
13	NYY	1999	2	2	0	1.000	42	4	0.987
14	CIN	1999	2	2	0	1.000	46	15	0.886
15	OAK	2000	2	2	0	1.000	44	5	0.982
16	PHI	2008	2	2	0	1.000	40	7	0.960
17	STL	1901	1	1	0	1.000	20	6	0.901
18	WSH	1901	1	1	0	1.000	20	8	0.842
19	MLA	1901	1	1	0	1.000	21	7	0.882
20	DET	1901	1	1	0	1.000	21	0	1.000
21	BOS	1901	1	1	0	1.000	23	12	0.767
22	NYG	1901	1	1	0	1.000	25	13	0.768
23	BLA	1902	1	1	0	1.000	21	6	0.908
24	PHA	1902	1	1	0	1.000	22	9	0.837
25	CLE	1902	1	1	0	1.000	23	7	0.898
26	CIN	1902	1	1	0	1.000	24	2	0.990
27	NYG	1903	1	1	0	1.000	20	2	0.985
28	NYY	1904	1	1	0	1.000	21	3	0.972
29	NYY	1906	1	1	0	1.000	20	5	0.927
30	PIT	1907	1	1	0	1.000	20	5	0.927
31	DET	1908	1	1	0	1.000	21	2	0.987
32	BSN	1910	1	1	0	1.000	20	7	0.872
33	CHC	1911	1	1	0	1.000	20	2	0.985
34	CHW	1911	1	1	0	1.000	20	6	0.901
35	PHI	1911	1	1	0	1.000	21	5	0.933
36	CIN	1911	1	1	0	1.000	26	3	0.981
37	STL	1912	1	1	0	1.000	20	5	0.927
38	BOS	1912	1	1	0	1.000	21	8	0.854
39	PIT	1912	1	1	0	1.000	23	4	0.961
40	PHA	1912	1	1	0	1.000	24	2	0.990
41	PHA	1913	1	1	0	1.000	21	8	0.854
42	IND	1914	1	1	0	1.000	21	6	0.908
43	WSH	1915	1	1	0	1.000	20	5	0.927
44	BSN	1915	1	1	0	1.000	20	1	0.996
45	CLE	1917	1	1	0	1.000	20	6	0.901
46	STL	1918	1	1	0	1.000	22	7	0.890

Frequency Distribution



RARE EVENT (< 1%)
1901-2008: 222/171,797 games (0.13%)

Data to JUNE 2009

The Memoryless Property

If the data indeed follows a Poisson process, the exponential distribution can be used as a model for the distribution of times between the occurrence of successive no-hitters or cycle events. As a definition, a random variable X is said to have an Exponential Distribution if its probability density function is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \lambda > 0.$$

Additionally, if X has an exponential distribution with parameter λ, then the expected value of X equals $1/\lambda$ and the variance of X is equal to $1/\lambda^2$. Both the mean and standard deviation are the same.

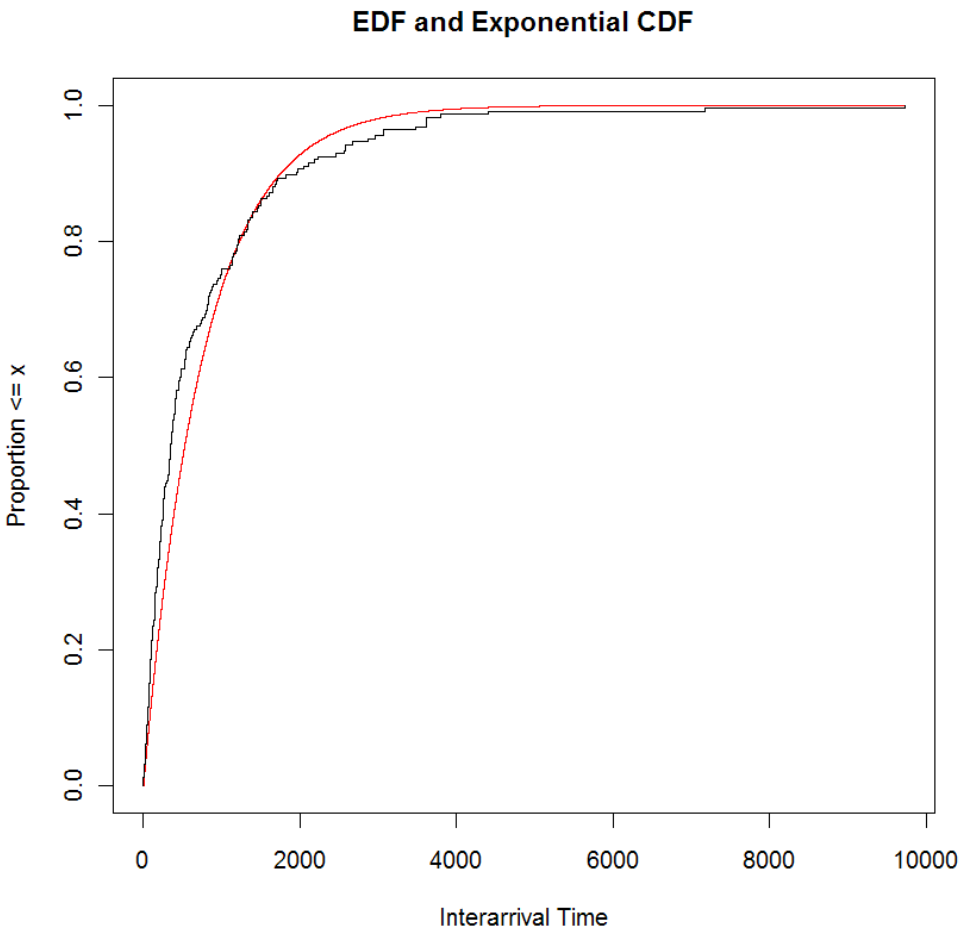
Calculate Inter-Arrival times, plot cdf with exponential models.

Calculating Inter-Arrival Times

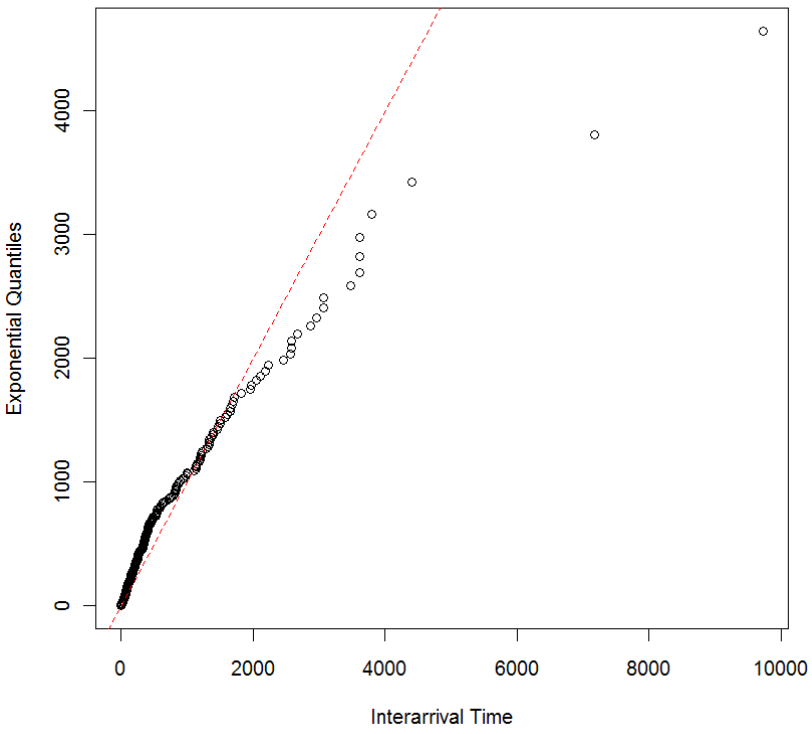
	A	B	C	D	E	F	G	H
1	Year	Team	Date	RS	Game #	IAT	G/Season	Notes
2	1901	BOS	2-May-1901	23	55	55		
3		MLA	5-May-1901	21	77	22		
4		NYG	9-Jun-1901	25	284	207		
5		BRO	21-Jun-1901	21	365	81		
6		STL	5-Aug-1901	20	678	313		
7		WSH	7-Sep-1901	20	931	253		7 innings
8		DET	15-Sep-1901	21	995	64		8 innings
9		BRO	23-Sep-1901	26	1037	42	1110	
10	1902	CIN	13-May-1902	24	143	216		
11		PHA	8-Jul-1902	22	507	364		
12		BLA	25-Aug-1902	21	842	335		
13		CLE	2-Sep-1902	23	906	64	1117	
14	1903	NYG	6-May-1903	20	109	320	1114	
15	1904	NYY	14-Jul-1904	21	568	1573	1249	
16	1905	None					1237	
17	1906	NYY	31-Aug-1906	20	955	2873	1228	Game 2
18	1907	PIT	22-Aug-1907	20	878	1151	1233	
19	1908	DET	17-Jul-1908	21	632	987	1244	
20	1909	None					1241	
21	1910	BSN	6-Oct-1910	20	1214	3067	1249	
22	1911	CHW	11-May-1911	20	173	208		

An Exponential Model for the data

GOF test	Test Statistic	p-value
KS Test	0.1581	<< 0.0001
A-D Test	8.207	<< 0.0025
χ^2 Test	44.616	<< 0.05

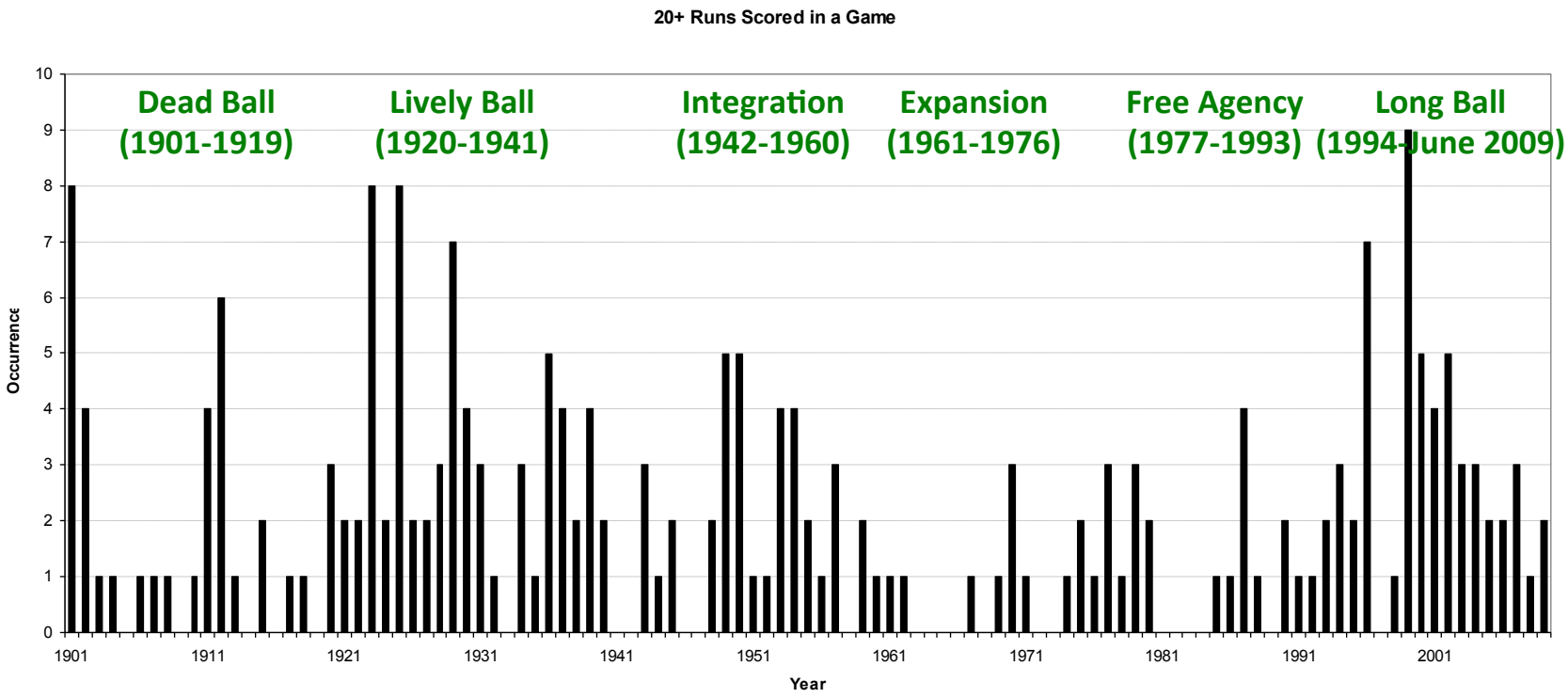


10 worst fit data points



Year	Team	Games since previous 20+ run event
1985	Philadephia Phillies	9723
1967	Chicago Cubs	7181
1974	Kansas City Royals	4408
1975	Boston Red Sox	3608
1992	Milwaukee Brewers	3471
1969	Oakland A's	3612
1910	Boston Doves	3067
1948	St. Louis Cardinals	3620
1986	Boston Red Sox	2960
1990	SF Giants	3795

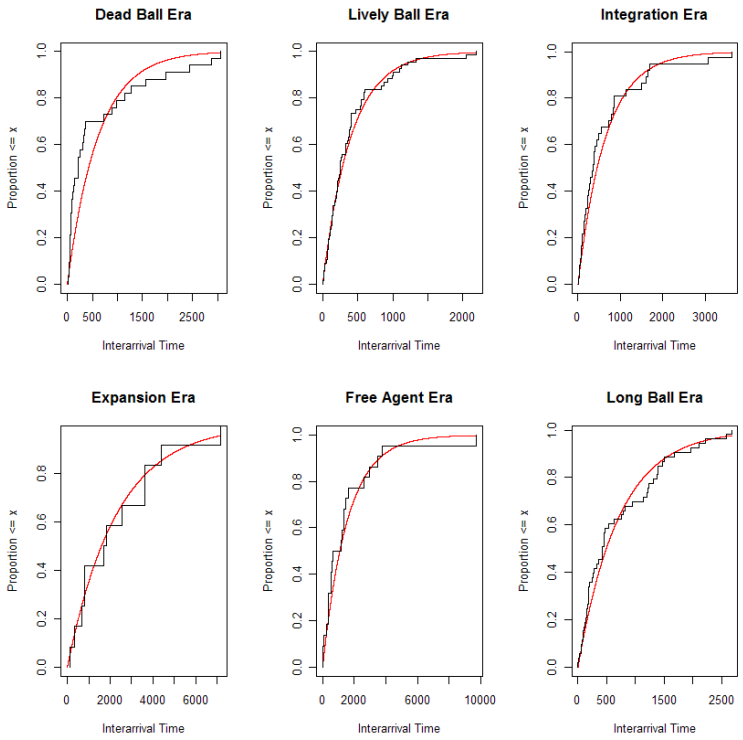
Era's


20+ Runs Scored in a Game
Dead Ball (1901-1919)
Lively Ball (1920-1941)
Integration (1942-1960)
Expansion (1961-1976)
Free Agency (1977-1993)
Long Ball (1994-June 2009)
Occurrence
Year
0
1
2
3
4
5
6
7
8
9
10
1901
1911
1921
1931
1941
1951
1961
1971
1981
1991
2001

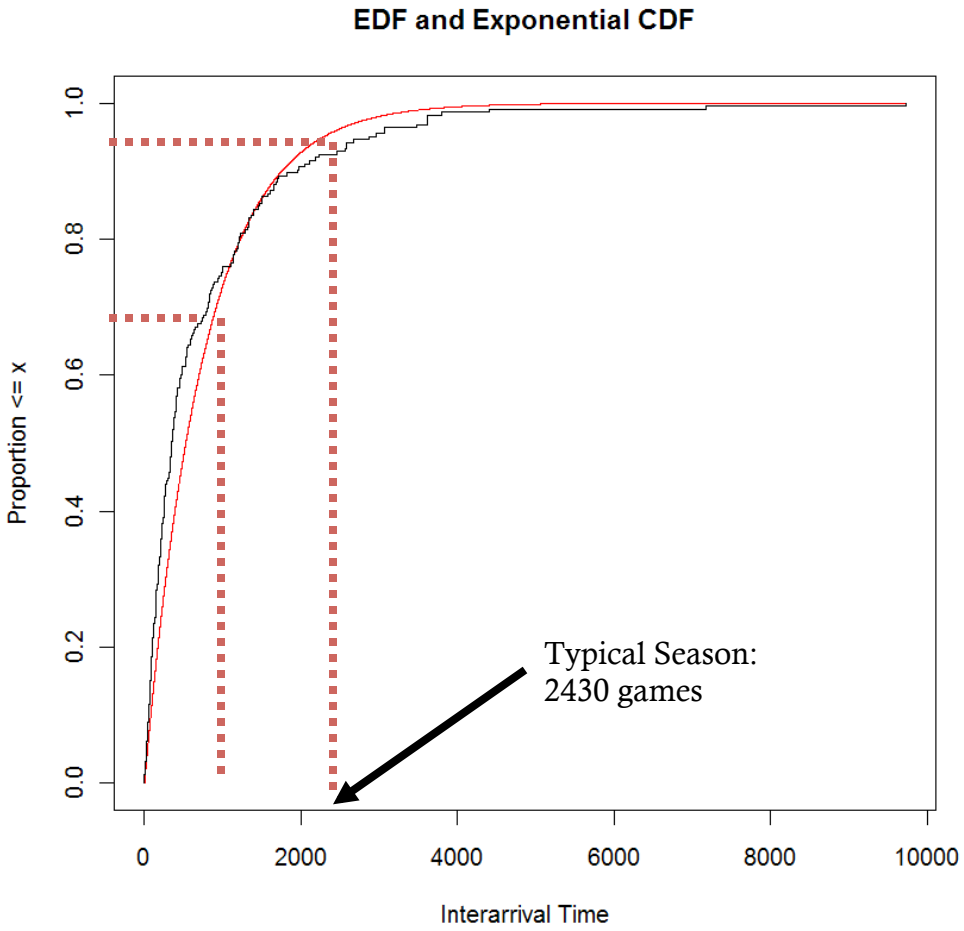
Baseball Eras

Era	Dates
Dead Ball	1901-1919
Lively Ball	1920-1941
Integration	1942-1960
Expansion	1961-1976
Free Agency	1977-1993
Long Ball	1994-June 2009

Era	20+ games in Era	Mean IAT
Dead Ball	33	612.5152
Lively Ball	68	400.7941
Integration	37	650.1351
Expansion	12	2307.3333
Free Agency	22	1561.4091
Long Ball	53	709.7547



When Will We See the Next Game?



Michael R. Huber and Rodney X. Sturdivant, "Building a Model for Scoring 20 or More Runs in a Baseball Game," *Annals of Applied Statistics,* Volume 4, Number 2, 2010.

The "Survival" Function

- Time until the event of interest occurs, *T*
  - e.g. time until next 20 run game
- We modeled using the CDF: $F(t) = P(T \leq t)$
- The SURVIVAL FUNCTION:

$$S(t) = \Pr(T > t)$$

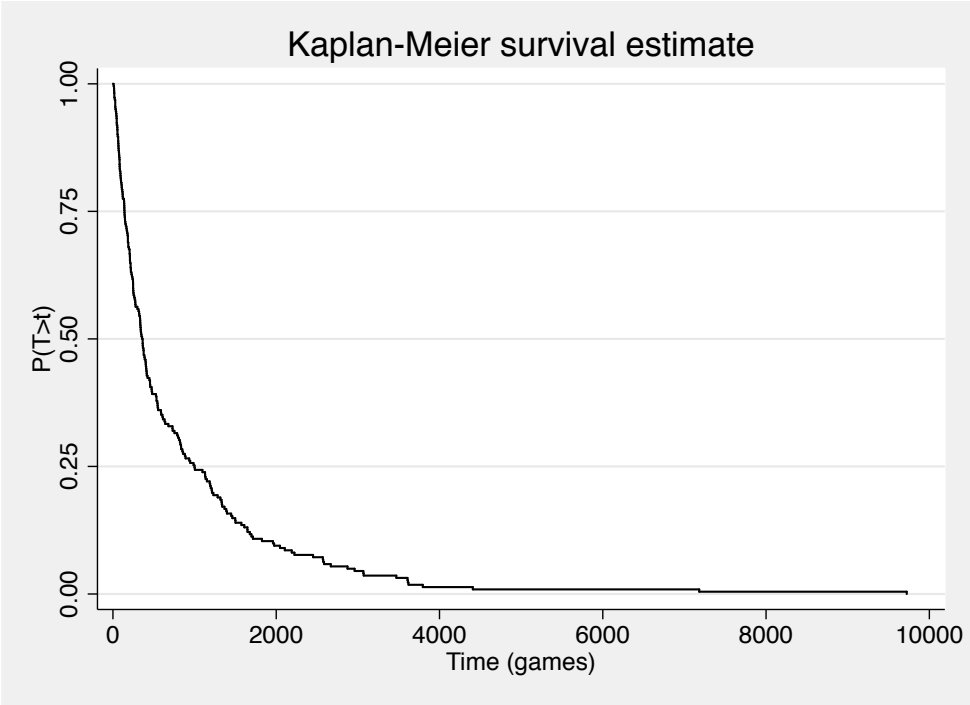
$$= \Pr(\text{event has not occurred by time } t)$$

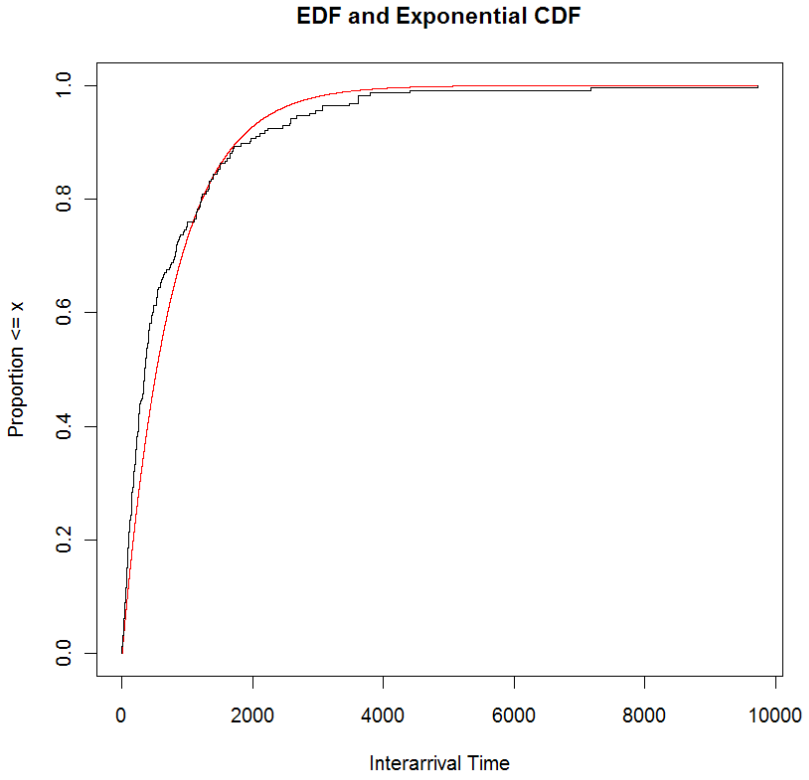
$$S(t) = 1 - F(t)$$

- Note that *S(t)* decreases with time and that *S(0)* = 1

An empirical estimator

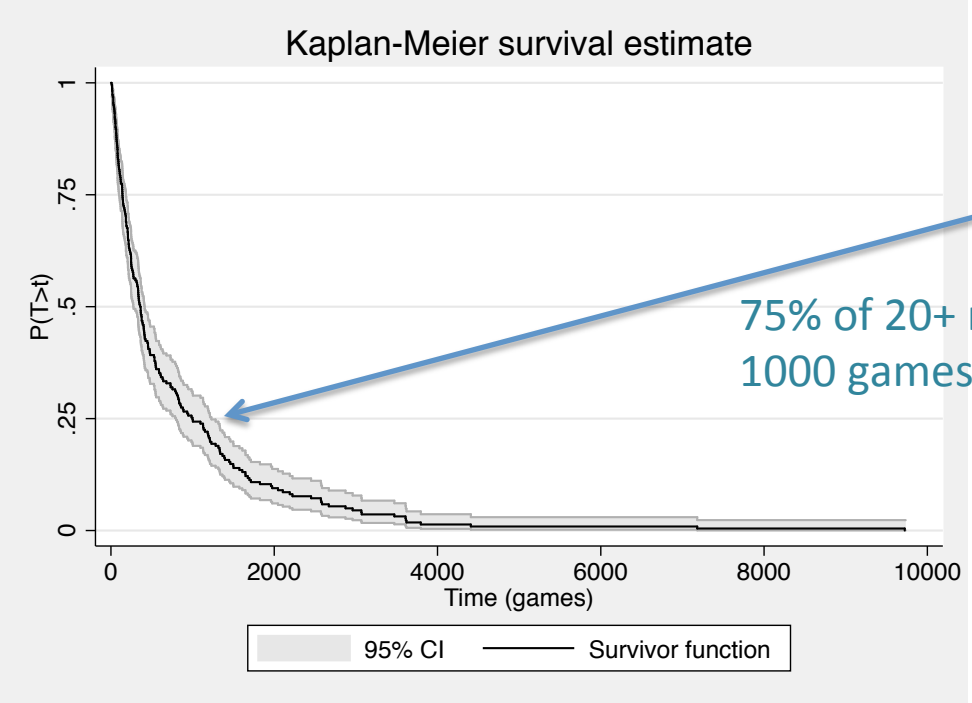
- Data based estimators of probability of "surviving" beyond *t*
  - Plays the role of the EDF





Immediate advantages

- Available in software including:
  - Confidence interval estimates
  - Estimates of means, percentiles



Percentile	Estimate	95% CI
25th	141	(99, 183)
50th (median)	357	(274, 413)
75th	999	(797, 1230)

75% of 20+ run games were less than 1000 games apart

Statistic	Estimate	95% CI
Mean	771.1	(622.9, 919.2)

- Mean is the area under the survival function curve
- Note the data is skewed
  - mean > median

Building a Model

- We could use *S(t)* as we did *F(t)*
  - Want more flexibility in accounting for things like baseball era in the model ("covariates")
  - Want readily available tools/statistics
    - Confidence intervals, tests of significance, assessment of model fit
- ONE more function: regression models of the HAZARD function

The "Hazard" Function

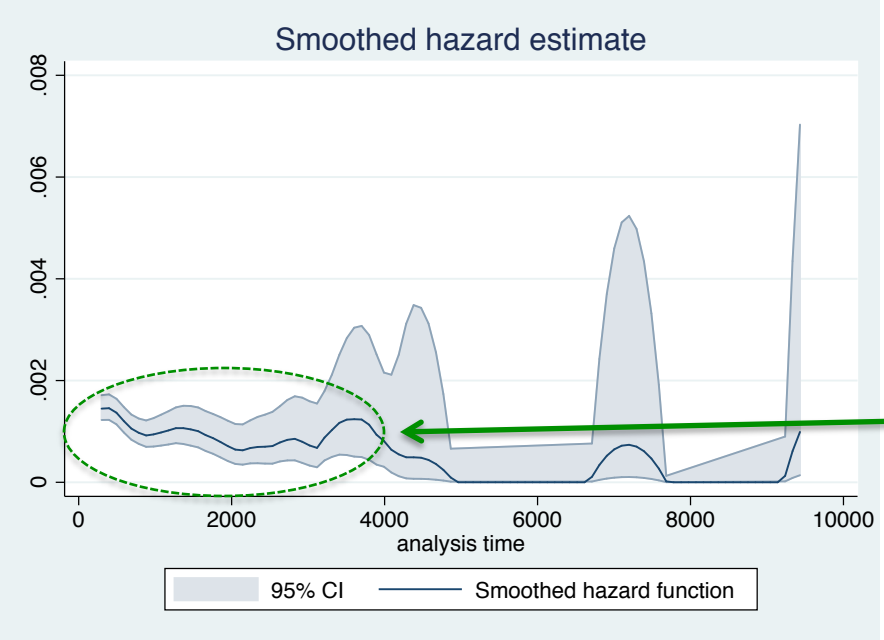
- The hazard at time t is the event rate "per time unit" conditional on the fact that the event has yet to occur at that time, e.g. is "at risk" at time t.
- In our example, consider 2000 games in between
  - Conditional rate of 20+ run games after 2000 games without event
  - Applies only to subset for which the interval is not less than 2000 games
- Related to the survival function

$$h(t) = \frac{f(t)}{S(t)}$$

$f(t)$ is the probability density function

The "Hazard" Function

- From data, estimated hazard function tends to jump around due to sampling variability



In our data appears to get lower as time goes by

- Note much less data for longer times between 20+ run games

Exponential hazard function is a CONSTANT

- Reasonable to assume in this data?
- Note the rate appears to be slightly higher than 0.001 (on average)

The Exponential Hazard Model

- Software (SAS) can fit this model

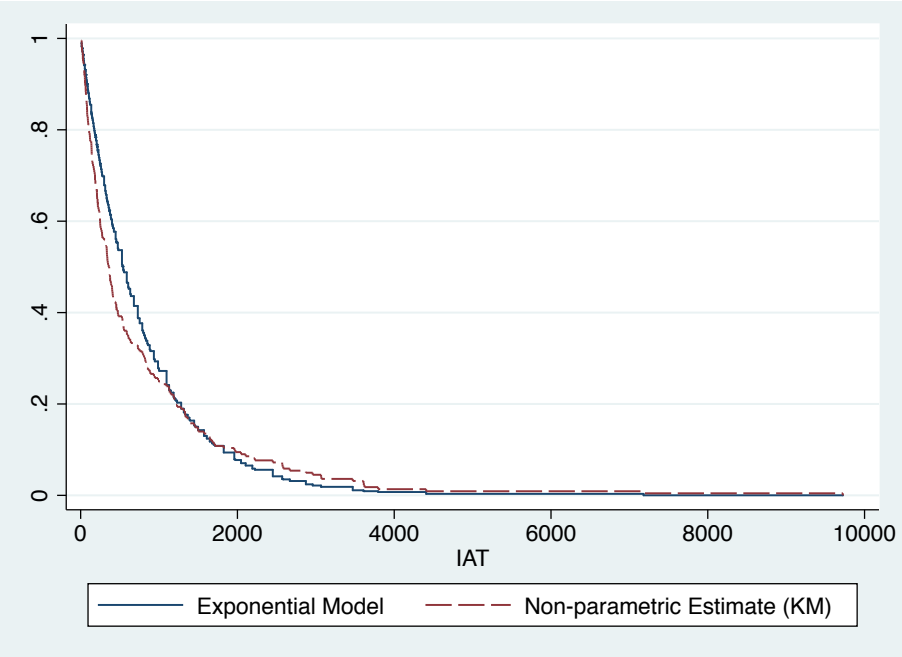
Hazard rate	95% CI
0.0013	(0.0011, 0.0015)

Can then estimate survival function through cumulative hazard:

$$H(t) = \int_0^t h(u)du$$

$$S(t) = e^{-H(t)}$$

Rate estimate of 0.0012969:
Mean = 1/0.0012969
= 771 games between 20+ run events
= empirical estimate (KM)



This is our CDF/EDF modeling "flipped" to 1-0 instead of 0-1

Is this helpful? Yes ☺

A Regression model

- We began with a simple model (exponential, constant hazard):

$$h(t) = \lambda$$

IN GENERAL WHAT WE NEED:

1. Hazard function is a rate ⇨ must be >0
2. A desirable property for a statistical model to have is to be parameterized in such a way that the allowable range of parameter values is infinite

A Regression model

- One approach to handle both issues $h(t) = \lambda = e^{\beta_0}$

- Given this form a natural way to include covariates is to have them be additive on the log scale, specifically for a covariate *x* the log-hazard function is

$$\ln\left[h(t,x)\right] = \beta_0 + \beta_1 x$$

- and the hazard function is

$$h(t,x) = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x} = \lambda e^{\beta_1 x}$$

- Does not depend on time (constant) in this case
- May be plausible for some, but not so in other settings
  - If hazard changes over time, could model with a function of time λ(t)
- Can use to get the cumulative hazard function and then survival function

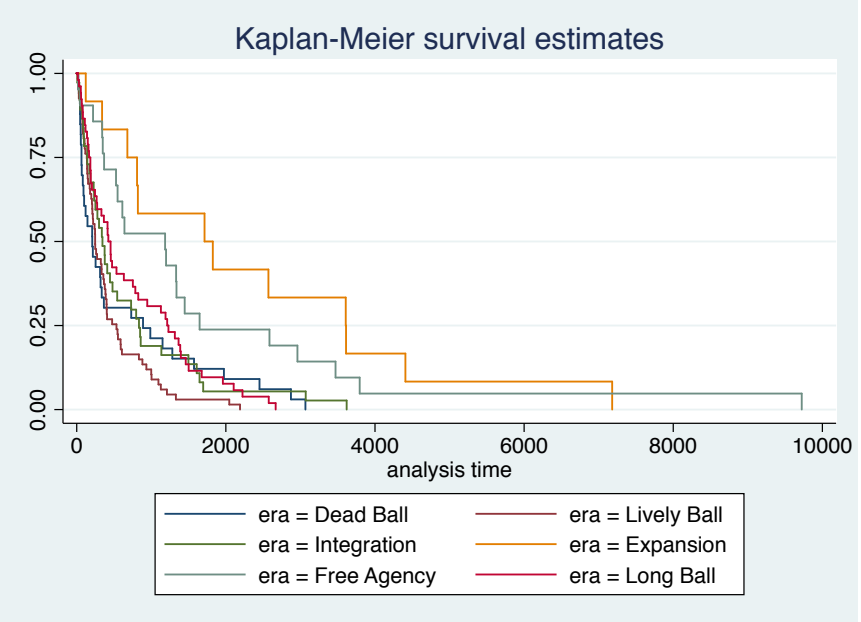
Including ERA in the model

- Additional terms in the model:

$$\ln\left[h(t,x)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

$$h(t,x) = \lambda e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5}$$

Era	x_1	x_2	x_3	x_4	x_5
Dead Ball (1901-1919)	0	0	0	0	0
Lively Ball (1920-1941)	1	0	0	0	0
Integration (1942-1960)	0	1	0	0	0
Expansion (1961-1976)	0	0	1	0	0
Free Agency (1977-1993)	0	0	0	1	0
Long Ball (1994-June 2009)	0	0	0	0	1



Non-parametric Results

- Kaplan-Meier estimates:

Era	Events	Median	CI	Mean
Dead Ball (1901-1919)	33	208	(81,335)	612.5
Lively Ball (1920-1941)	67	247	(195,365)	406.8
Integration (1942-1960)	37	345	(227,543)	650.1
Expansion (1961-1976)	12	1714	(340,3612)	2307.3
Free Agency (1977-1993)	21	1185	(367,1448)	1635.7
Long Ball (1994-June 2009)	52	424	(245,753)	723.1

- Non-parametric tests of equality of survival functions reject ($p<0.0001$)
  - Log rank, Wilcoxon, Tarone-Ware, Peto-Prentice
  - Another immediate advantage of approach
  - Statistical difference between era's

Exponential Regression Model Results

- Parameter estimates:

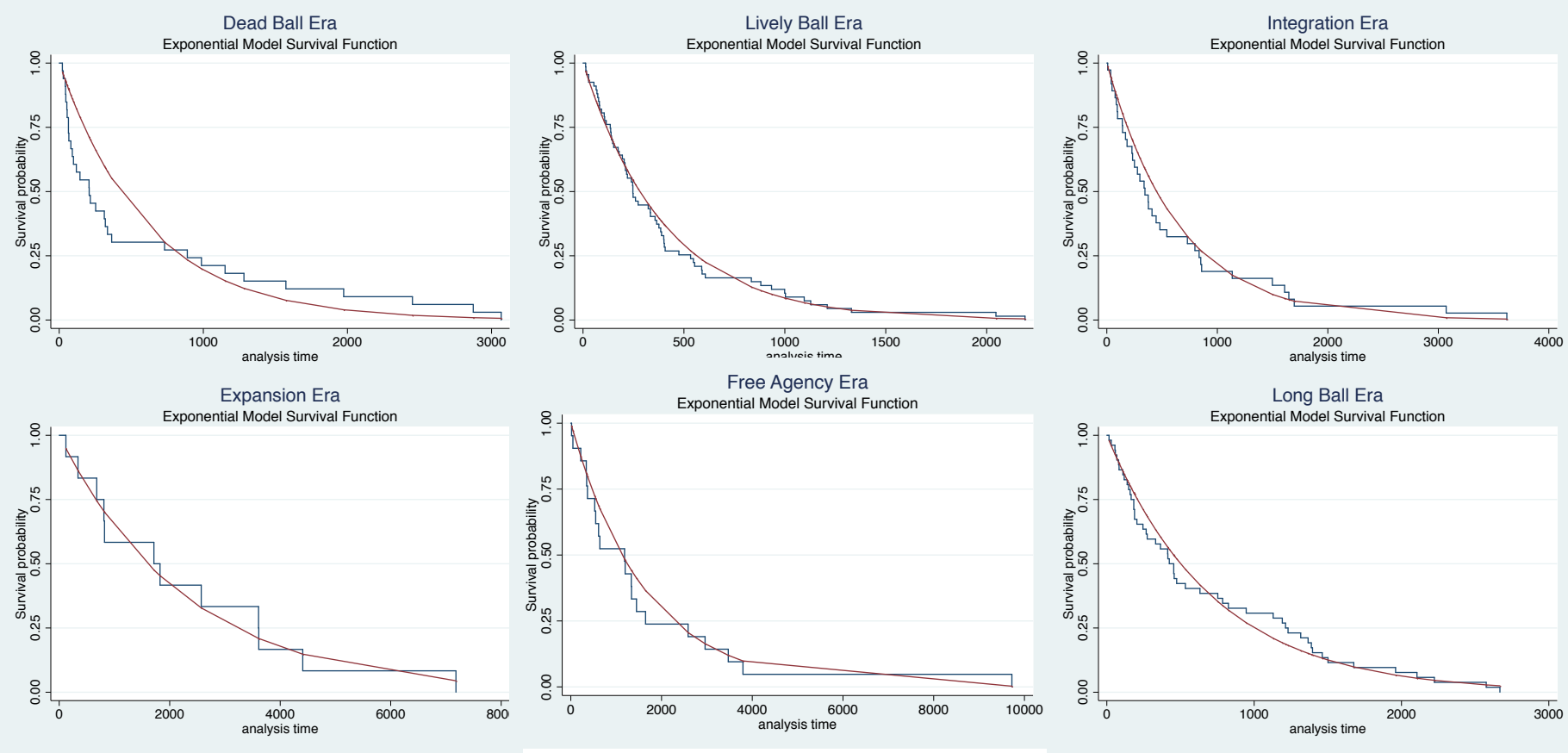
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
era						
Lively Ball	.4093474	.2126697	-1.92	0.054	-.8261724	.0074775
Integration	-.0596066	.239437	0.25	0.803	-.4096813	.5288944
Expansion	-1.326274	.3370999	3.93	0.000	.6655702	1.986978
Free Agency	-.9822612	.2791453	3.52	0.000	.4351465	1.529376
Long Ball	-.1663674	.2225619	0.75	0.455	-.2698459	.6025807
Constant	6.417574	.1740777	36.87	0.000	6.076388	6.75876

Likelihood Ratio test (overall model significance) p < 0.0001

DEAD BALL era the comparison group (all indicator variables = 0)

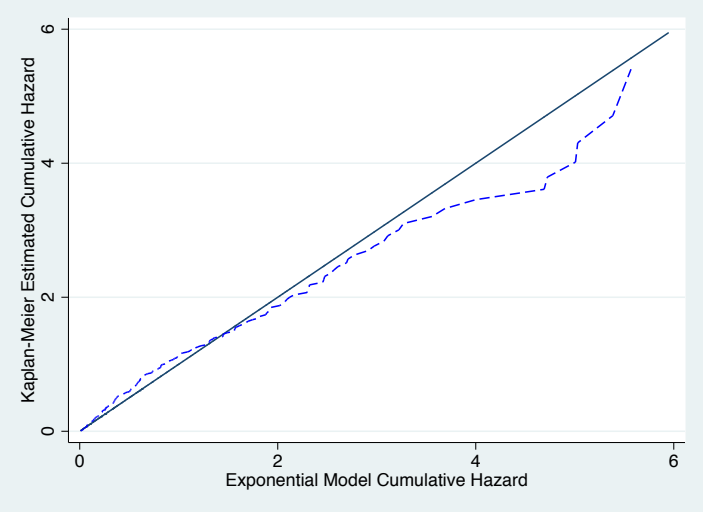
- Lively Ball slightly increased "hazard" of 20+ run game (p = 0.054)
- Integration/Long Ball decrease "hazard" but not significant (p = 0.8/0.46)
- Expansion/Free Agency significantly decrease "hazard" (p < 0.001)

Model Fit



Reasonable fit in all but DEAD BALL era – model rate too low early, too high late

More on Model Fit



Test of exponential hazard specification:

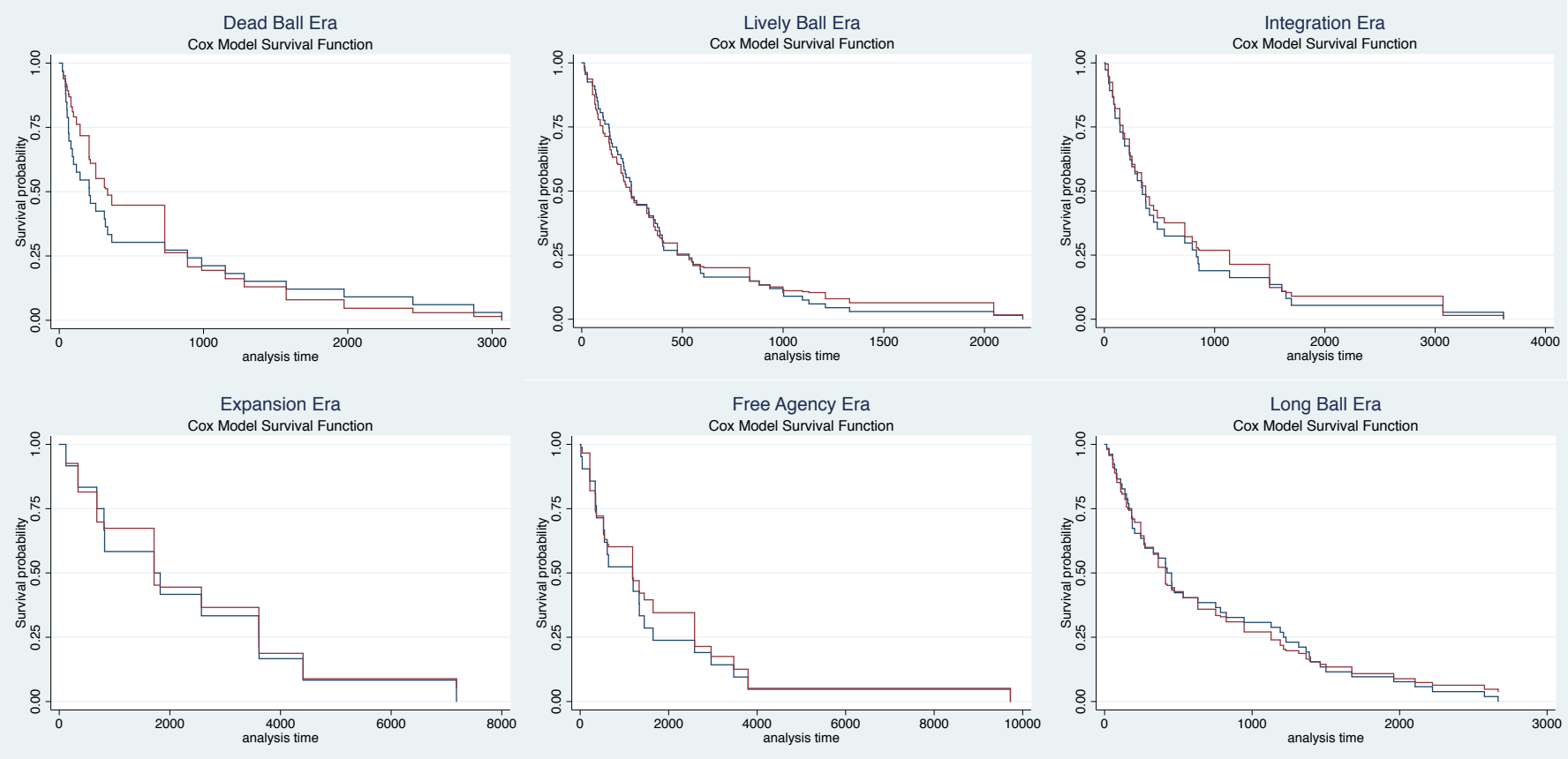
- Late departure from line lack of data
- Some evidence earlier of systematic departure – hazard too low early, too high late
- Recall the data suggested slightly decreasing hazard over time

$$h(t,x) = \lambda(t)e^{\beta x}$$

Could update model so "baseline hazard" a function of time

- Other fully parametric choices include Weibull, logistic
- "Semi-parametric" do not specify the baseline hazard function (Cox model)

A “Semi-parametric” Approach

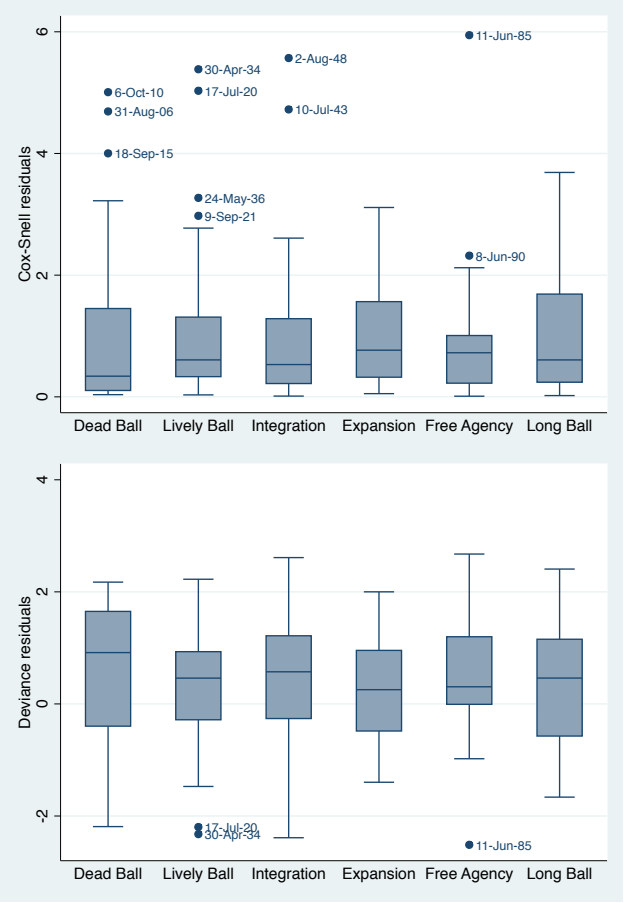


- Baseline hazard estimated from data
- Models focus of much of survival analysis texts/courses

Proceeding with Exponential Model

Can perform MODEL ASSESSMENT:

- Various residuals produced by software
- Indicators of subjects with potential leverage, influence
- Two examples below: Cox-Snell and Deviance residuals



Year	Team	Games since previous 20+ run event
1910	Boston Doves	3067
1906	NY Yankees	2873
1915	Boston Braves	2451
1934	Chicago White Sox	2190
1920	NY Yankees	2046
1936	NY Yankees	1330
1921	Chicago White Sox	1210
1943	Brooklyn Dodgers	3071
1948	St. Louis Cardinals	3620
1990	SF Giants	3795
1985	Philadephia Phillies	9723

- *Longest times between 20+ Run games in era's*
- *1985: previous 20+ game April 27, 1980*

Interpreting the Model

- Recall the model is a constant hazard function (exponential distribution):

$$h(t,\mathbf{x}) = \lambda e^{\beta_1 x_1 + \cdots + \beta_5 x_5}$$

- The survival function for this model is then:

$$S(t) = \exp\left(-H(t)\right) = \exp\left(-\lambda t e^{\beta_1 x_1 + \cdots + \beta_5 x_5}\right)$$

The *median* survival time is then (set *S(t)* = 0.5 and solve for *t*):

$$t_{50}(\mathbf{x},\boldsymbol{\beta}) = -\frac{1}{\lambda} e^{-\beta_1 x_1 - \cdots - \beta_5 x_5} \times \ln(0.5)$$

Time Ratios

- Let's consider two era's, Dead Ball and Free Agency (recall these differed statistically. Their median times are:

$$t_{50,fa}\left(\mathbf{x}=(0,0,0,1,0),\boldsymbol{\beta}\right)=-\frac{1}{\lambda}e^{-\beta_1 0-\cdots-\beta_4 1-\beta_5 0}\times\ln(0.5)$$

$$t_{50,db}\left(\mathbf{x}=(0,0,0,0,0),\boldsymbol{\beta}\right)=-\frac{1}{\lambda}e^{-\beta_1 0-\cdots-\beta_5 0}\times\ln(0.5)$$

- The ratio of these times is: $\mathrm{TR}\left(x_4=1,x_4=0\right)=\dfrac{-\lambda e^{-\beta_4}\times\ln(0.5)}{-\lambda e^{0}\times\ln(0.5)}=e^{-\beta_4}$

- The exponentiation of other parameters similarly TR compared to Dead Ball era.
- This result holds not just for the median but for all percentiles.

The quantity exp(-β) is called the ***acceleration factor***

* Some software uses the parameterization θ = -β so the estimates are in terms of this "accelerated failure time" interpretation instead of the hazards; see Hosmer, Lemeshow and May (2008) section 8.1 for details

Time Ratio Estimates

	Tm. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
era						
Lively Ball	.6640835	.1412304	-1.92	0.054	.4377215	1.007506
Integration	1.061419	.2541429	0.25	0.803	.6638618	1.697055
Expansion	3.766981	1.269849	3.93	0.000	1.9456	7.293456
Free Agency	2.670488	.745454	3.52	0.000	1.545189	4.615295
Long Ball	1.181007	.2628471	0.75	0.455	.7634971	1.826827
_cons	612.5152	106.6252	36.87	0.000	435.4534	861.5729

- Comparing the Free Agency era to the Dead Ball era:
  - The estimated median games between 20+ runs scored is 2.7 times as long in Free Agency compared to Dead Ball
  - The confidence interval suggests this increase could be as little as 1.5 times or as much as 4.6 times as long

Comparing other era's

- We see that 20+ games seem to pick up in the Long Ball era – suppose we want to compare this era to the Free Agency era
- Start by getting the expressions of the median survival time for each from the model:

$$t_{50,fa}\left(\mathbf{x}=(0,0,0,1,0),\boldsymbol{\beta}\right)=-\frac{1}{\lambda}e^{-\beta_1 0-\cdots-\beta_4 1-\beta_5 0}\times\ln(0.5)$$

$$t_{50,lb}\left(\mathbf{x}=(0,0,0,0,1),\boldsymbol{\beta}\right)=-\frac{1}{\lambda}e^{-\beta_1 0-\cdots-\beta_4 0-\beta_5 1}\times\ln(0.5)$$

- The time ratio is then: $\mathrm{TR}\left(1b,fa\right)=\dfrac{-\lambda e^{-\beta_5}\times\ln(0.5)}{-\lambda e^{-\beta_4}\times\ln(0.5)}=e^{\beta_4-\beta_5}$

- Software can produce this estimate and confidence interval:

_t	Time Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lb v. Fa	.4422439	.1143436	-3.16	0.002	.2664286	.734079

Time to 20+ run games shortened by 55% in Long Ball era compared to Free Agency

A Second Interpretation: the Hazard Ratio

- Back to the Dead Ball and Free Agency case, the hazard function:

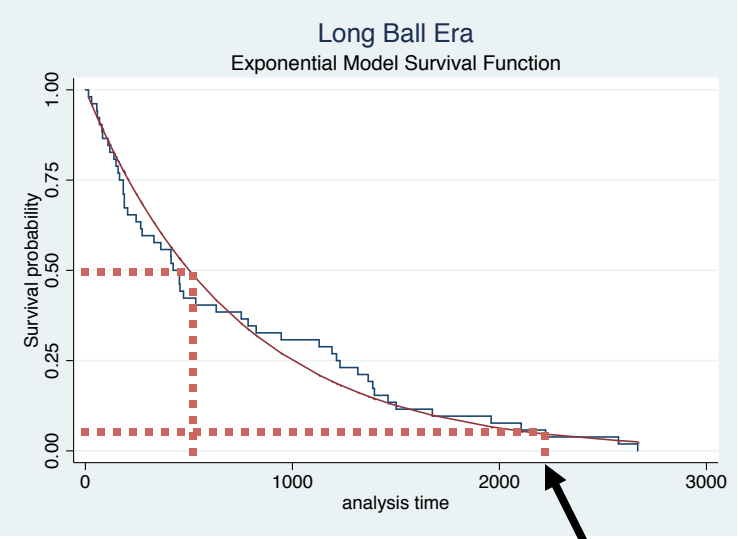
$$h(t,\mathbf{x}) = \lambda e^{\beta_1 x_1 + \cdots + \beta_5 x_5}$$

- Leads to the ratio of hazards (FA vs. DB) of: $HR(x_4 = 1, x_4 = 0) = e^{\beta_4}$

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
era						
Lively Ball	1.505835	.3202454	1.92	0.054	.9925503	2.284558
Integration	.9421351	.225582	-0.25	0.803	.5892561	1.506338
Expansion	.2654646	.0894881	-3.93	0.000	.1371092	.5139804
Free Agency	.3744634	.1045297	-3.52	0.000	.2166709	.6471698
Long Ball	.8467351	.188451	-0.75	0.455	.5473972	1.309763
_cons	.0016326	.0002842	-36.87	0.000	.0011607	.0022965

The interpretation is that 20+ run games in the Free Agency era occur at a rate that is estimated to be 0.37 times that of the Dead Ball era and the hazard ratio could be as low as 0.22 times or as high as 0.65 times with 95% confidence.

When Will We See the Next Game?



Typical Season:
2430 games

Median "survival" estimate 501 games

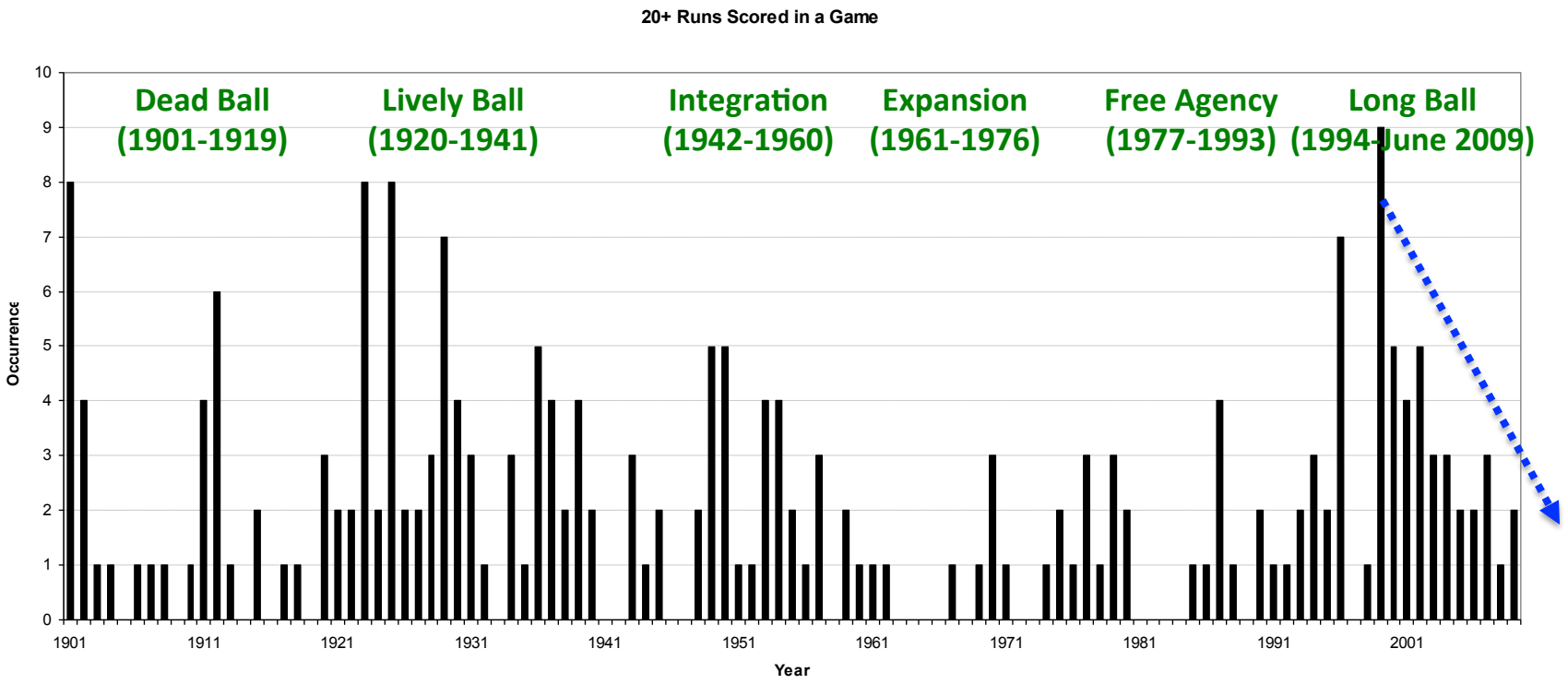
$$t_{50,lb}(\mathbf{x}=(0,0,0,0,1),\boldsymbol{\beta}) = -e^{\beta_0-\beta_5 1}\times\ln(0.5)$$
$$= -e^{6.4+0.17}\times\ln(0.5)$$
$$\approx 501$$

50/50 chance we see 20+ runs in next 500 games

$$S(2430) = \exp\left(-2430e^{\beta_0+\beta_5 1}\right)$$
$$\approx 0.035$$

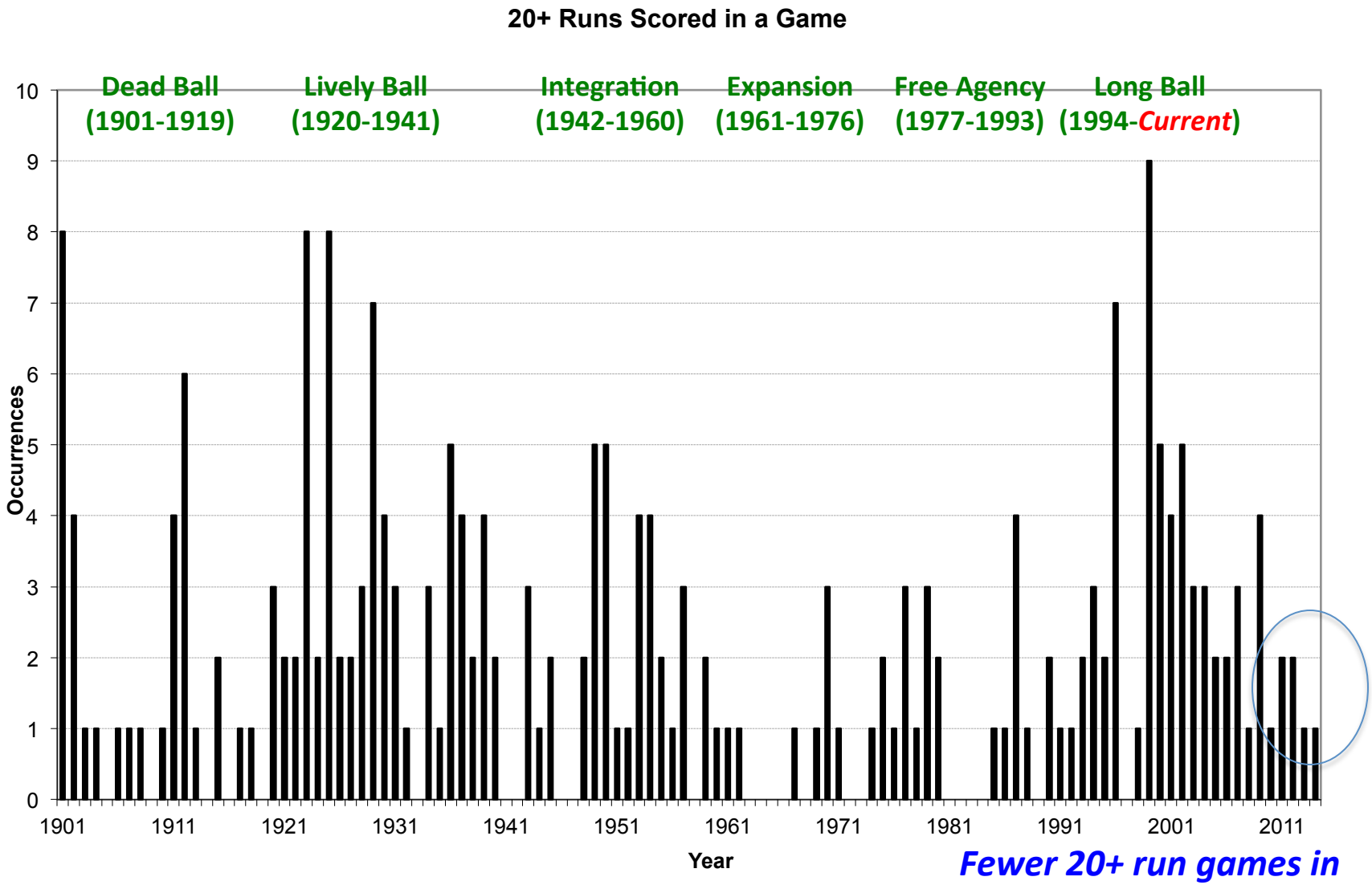
< 5% probability of no 20+ run games this season

Era's



But… Is Long Ball Era over ?

New Data



Fewer 20+ run games in recent years?

Non-parametric Results

- Kaplan-Meier estimates:

Era	Events	Median	CI	Mean
Dead Ball (1901-1919)	33	208	(81,335)	612.5
Lively Ball (1920-1941)	67	247	(195,365)	406.8
Integration (1942-1960)	37	345	(227,543)	650.1
Expansion (1961-1976)	12	1714	(340,3612)	2307.3
Free Agency (1977-1993)	21	1185	(367,1448)	1635.7
Long Ball (1994-June 2009)	52	424	(245,753)	723.1
Long Ball (1994-Current)	61	457	(364, 825)	837.0

9 New events in 5 ½ years since our original analysis

Changes empirical estimates of median/ mean time to next event when added to data…

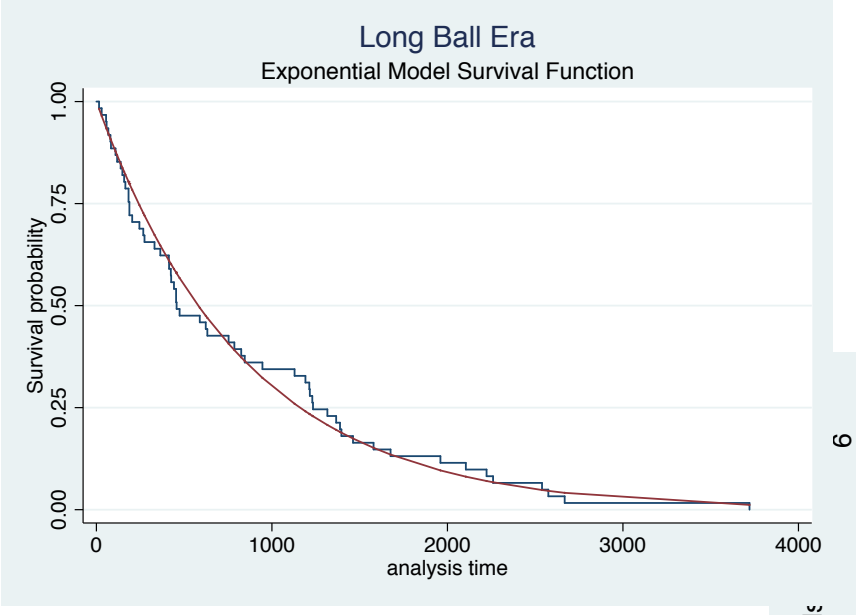
Exponential Model Hazard Ratios

- Long Ball Era Hazard ratio (reference group Dead Ball) changes with new data…

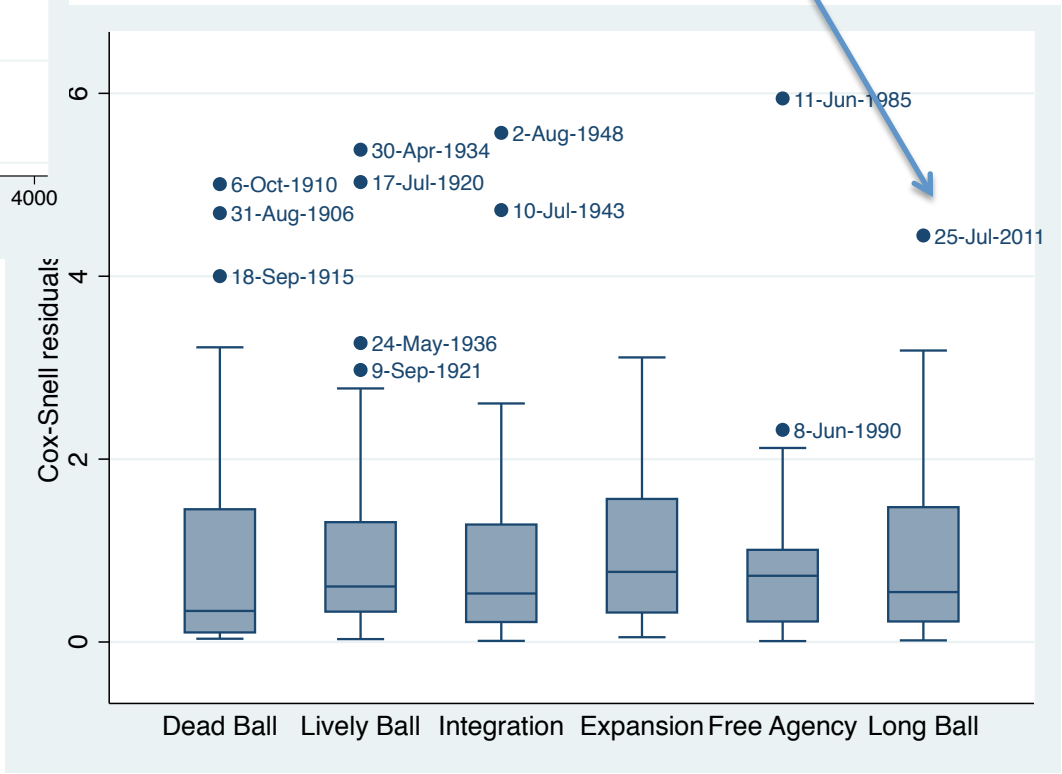
_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
era						
Lively Ball	1.505835	.3202454	1.92	0.054	.9925503	2.284558
Integration	.9421351	.225582	-0.25	0.803	.5892561	1.506338
Expansion	.2654646	.0894881	-3.93	0.000	.1371092	.5139804
Free Agency	.3744634	.1045297	-3.52	0.000	.2166709	.6471698
Long Ball(09)	.8467351	.188451	-0.75	0.455	.5473972	1.309763
Long Ball	.7318126	.1581401	-1.44	0.148	.479138	1.117736

- 20+ run games in the Long Ball era occur at a rate that is estimated to be 0.73 times that of the Dead Ball era
- Rate previously was 0.85 times – now closer to significant difference

Model fit quite good…not a new era?



Fit actually better overall
One new outlier identified



July 25, 2011 Texas Rangers score 20

- Previous 20 run game 3721 days prior (Brewers, April 22, 2010)

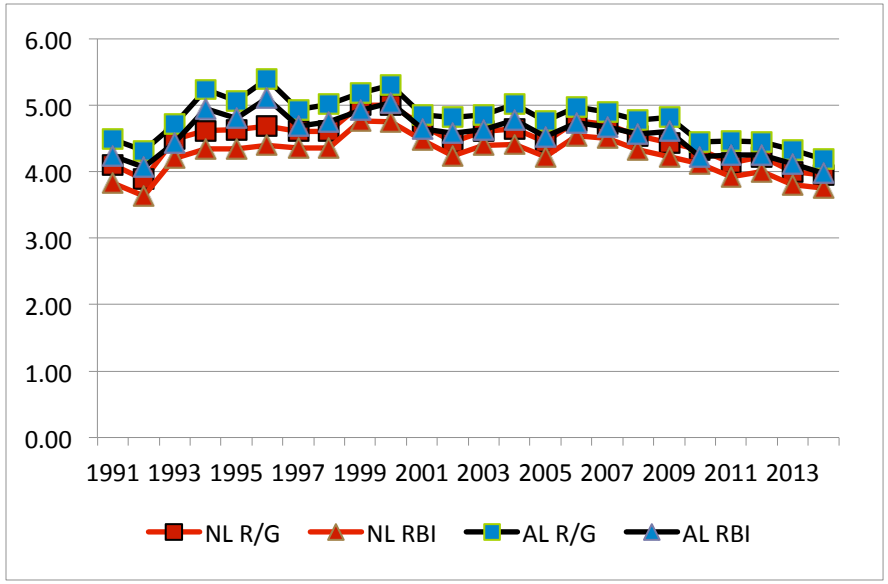
Are we in a new era?

HITTING DOWN

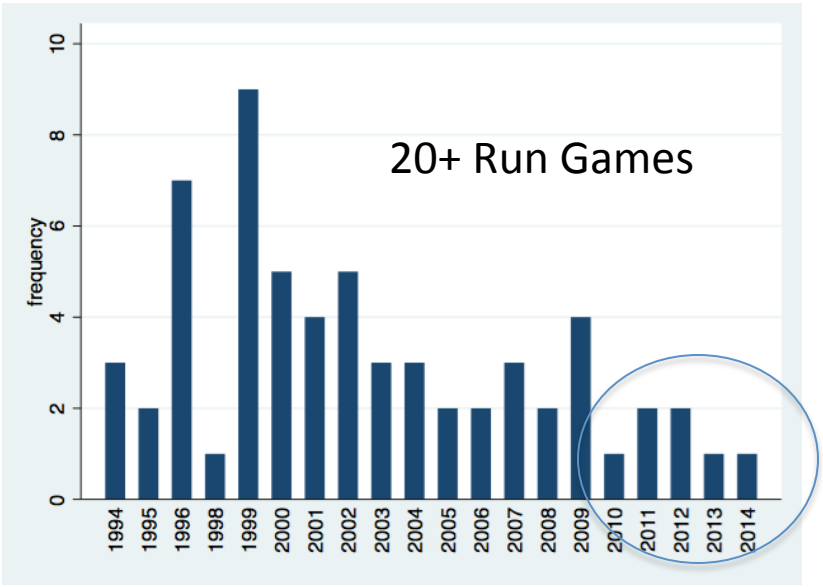
Batting Average
1991-2009: 0.269
2009-2015: 0.256

Year	R/G	RBI	BA	OPS
2015	4.29	4.06	0.251	0.714
2014	4.18	3.97	0.253	0.706
2013	4.33	4.12	0.256	0.725
2012	4.45	4.25	0.255	0.731
2011	4.46	4.25	0.258	0.730
2010	4.45	4.23	0.260	0.734
2009	4.82	4.60	0.267	0.764
2008	4.78	4.56	0.268	0.756
2007	4.90	4.67	0.271	0.761
2006	4.97	4.74	0.275	0.776
2005	4.76	4.53	0.268	0.755
2004	5.01	4.77	0.270	0.771
2003	4.86	4.63	0.267	0.761
2002	4.81	4.58	0.264	0.755
2001	4.86	4.64	0.267	0.762
2000	5.30	5.04	0.276	0.792
1999	5.18	4.92	0.275	0.786
1998	5.01	4.75	0.271	0.771
1997	4.93	4.68	0.271	0.768
1996	5.39	5.11	0.277	0.795
1995	5.06	4.80	0.270	0.771
1994	5.23	4.95	0.273	0.779
1993	4.71	4.44	0.267	0.745
1992	4.32	4.07	0.259	0.713
1991	4.49	4.24	0.260	0.724

American League Data (NL similar)



RBI's and Runs per game lower



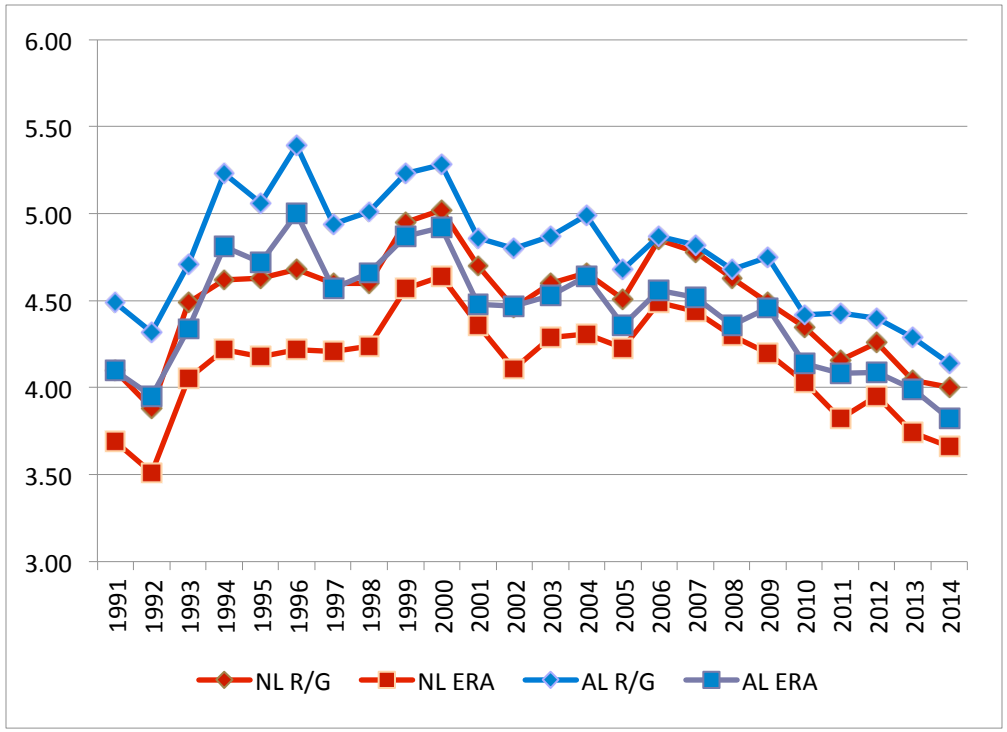
THE OHIO STATE UNIVERSITY
COLLEGE OF PUBLIC HEALTH

Are we in a new era?

PITCHING UP

Earned Run Average
1991-2009: 4.498
2009-2015: 3.987

Year	R/G	ERA	H	ER	WHIP
2015	4.23	3.94	8.40	3.91	1.276
2014	4.14	3.82	8.63	3.81	1.284
2013	4.29	3.99	8.76	3.98	1.318
2012	4.40	4.09	8.69	4.06	1.308
2011	4.43	4.08	8.78	4.06	1.325
2010	4.42	4.14	8.79	4.10	1.346
2009	4.75	4.46	9.13	4.41	1.403
2008	4.68	4.36	9.11	4.32	1.391
2007	4.82	4.52	9.28	4.46	1.412
2006	4.87	4.56	9.40	4.50	1.414
2005	4.68	4.36	9.13	4.31	1.362
2004	4.99	4.64	9.33	4.60	1.416
2003	4.87	4.53	9.21	4.48	1.385
2002	4.80	4.47	9.10	4.42	1.383
2001	4.86	4.48	9.20	4.44	1.391
2000	5.28	4.92	9.55	4.86	1.49
1999	5.23	4.87	9.51	4.80	1.486
1998	5.01	4.66	9.35	4.61	1.432
1997	4.94	4.57	9.40	4.53	1.443
1996	5.39	5.00	9.67	4.97	1.505
1995	5.06	4.72	9.30	4.67	1.467
1994	5.23	4.81	9.44	4.77	1.475
1993	4.71	4.34	9.11	4.29	1.418
1992	4.32	3.95	8.82	3.93	1.363
1991	4.49	4.10	8.90	4.10	1.37



American League Data (NL similar)

Evidence from Player Performance

HOME RUN LEADERS 2010 – 2014 (MLB)

2014	
1 Cruz (BAL)	40
2 Stanton (MIA)	37
Carter (HOU)	37
4 Abreu (CHW)	36
Trout (LAA)	36
6 Bautista (TOR)	35
Ortiz (BOS)	35
8 Encarnacion (TOR)	34
9 Martinez (DET)	32
Rizzo (CHC)	32

2013	
1 Davis (BAL)	53
2 Cabrera (DET)	44
3 Goldschmidt (ARI)	36
Encarnacion (TOR)	36
Alvarez (PIT)	36
6 Dunn (CHW)	34
Trumbo (LAA)	34
Soriano (2TM)	34
9 Jones (BAL)	33
10 Longoria (TBR)	32

2012	
1 Cabrera (DET)	44
2 Granderson (NYY)	43
Hamilton (TEX)	43
4 Encarnacion (TOR)	42
5 Dunn (CHW)	41
Braun (MIL)	41
7 Stanton (MIA)	37
8 Beltre (TEX)	36
9 Willingham (MIN)	35
10 Bruce (CIN)	34

2011	
1 Bautista (TOR)	43
2 Granderson (NYY)	41
3 Teixeira (NYY)	39
Kemp (LAD)	39
5 Fielder (MIL)	38
6 Pujols (STL)	37
Reynolds (BAL)	37
8 Uggla (ATL)	36
9 Stanton (FLA)	34
10 2 tied	33

2010	
1 Bautista (TOR)	54
2 Pujols (STL)	42
3 Konerko (CHW)	39
4 Dunn (WSN)	38
Cabrera (DET)	38
6 Votto (CIN)	37
7 Gonzalez (COL)	34
8 Uggla (FLA)	33
Teixeira (NYY)	33
10 Ortiz (BOS) + 3	32

Compare this to…

> 8 players with more than 40 HRs every season from 1996-2006

2006	
1 Howard (PHI)	58
2 Ortiz (BOS)	54
3 Pujols (STL)	49
4 Soriano (WSN)	46
5 Berkman (HOU)	45
6 Dye (CHW)	44
7 Thome (CHW)	42
Hafner (CLE)	42
9 Beltran (NYM)	41
Jones (ATL)	41

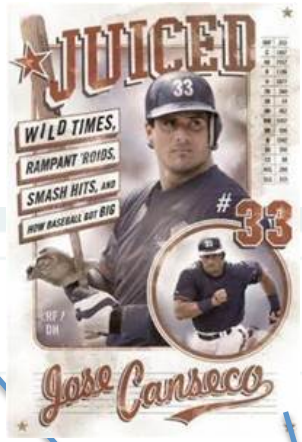
2001	
1 Bonds (SFG)	73
2 Sosa (CHC)	64
3 Gonzalez (ARI)	57
Rodriguez 4 (TEX)	52
5 Thome (CLE)	49
Helton (COL)	49
Green (LAD)	49
8 Palmeiro (TEX)	47
9 Sexson (MIL)	45
10 Glaus (ANA)	41
Ramirez (BOS)	41
Nevin (SDP)	41

1998	
1 McGwire (STL)	70
2 Sosa (CHC)	66
3 Griffey (SEA)	56
4 Vaughn (SDP)	50
5 Belle (CHW)	49
6 Castilla (COL)	46
Canseco (TOR)	46
8 Gonzalez (TEX)	45
Ramirez (CLE)	45
10 Galarraga (ATL)	44

1996	
1 McGwire (OAK)	52
2 Anderson (BAL)	50
3 Griffey (SEA)	49
4 Belle (CLE)	48
5 Gonzalez (TEX)	47
Galarraga (COL)	47
7 Buhner (SEA)	44
Vaughn (BOS)	44
9 Bonds (SFG)	42
Sheffield (FLA)	42

Can Tie to Events

Ephedra Tied To Pitcher's Death



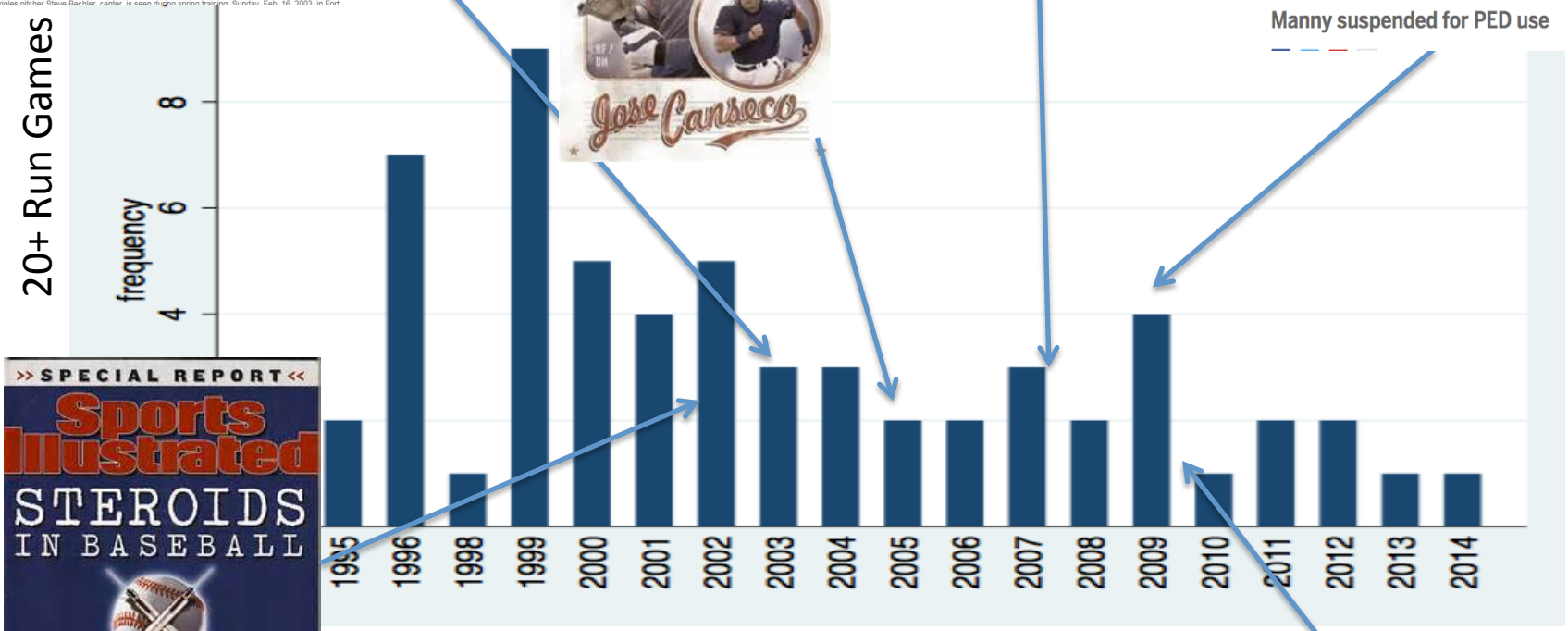
Mitchell report: Baseball slow to react to players' steroid use

12/14/2007 - MLB, MINNESOTA TWINS +29

Share with Facebook Share with Twitter

Manny suspended 50 games for PED use

Manny suspended for PED use



Biggest sports stories of the decade

Sports Illustrated

2000s: Top 10 Stories

By Lee Jenkins, SI.com

1 Baseball's steroid scandal

THE OHIO STATE UNIVERSITY

COLLEGE OF PUBLIC HEALTH

The “Post PED” Era (start in 2010)

- Kaplan-Meier estimates:

Length (seasons)	Era	Events	Median	CI	Mean
19	Dead Ball (1901-1919)	33	208	(81,335)	612.5
22	Lively Ball (1920-1941)	67	247	(195,365)	406.8
19	Integration (1942-1960)	37	345	(227,543)	650.1
16	Expansion (1961-1976)	12	1714	(340,3612)	2307.3
17	Free Agency (1977-1993)	21	1185	(367,1448)	1635.7
16	Long Ball (1994-2009)	54	442	(267,632)	718.8
	Post PED (2010-Current)	7	1234	(426, 2539)	1748.7

Reasonable time for an Era

Appears to differ from Long Ball Era

Post PED Model Hazard Ratios

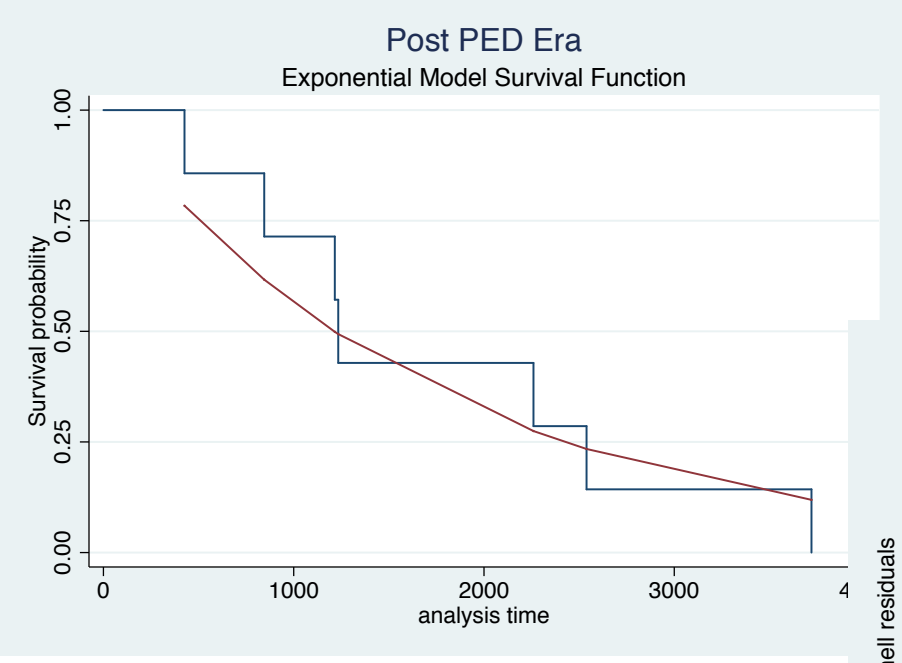
Hazard ratio (reference group Dead Ball) quite different than Long Ball:

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
era						
Lively Ball	1.505835	.3202454	1.92	0.054	.9925503	2.284558
Integration	.9421351	.225582	-0.25	0.803	.5892561	1.506338
Expansion	.2654646	.0894881	-3.93	0.000	.1371092	.5139804
Free Agency	.3744634	.1045297	-3.52	0.000	.2166709	.6471698
Long Ball	.8521401	.1882855	-0.72	0.469	.5526281	1.313981
Post PED	.350266	.1457545	-2.52	0.012	.15495	.7917797
_cons	.0016326	.0002842	-36.87	0.000	.0011607	.0022965

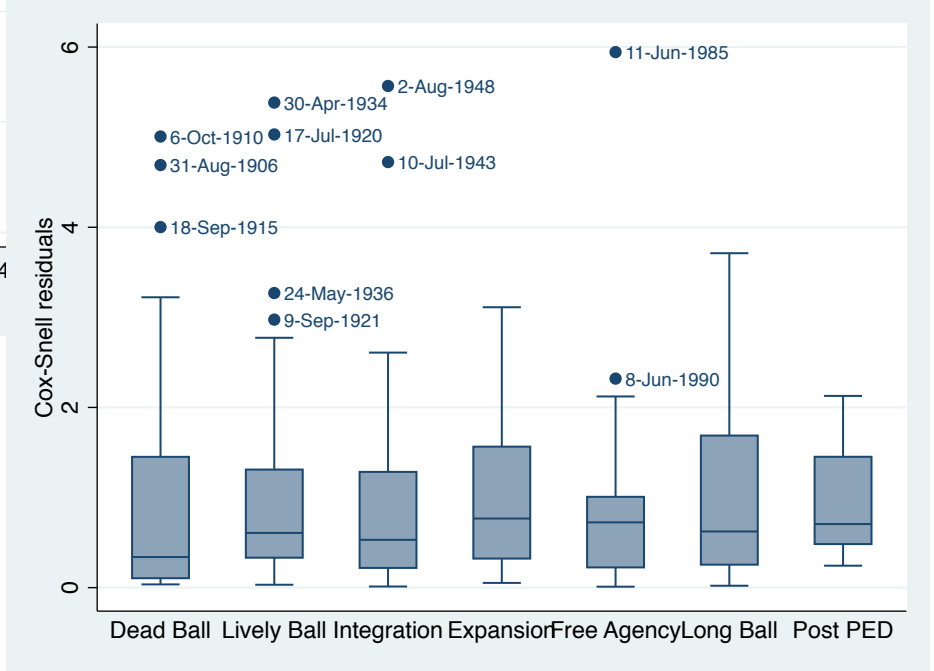
- Significantly higher hazard of 20+ run games in Long Ball than our "Post PED" era:

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Long Ball - Post PED	2.432837	.9773093	2.21	0.027	1.107062	5.346311

A New Era Dawns…



Not much data, but fit appears reasonable



Summary, Extensions, Future

- Choice of parametric form (or non-parametric) of the "baseline hazard"
- Easy to add other predictors to the model!
- Framework allows:
  - Ready model assessment
  - Interpretation and comparison of groups (hazard ratio, time ratios for exponential model)
  - Testing of significance and quantification of precision of estimates
- Other rare events: triple plays, perfect games, other sports
- A different approach: treat seasons as "subjects"
  - Censored if no event
  - Possibility of more than one event: "recurrent event" model
- Explore the "new Era" more; determine more rigorously where things changed

REFERENCES and SOURCES

- Hosmer, D.W., Lemeshow, S., and May, S. (2008), Applied Survival Analysis, Second Edition, Wiley, New York.
- Michael Huber and Andrew Glen, "Modeling Rare Baseball Events – Are They Memoryless?" *Journal of Statistics Education*, Volume 15, Number 1, 2007.
- Michael R. Huber and Rodney X. Sturdivant, "Building a Model for Scoring 20 or More Runs in a Baseball Game," *Annals of Applied Statistics,* Volume 4, Number 2, 2010.
- *www.baseball-reference.com*. Accessed May 2009, April/May 2015.
- *www.retrosheet.org*. Accessed May 2009.
- *www.netshrine.com/era.html*. Accessed June 2009.
- R. D'Agostino and M. Stephens. *Goodness-of-Fit Techniques*,. Marcel Dekker, New York: 1986.