Modelling the Lifetime of Toronto Blue Jays Games by Making Use of Parametric Log Location Scale Models

February 04, 2016

Abstract

Considering the popularity of watching baseball games among Nourth American people, especially for those of Toronto Blue Jays (TBJ), we are interested in doing a survival analysis on the length of the games of TBJ and see the effects of other covariates such as innings, hits, etc, on the length of the games. Different parametric models like Weibell, lognormal and loglogistic have been studied for obtaining an efficient proportional hazard model, while we did variable selection and observed the residual behaviours of such models. Finally, by graphical model checking, the lognormal was has been selected as the best parametric model and its relevent significant covariates are showed up.

1 Introduction

This past season the TBJ broke several records for viewership, with a couple of games breaking the four million mark. For reference, not one Stanley Cup Playoff Game could compete with this number. In fact, the only hockey game to see more viewers was the World Junior Hockey Championship final. The popularity of major league baseball, especially in Toronto and the surrounding areas, is undeniable. As two people new to this area and to watching this sport, we were interested to consider the liftetime of TBJ games.

Baseball is one of very few sports to not have a game clock. This past season, the Jays had a game just 1 hour and 54 minutes long and another one of 4 hours and 32 minutes long. Most games are over in 9 innings, although the number of Innings ranged ebtween 8.5 and 12 last season. We are not motivated to build a model for predicting the length of games (as the data we have only comes with having the length of a game), but rather we interested in investigating more closely how other factors affect the length of games.

2 The Data

The Jays played a total of 162 games last season. Data for each of these games was collected from mlb.com. For each observation, we have 18 variables including the length of the game, the number of innings, runs, hits, walks (or base on balls), stolen bases, opposing errors, strike outs thrown, runs against, hits against, walks against, errors, and number of Jays struck out. Additionally, we have the date of each game, the opposing team, whether it was a home or away game, and whether the Jays won or not. We did not anticipate using all of these variables (such as Date, Versus, Away, and Win), but collected these variables in an effort to be complete and thorough. The question of whether temperature could play a role in outliers or influential cases was raised. The following is a short representation of the dataset, which including the lifetime objective (time) as well as 15 listed covariates.

```
'data.frame':        162 obs. of  16 variables:
 $ time          : num  179 176 172 165 154 197 150 186 182 180 ...
 $ Away          : Factor w/ 2 levels "Away","Home": 1 1 1 1 1 1 2 2 2 2 ...
 $ Innings       : num  9 8.5 9 9 8.5 9 9 9 8.5 9 ...
 $ Temp          : int  54 44 42 51 61 60 68 68 68 68 ...
 $ Win           : Factor w/ 2 levels "Loss","Win": 2 1 2 2 1 2 1 1 2 1 ...
 $ Runs          : int  6 3 6 12 1 10 1 2 12 2 ...
 $ Hits          : int  6 7 9 16 3 10 2 8 14 4 ...
 $ Walks         : int  5 3 0 3 1 6 4 1 5 4 ...
 $ StolenBases   : int  2 0 0 0 1 1 0 0 0 0 ...
 $ OpposingErrors: int  1 2 1 1 0 0 0 0 0 0 ...
 $ StrikesThrown : int  5 7 10 2 8 9 8 11 7 8 ...
 $ RunsAgainst   : int  1 4 3 5 7 7 2 3 7 4 ...
 $ HitsAgainst   : int  3 7 7 13 9 8 3 7 11 8 ...
 $ Walked        : int  3 4 4 2 6 2 6 5 4 3 ...
 $ Errors        : int  0 0 0 1 1 0 0 1 1 1 ...
 $ StruckOut     : int  12 7 10 6 10 5 4 4 5 12 ...
```

Before attempting to fit a model, some basic exploratory data analysis was performed. A scatter plot of the numerical variables suggest that there is a relationship between Time and several of the other variables including Innings, Hits, Walks, HitsAgainst, Walked, and StruckOut. The variables Hits and Runs as well as HitsAgainst and RunsAgainst appear to be strongly correlated, which was to be expected, suggesting that the final model should not include either pair. Although this is perhaps irrelevant to our analysis, the Time variable was found to be approximately normal.

3 Model Fitting

First, the Kaplan-Meier and Nelson-Alan estimators of "time" variable are estimated to see the performances of nonparametric models applied on the time

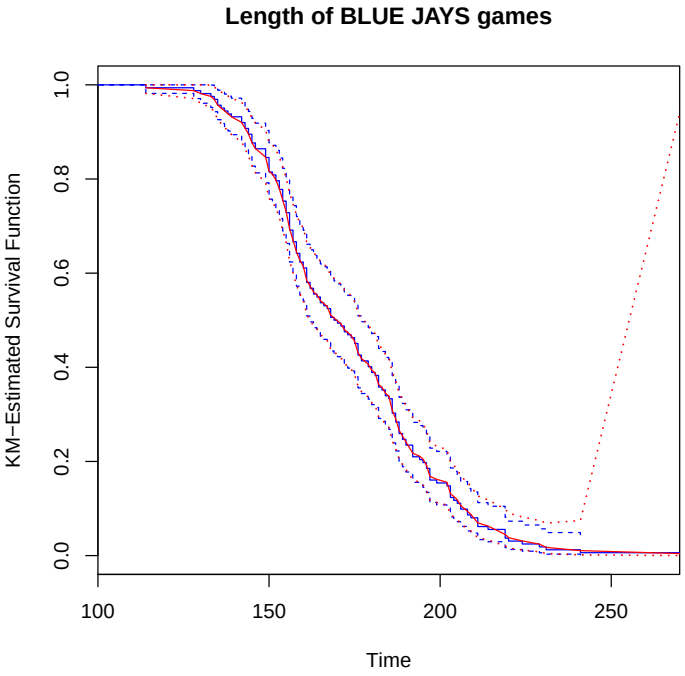

Figure 1: KM (blue) and NA (red)-Estimated Survival Function

while the summary of the KM estimator is as follows:

```
Call: survfit(formula = surv.time ~ 1)

   records       n.max     n.start      events      *rmean *se(rmean)      median
       162         162         162         162         173           2         170
   0.95LCL     0.95UCL
       162         177
    * restricted mean with upper limit =  272
```

So, the median length of the games was 170mins which is very close to the mean survival time of 173mins of the games. The following, Figure 1, illustrates the behaviour of KM and NA survival functions. There is a strong consistency between the two estimators, possibly because we have no censorred observation. It is seen, from Figure 1, between time 150 and 200, the survival function drops down sharply while it fall more smoothly out of this range. By having a look at the corresponding cumulative hazard functions of these two nonparametric estimators, we see that the hazard rate has an increasing, almost linearly, pattern after time 150, see Figure 2.

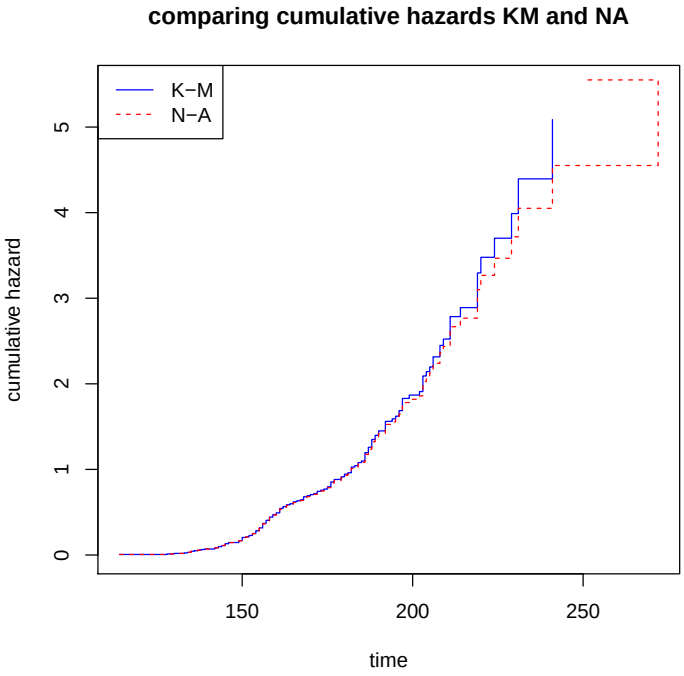

Figure 2: KM (blue) and NA (red)-Estimated Survival Function

Now, we are interested in finding out a good parametric survival model which fits more to the data. For this aim, the following plots (see Figure 3), based on the estimated KM survival function, are illustrated to check the suitability of using Weibull, lognormal, or loglogistic distributions for fitting parametric models to the data.

The most linear relationship appears to be with the lognormal distribution, although the logllogistic is a close second. For this reason, we compare the two. Now, we fit the full model (using all covariates excluding Date, Win, Away, and Versus) with a lognormal distribution. The outputs are shown below.

```
# fit on all covariates of interested
> lognorm=survreg(surv.time~., data=AB, dist="lognormal")
> # variable selection by step function based on BIC
> step(lognorm,k=log(length(time)), trace=0)

Call:
survreg(formula = surv.time ~ Innings + Hits + Walks + HitsAgainst +
    Walked + StruckOut, data = AB, dist = "lognormal")

Coefficients:
(Intercept)     Innings        Hits       Walks HitsAgainst      Walked
3.986286418 0.086321397 0.012025803 0.022667698 0.011065960 0.020401449
  StruckOut
0.008424535

Scale= 0.08624595

Loglik(model)= -666.2   Loglik(intercept only)= -749.5
        Chisq= 166.65 on 6 degrees of freedom, p= 0
n= 162
```

Obviously, many of the variables were insignificant in the full model. Using the step function and BIC criteria, stepwise variable selection was performed. The following is the summary of fitting a lognormal model to the selected covariates. It is seen that all the coefficients are highly significant. It meants that all these covariates has significant effects on the time of the games, i.e., by changing their corresponding values, the length of the games may change sensibly.

```
Call:
survreg(formula = surv.time ~ Innings + Hits + Walks + HitsAgainst +
    Walked + StruckOut, data = AB, dist = "lognormal")
              Value Std. Error      z        p
(Intercept)  3.98629    0.10469  38.08 0.00e+00
Innings      0.08632    0.01212   7.12 1.06e-12
Hits         0.01203    0.00195   6.18 6.30e-10
```

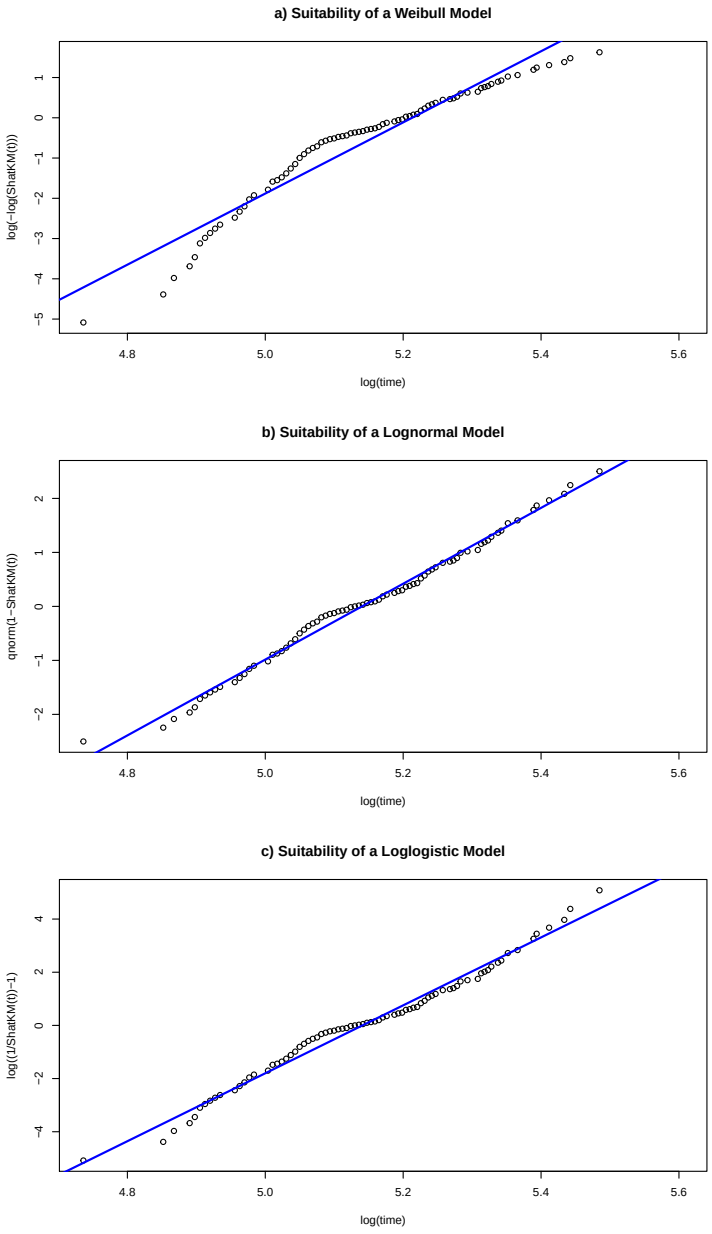

Figure 3: Graphical model checking for parametric models

```
Walks         0.02267    0.00326   6.95 3.77e-12
HitsAgainst   0.01107    0.00208   5.31 1.07e-07
```

```
Walked       0.02040    0.00376   5.43 5.60e-08
StruckOut    0.00842    0.00257   3.28 1.03e-03
Log(scale)  -2.45055    0.05556 -44.11 0.00e+00

Scale= 0.0862

Log Normal distribution
Loglik(model)= -666.2   Loglik(intercept only)= -749.5
        Chisq= 166.65 on 6 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 162
```

As we are interested in comparing "lognormal" and "loglogistic" survival models, by repeating the same process for fitting a loglogestic model as well as the variable selection similarly, the summary of the outputs for the "loglogistic" model is as follows:

```
Call:
survreg(formula = surv.time ~ Innings + Hits + Walks + HitsAgainst +
    Walked + StruckOut, data = AB, dist = "loglogistic")
              Value Std. Error      z        p
(Intercept)  3.9749    0.10307  38.56 0.00e+00
Innings      0.0884    0.01192   7.42 1.17e-13
Hits         0.0125    0.00200   6.23 4.77e-10
Walks        0.0219    0.00334   6.55 5.65e-11
HitsAgainst  0.0103    0.00215   4.81 1.49e-06
Walked       0.0201    0.00381   5.28 1.31e-07
StruckOut    0.0083    0.00250   3.32 8.94e-04
Log(scale)  -2.9879    0.06463 -46.23 0.00e+00

Scale= 0.0504

Log logistic distribution
Loglik(model)= -670   Loglik(intercept only)= -752.5
        Chisq= 165.08 on 6 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 162
```

Using "survreg" function, we note that both models contain the same predictors and similar estimates for the intercept, coefficients, and scale parameter. The p-values are all very small. We take a closer look at how these two models are performing. The residual plots of these two models are shown in Figure 4. They both follow the $exp(-x)$ curve reasonably well, leaving no cause for concern. We see a little less departure from the curve for the lognormal model. Below we see the residuals plotted versus the covariate coefficients, intercept

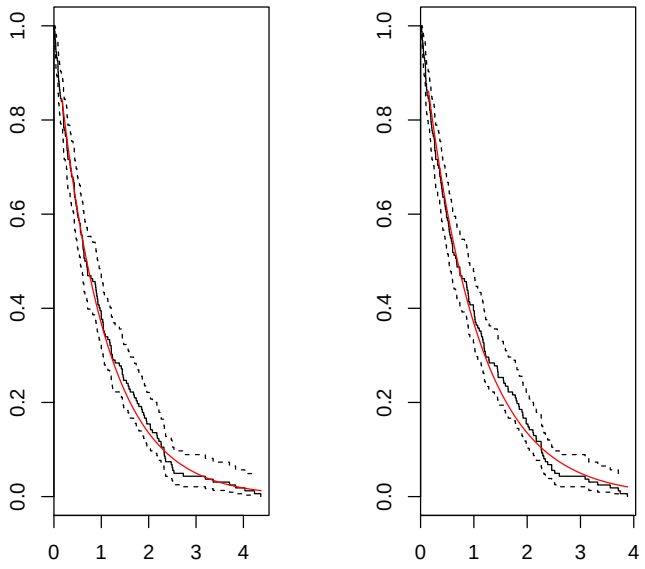

Figure 4: Residual plots for the lognormal (left) and loglogistic (right) models.

and the slop for each model seperately. Each graph shows the effect of removing each observation on that parameter, see Figures 5, 6.

Although both models appear acceptable (with the absolute values all being less than 0.4), the lognormal model seems to have slightly smaller values overall.

Next, the cox proportional hazard model based on both "breslow" and "efron" methods applied on all the covariates and then on the slelected covariates were tested. In both cases, only the selected covariates were highly significant. The following is the summery of "coxph" on the selected covariates fitting based on the "efron" method.

```
Call:
coxph(formula = surv.time ~ Innings + Hits + Walks + HitsAgainst +
    Walked + StruckOut, data = AB, method = "efron")

  n= 162, number of events= 162

               coef exp(coef) se(coef)      z Pr(>|z|)
Innings     -0.91480   0.40060  0.16040 -5.703 1.18e-08 ***
Hits        -0.14745   0.86291  0.02477 -5.953 2.63e-09 ***
```

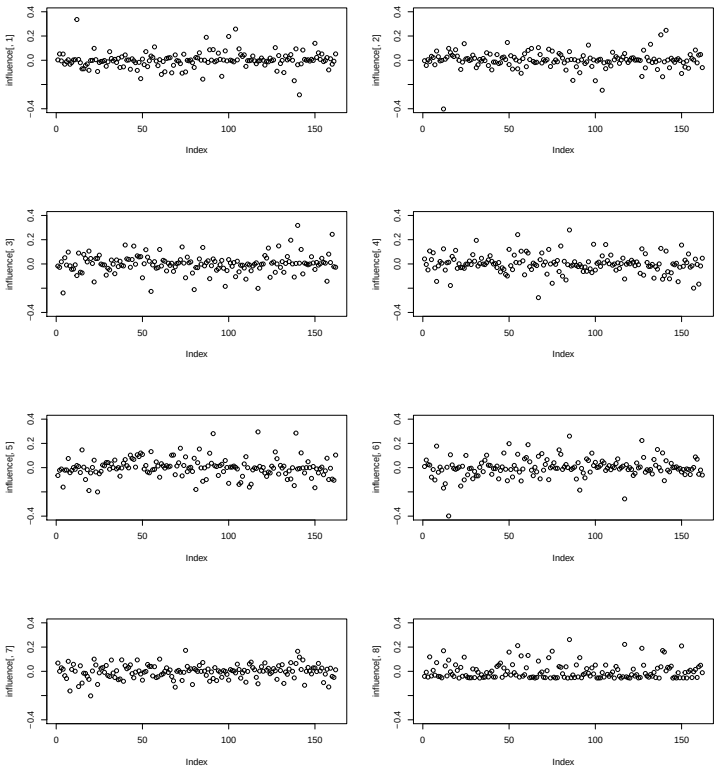

Figure 5: Residual plots versus estimated parameters for lognormal distribution model to distinguish influential observations.

```
Walks       -0.26774   0.76511  0.04478 -5.979 2.25e-09 ***
HitsAgainst -0.15961   0.85247  0.02662 -5.995 2.03e-09 ***
Walked      -0.23397   0.79138  0.04789 -4.886 1.03e-06 ***
StruckOut   -0.12049   0.88648  0.03547 -3.397  0.00068 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
Innings        0.4006      2.496    0.2925    0.5486
Hits           0.8629      1.159    0.8220    0.9058
Walks          0.7651      1.307    0.7008    0.8353
HitsAgainst    0.8525      1.173    0.8091    0.8981
Walked         0.7914      1.264    0.7205    0.8693
StruckOut      0.8865      1.128    0.8270    0.9503
```

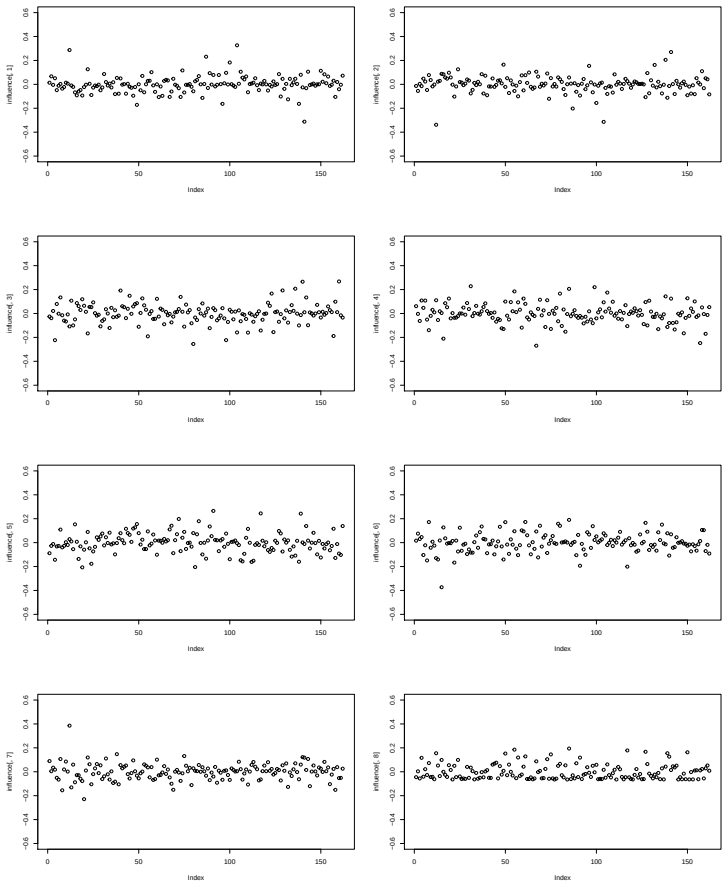

Figure 6: Residual plots versus estimated parameters for loglogistic distribution model to distinguish influential observations.

```
Concordance= 0.793  (se = 0.027 )
Rsquare= 0.619   (max possible= 1 )
Likelihood ratio test= 156.2  on 6 df,   p=0
Wald test            = 118.1  on 6 df,   p=0
Score (logrank) test = 125.3  on 6 df,   p=0
```

Here we are interested in testing PH assumption. As it is seen from the numerical results, the p-values of almost all the covariates are greater than 0.05, which means there is no conflict for incloding those covariate in our cox PH model.

```
> cox.zph(inter.coxph)
```

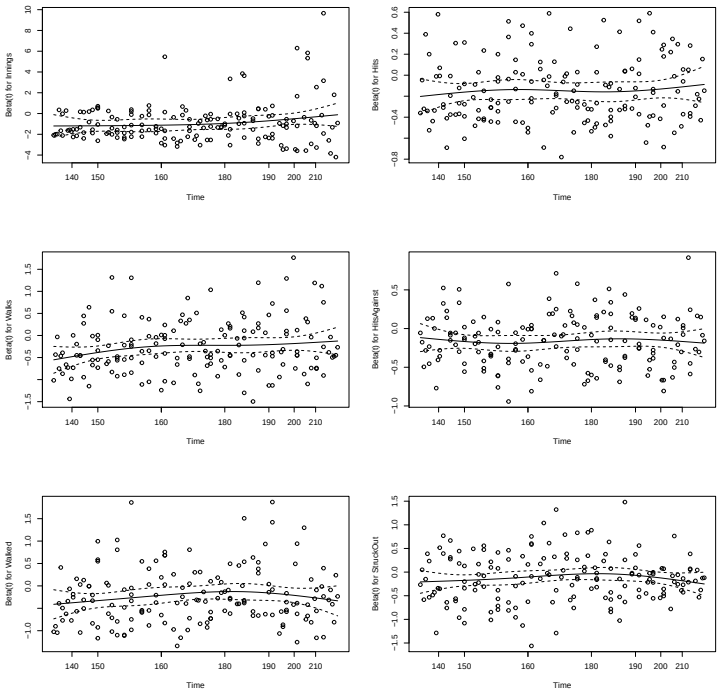

Figure 7: Testing COX PH assumption for each covariate.

	rho	chisq	p
Innings	0.14672	3.174	0.0748
Hits	0.05312	0.417	0.5182
Walks	0.15580	4.348	0.0371
HitsAgainst	0.00798	0.010	0.9201
Walked	0.07635	0.974	0.3236
StruckOut	0.03418	0.245	0.6204
GLOBAL	NA	7.847	0.2495

The visualized results of the effects of each observation on the covariates based on the cox PH model are illustrated, in Figure 7, to see the influential observations.

4 Discussion

The residual and influence plots both suggest the lognormal distribution provides a slightly better fit.

It is widely believed that the starting pitcher has a large impact on the length

of a baseball game. We did not have the data to consider this here, but it would be interesting in future to see if this belief holds true.

5 Conclusion

After doing such analysis, the lognormal is preferred for fitting on the data, however, the performance of th loglogistic model was acceptable as well. It means that a good analysis for this sort of data would be obtained by considering a parametric lognormal distribution.

References

[1] http://web.stanford.edu/class/stats191/notebooks

[2] The Lecture notes of professor Jones.

[3] https://ca.sports.yahoo.com/blogs/eh-game/blue-jays-continue-to-set-records-for-sportsnet-024554485.html.

[4] mlb.com.