SURVIVAL ANALYSIS

FINAL PROJECT

NBA Players' Time to Retirement

Author:
Che-shun CHANG, Minzhao LIU , Xuan LUO

Dec. 2, 2010

Contents

Abstract

Basketball is always one of the most popular sports. In other words, the economic effect is what we cant ignore. The most important property for those owners should be the players. How to evaluate those players is the issue which we want to figure out. We try to use all the information we have to predict their mean residual life which means how long they can still play, and to find out which factors are important to improve players' career lifespan.

We use weibull, log-logistic, lognormal, and coxph model to fit the model with stepwise method. For each method, we check if the model is fitted by Nelson-Aalen estimator and assess linearity, Cox-Snell residuals, deviance residuals, and score residuals. Proportional assumption is also verified.

For the sign of the coefficients, we can see what is of benefit to a players career length. Base on our assumption, we are able to get an estimation of mean residual life. It will be a useful indicator for the team owners to offer contracts to players. Also the method may help players to decide their training plan, in order to extend their career length.

Key words: nonparametric model, parametric model, mean residual life.

1 Introduction

1.1 Motivation

Through this article, we want to fit a survival model for the NBA players' career lifespan. The event could be retirement, laid-off or whatever cause player leave the professional career. A useful survival model was expected to provide us explanatory factors, accurate prediction, and to help team managers to make contracts with players, in terms of salary and length, and for players deciding training plan and extending career lifespan.

1.2 Responces and Censor

In National Basketball Association, players enter the league when they finish their college or high school study. After few rookie years, they could be traded to other team, get laid off, have a longer and better contract, or retire. Their professional career lifespan is the main goal we want to predict, thus how many years they played as professional players is regarded as our responce. In the data, the difference between their last season and first season is taken as their 'life time'. This could be problematic since some players could quit and came back later. For example, Micheal Jordan went to MLB for certain years, but in our settings, we still take the difference between last and first season as his 'life span'. Models for recurrent events are good approaches dealing with this kind of situation, which is not covered in this article.

The reason we fit the survival model is that it deals with censored data well. In our scenario, active players are regarded as censored observations, since their rest career life is unknown.

We do not know how many seasons they could still play in the league. And obviously, the censor system can be thought independent with the event, which ensures us fitting those various kind of survival models.

1.3 Covariates Description

The dataset is from www.databasebasketball.com, and totally we have 3435 players from 50s upto 00s. The data could be divided into two parts, one is personal information, and the other is about players' performance on court.

For personal information, individual weight and height are recorded. Players' position could be center, forward, and guard, which is a three-level categorical variable. In addition, the age when they became professional athletes is also a potential factor. Also, we expect the decade (50s, 60s, etc) they entered the league might be an important explanatory for their survival life. However, athletes may change their positions, and their career life may cross two decades. Thus strictly speaking, those variables may be time dependent.

As to their performance, the total summary statistics, such as total points, total minutes, etc, are quite highly correlated with life time. The more seasons they play, the more statistics they get. Hence, to avoid time dependent variables, we take average statistics instead of those total statistics. For instance, how many games they attended per season, how many points they grabbed per game, etc. Also, shooting percentages are adopted rather than average shoots per game.

1.4 Missing Value

Due to historic problems, some statistics had not been used before. Minutes, offense rebounds are such kind of numbers not being recorded at that time. Another thing needs to mention is that althougn in the dataset, players' shooting percentage might be zero, this could be a missing value because it's due to his absence in shooting, it does not mean his shooting percentage is rather low.

2 Model Fit and Variable Selection

2.1 Semi Parametric and Parametric Approaches

When fitting the survival model, we try to use cox proportional hazard model as semi-parametric approach and later, several parametric models, say Weibull, Log-logistic, Log-normal models, are applied to the data.

Cox proportional hazard model makes strong assumption for the proportional hazard part, but makes no assumption on the baseline hazard function.

$$h(t|\boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z})$$

By partial likelihood method, estimates could be obtained.

While for parametric model, accelerated failure time (AFT) models are used at cost of additional parametric likelihood assumptions.

$$S(t|\boldsymbol{Z}) = S_0\left(t\exp\left(\boldsymbol{\theta}'\boldsymbol{Z}\right)\right)$$

Estimates can be solved through maximum likelihood method, since every thing is known if you assume the likelihood.

$$Lik = \prod_{i=1}^{n}\left(f\left(t_i\right)\right)^{\delta_i}\left(S\left(t_i\right)\right)^{1-\delta_i}$$

2.2 Variable Selection

Stepwise procedure is a very easy and convenient way to deal with variable selection. In our scenario, for each model, we ran the forward procedure and kept 0.05 level to choose the final 'best' model. Here's the results:

cox PH	decade, position, age, game, minutes, assists, steals, blocks, turnover, field goal, free throw, 3 points, height
Weibull	decade, position, age, game, minutes, assists, steals, blocks, turnover, field goal, free throw, 3 points, weight, foul, points
Log-logistic	decade, position, age, game, minutes, assists, steals, blocks turnovers, field goal, free throw, 3 points, height, foul
Log-normal	decade, position, age, game, minutes, assists, steals, blocks turnovers, field goal, free throw, 3 points, height, foul

From the table above, we can see the final results by stepwise procedure are slightly different with each other. The semiparametric way includes height as a significant variable, while parametric models generally provide more terms. One explanation could be the high correlation between height and weight, points, fouls, and games per season variables.

3 Diagnostics

3.1 Overall Fit

To check overall fit, use Cox-Snell residuals. If the model is correct, cox-snell residuals should look like a censored sample from a unit exponential. Here's the overall fit check by means of cox-snell residuals.

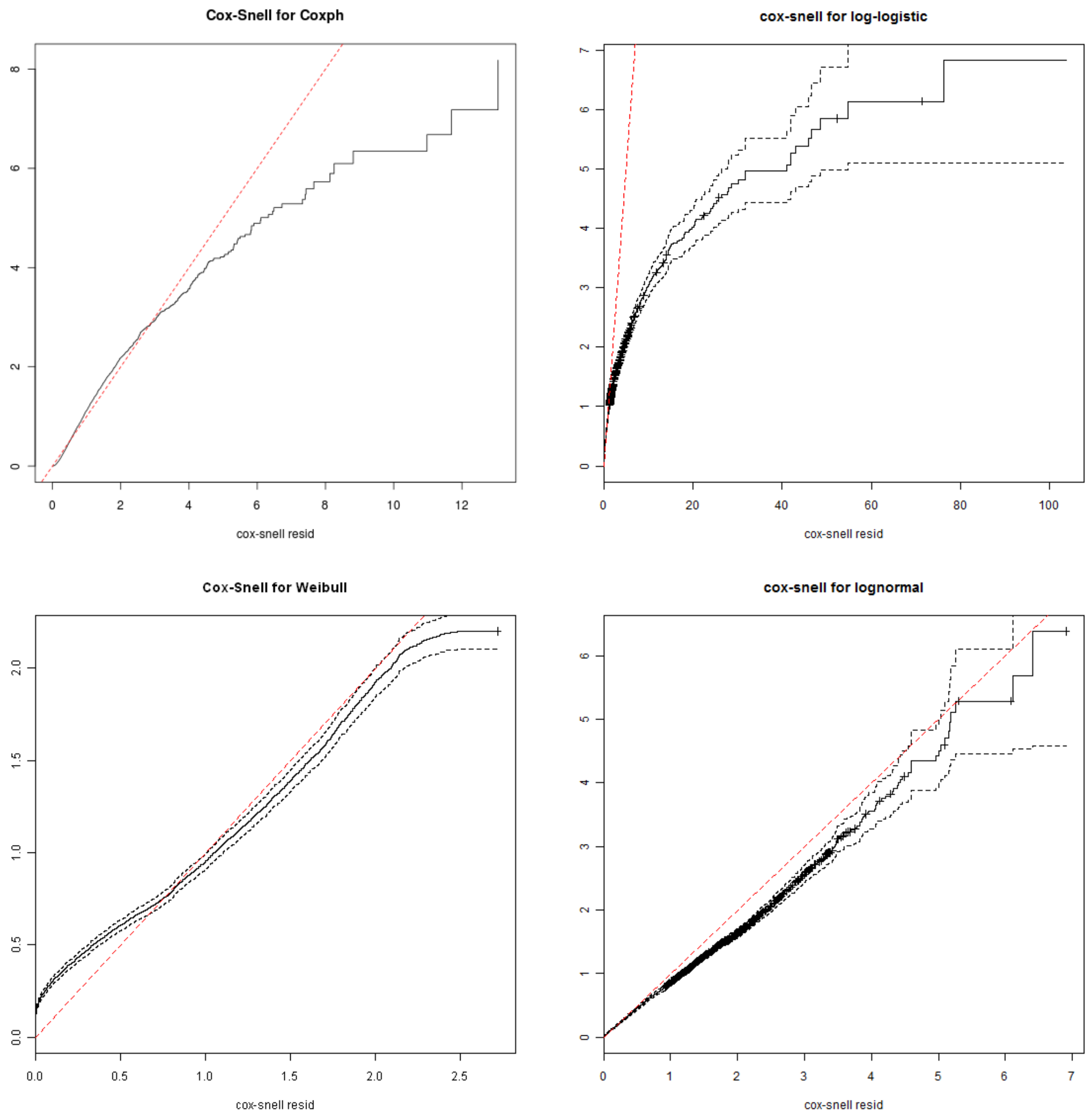

Figure 1: Cox-Snell Residuals For Coxph, Weibull, Loglogistic, Lognormal Models

From the plots, it seems that weibull distribution and log-normal distribution fits the model well, and log-logistic model is really bad for the model, with cox regression model lies in the middle.

3.2 PH Assumption Check (for Coxph)

Here, for categorical variables, graphical check is used. Fitting stratified cox model, $\hat{H}_{0g}(t)$ should be parallel if PH assumption is OK.

In our model, athlettes playing in different decades may behave different. We may check the proportional hazard assumption on the 'decade' variables.

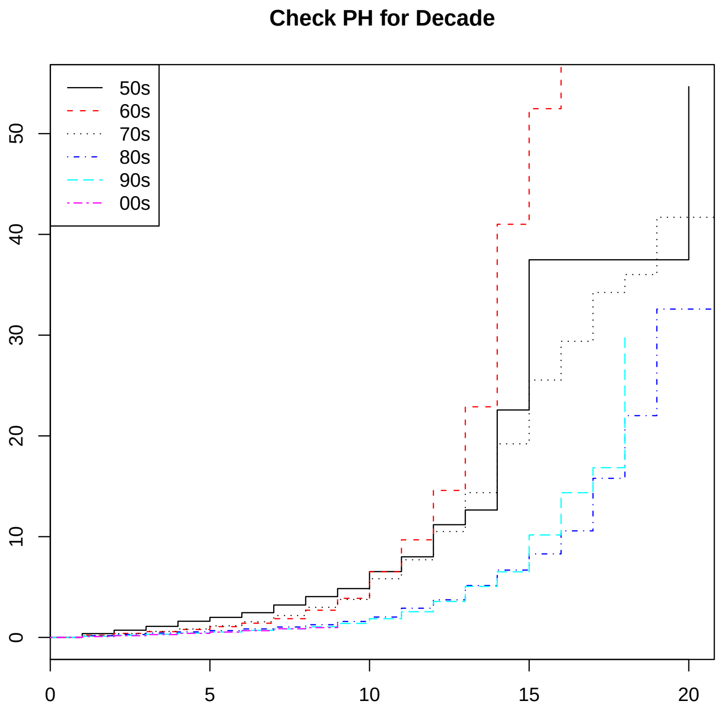

Figure 2: PH Check for Decade

Seen from Figure 2, cumulative hazard functions for different strata could be regarded as parallel. Thus the PH assumption could hold for 'decade' covariate.

For the same reason, we need to check PH assumption for 'position' variable.

By similar inference, we can also keep the assumption for 'position' variable.

As to the continuous variables, simply use *cox.zph* function to see if the assumption is accepted.

Since there are too many explanatory variables in our final model, presenting the 'coxzph' figure would be a mess, so we put the plot in the appendix. Anyhow, fortunately no continuous variables seem to be violate the proportional hazard assumption. See Figure 6 and 7 for detail.

3.3 Functional Form for Covariates

Functional form for covariates could be found by ploting martingale residuals when fitting without certain covariate. If the 'lowess' line presents discretization or quadratic pattern, we

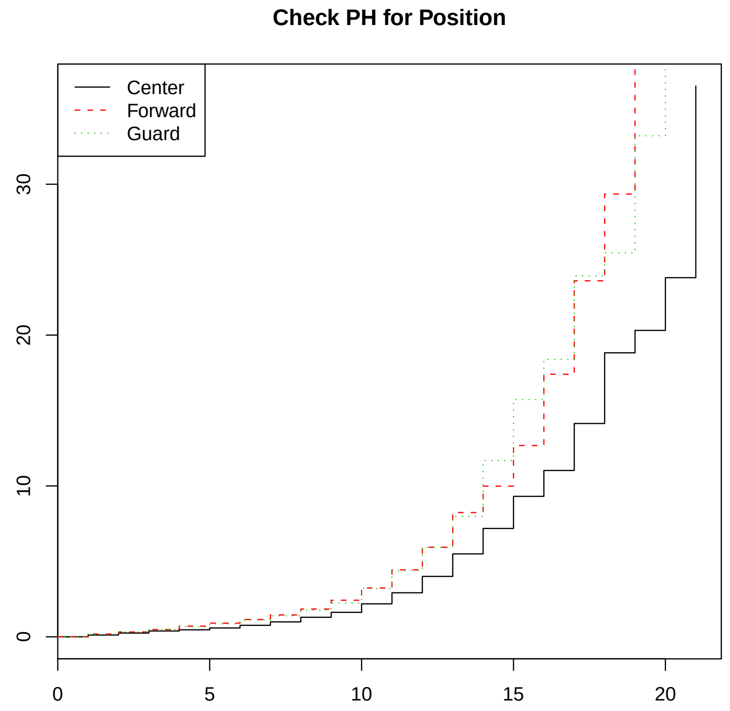

Figure 3: PH Check for Position

need to assign those kinds of pattern to that variable.

From Figure 4, we can see from these plots, there are no obvious pattern for the redline, they are almost flat, so we decide not to transform any covariates, just use the linear form.

3.4 Deviance Residuals and Outliers

Deviance residuals make more sense than martingale residuals since the latter ones are highly skewed. Deviance residuals tend to shrink large negative values. By looking for large values, we might have a brief idea about identities of outliers.

From Figure 5, it seems from the plots that we have a lot of outliers here, while considering we have a large sample size over 3000, this result might not be surprising. The weibull distribution fit and the cox model fit seem to have really similar result. And the log-logistic and log-normal model fit have similar outliers. Taking the weibull distribution fit as example, two obvious outliers are observaton 650 and 2897, both having deviance residuals exceeds 4. For observation 650, he played 4 years while the model predicts him to quit at 1.2 years; and for observation 2897, he played 8 years while the model predicts him to have a 2-year career life.

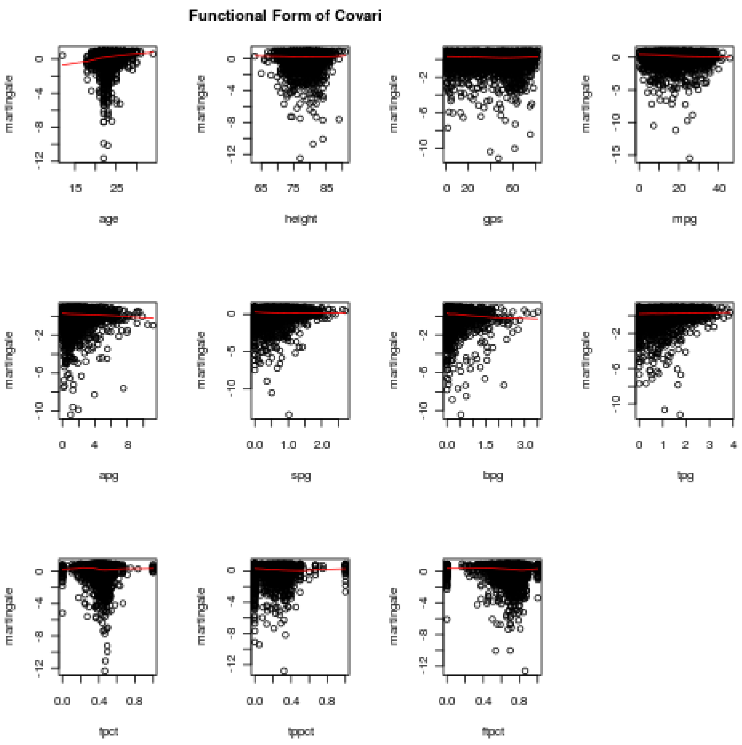

Figure 4: Functional Form Check

3.5 Dfbeta and Influence

Since we have lots of covariates, check dfbeta for each covariate is not feasible nor useful. Use likelihood displacement to check overall effect on the β vector.

From Figure 10, again, the cox model and the weibull model fit have similar results, both have obs 650 as influential one. It has biggest linear predictor, so we expect him to quit career life quickly, however, he played a relatively long time. Hence it impose large influence on the model fit.

4 Inference and Interpretation

4.1 Position Effect

From Figure 8, we can clearly find that players have a longer career if they play center instead of other two positions.

We believe that they are harder to be replaced, and there is less job competition at the position.

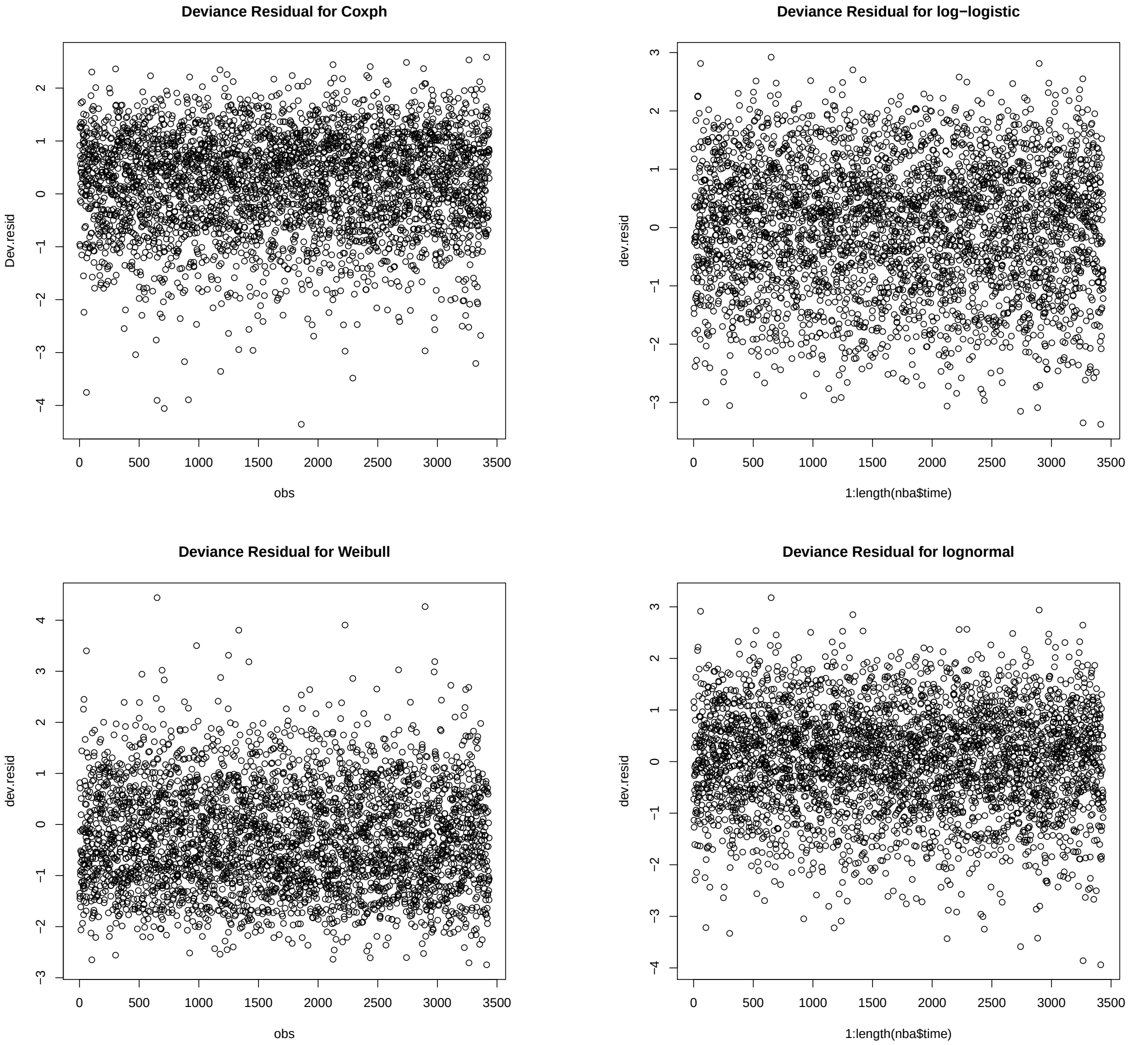

Figure 5: Deviance Residuals for Coxph, Weibull, Loglogistic, Lognormal models

4.2 Decade Effect

By our model, We learned that there is significant change in decades. Medical Technology is getting better and better. It helps players to maintain their status a lot. If they get hurt, they will have better treatment in the present years. See Figure 9 for detail.

4.3 Interpretation Coefficients

By the sign of coefficients, we can infer that the better they play, the longer they stay. The only tricky one is personal foul. It is surpriced us that the more personal foul they made, the lower survival risk they have. We would like to believe that it means the player devote themselves to the game really hard. All the defense cause conflicts, and that is really hard to avoid personal foul.

Fortunately, we can see that turn over is still a negative factor to their career, though it has high correlation with personal foul. It helps us to believe personal foul is meaningful. See Table 1 for detail.

4.4 Predict Mean Residual Life

We have more similar results between Weibull and Log-normal. Here's some examples from which coxph model seems to be relatively conservative. See Table 2 for detail.

5 Discussion

Due to the difficulties, we ignore the recurrent events, how we definde their career life is taking the last season substract the first season. It might cause overestimate of their career life. However, we think it's a more general defination of their career life. Nevertheless, by common sense, we know there are only few such kind of cases, which means we can nearly ignore that.

In addition, from our data, we don't have any details about if they change their position or not. Unfortunately, we are not able to fix it.

There are much difference between 40's and now. We can't really expect what's going to change in the future. The rules, equipments, training method, and medical tech might break through. Then the new model can be very different. Also, if player is just getting better, we will underestimate his career life because of his bad history. The time-dependent variables survival model might fit much better, with the cost of complicated assumptions.

6 Reference

www.databasebasketball.com

Term	Coxph	Weib	LN	Term	Cox	Weib	LN
minute	-0.099	0.058	0.051	3 points	-1.112	0.73	0.63
decade.f5	0.629	-0.42	-0.39	positionF	0.231	-0.15	-0.078
decade.f6	0.612	-0.33	-0.33	positionG	0.307	-0.22	-0.12
decade.f7	0.712	-0.38	-0.37	steal	-0.416	0.21	0.23
decade.f8	0.020	-0.019	-0.08	assist	-0.129	0.065	0.098
decade.f9	-0.051	0.058	0.05	game	-0.004	0.002	0.0047
block	-0.754	0.28	0.34	height	-0.027	NA	0.022
age	0.116	-0.066	-0.07	foul	NA	0.068	0.043
field goal	-1.382	1.026	0.75	point	NA	-0.014	NA
free throw	-0.615	0.51	0.63	weight	NA	0.0015	NA
turnover	0.451	-0.29	-0.28				

Table 1: Coefficients from Three Models

Name	Played	Coxph	Weibull	Log-normal
Nash	14	3.52	6.39	6.49
Love	2	9.68	11.14	9.46
Tmac	13	7.7	8.63	8.96
Rondo	4	9.73	11.88	11.75
Yao	8	10.7	12.04	12.29
Kobe	14	5.88	7.56	7.81
Shaq	18	4.25	4.66	4.64
Duncan	13	7.8	8.68	9.05
Ariza	6	5.9	7.22	6.86
Aaron Brooks	3	5.58	5.07	4.29
A.Iverson	14	5.29	7.04	7.27
Juwan Howard	16	2.34	4.98	4.96

Table 2: Predicted Mean Residual Life from Three Models for Sample Players

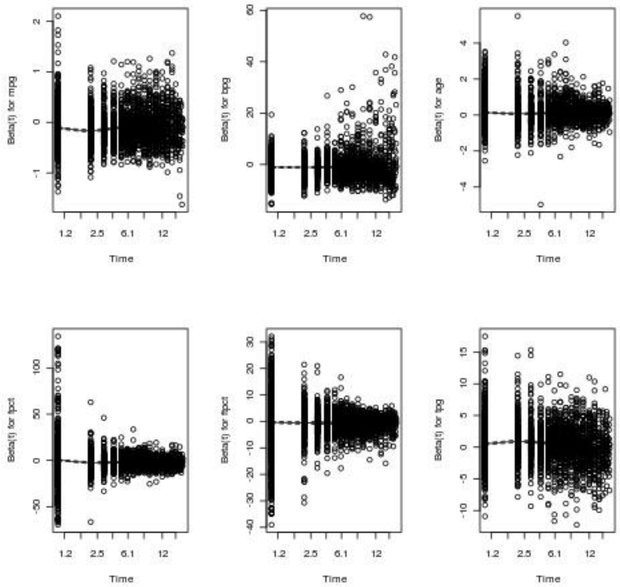

Figure 6: Test PH 1

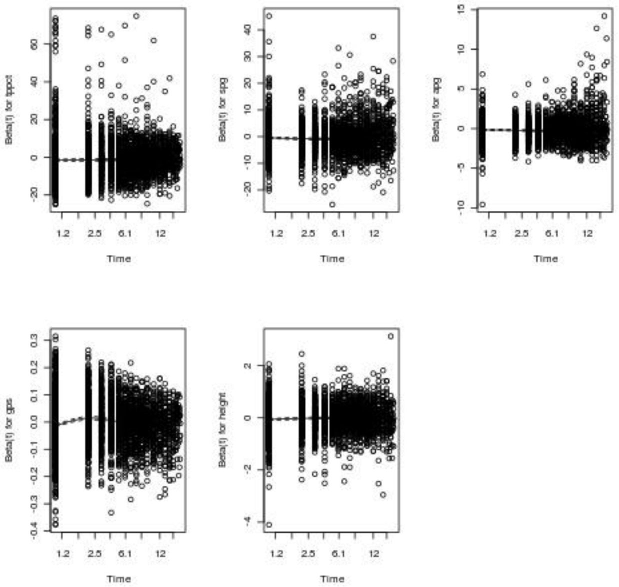

Figure 7: Test PH 2

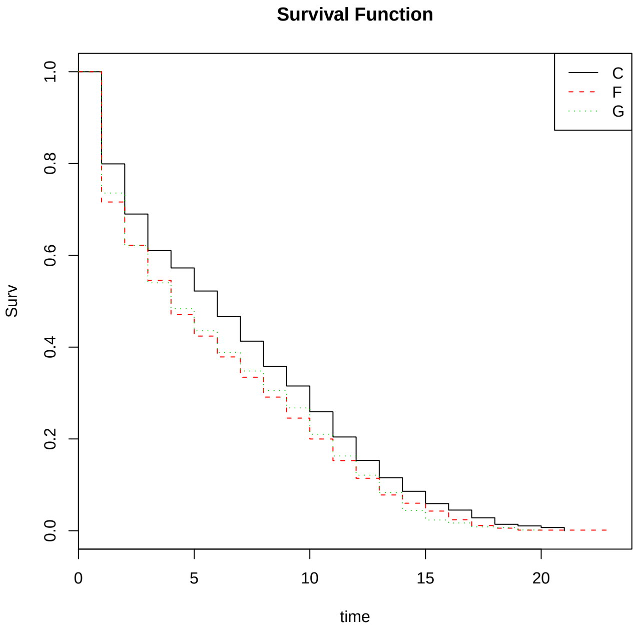

Figure 8: Strata for Position

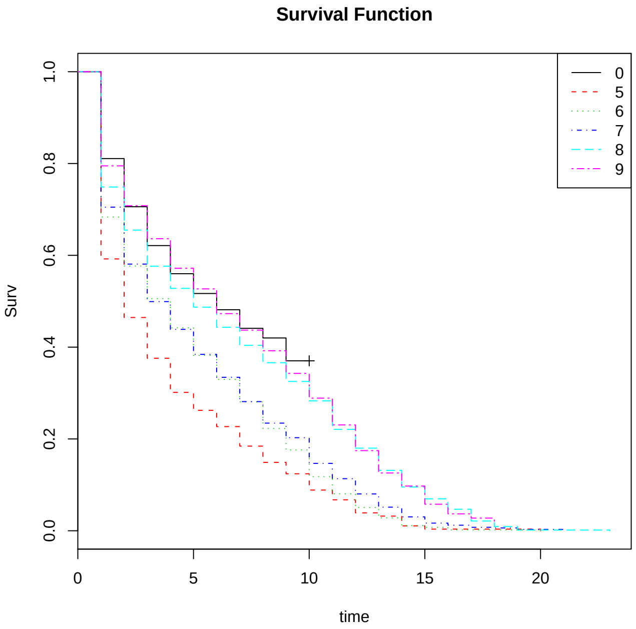

Figure 9: Strata for Decade

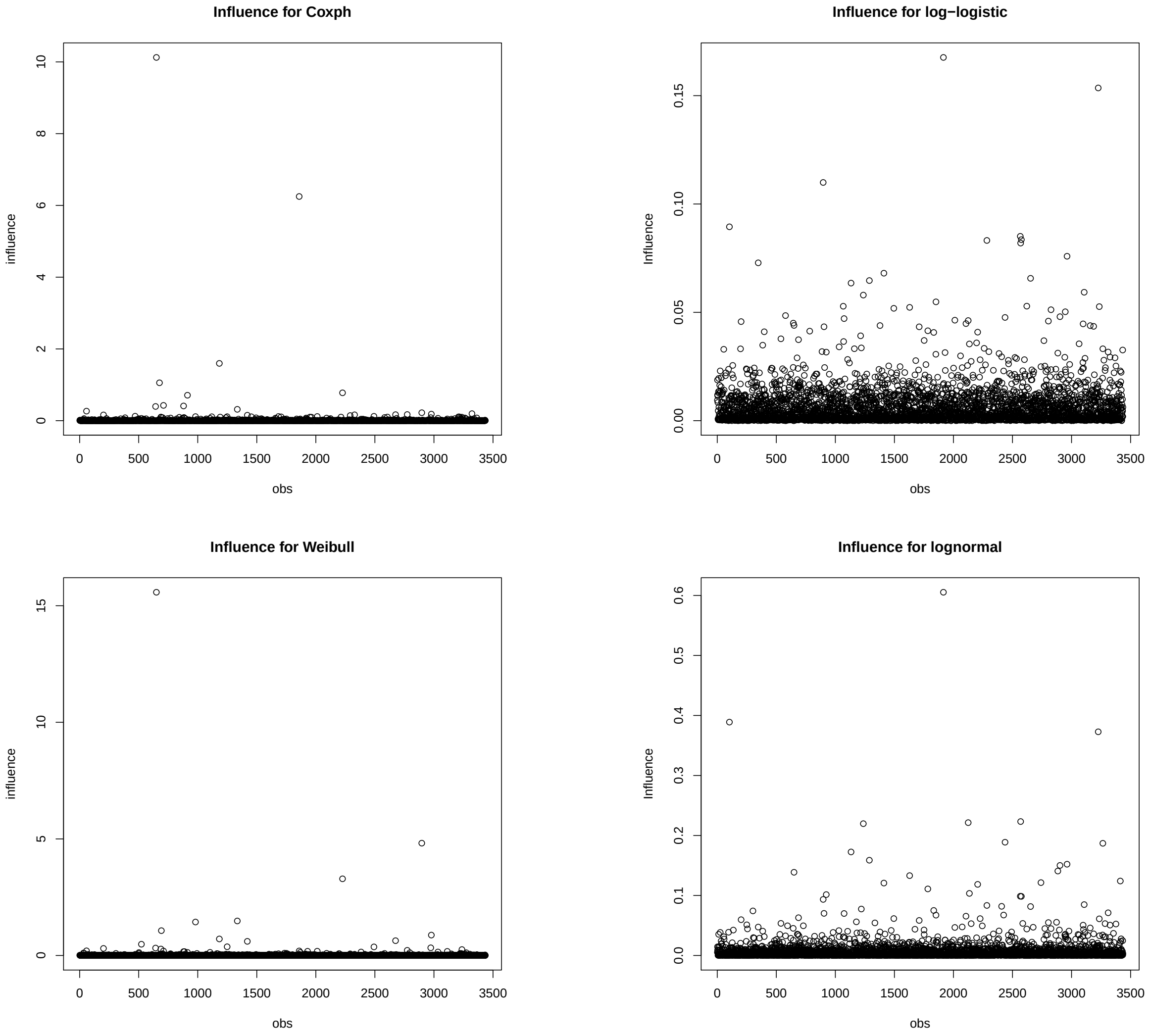

Figure 10: Influence for Four Models