Modelling the survival of elite Australian Rules Footballers

Samuel Glasson and Anthony Bedford

School of Mathematical & Geospatial Sciences, RMIT University, Bundoora, VIC, 3083, AUSTRALIA
anthony.bedford@rmit.edu.au

Abstract

The length of player careers is modelled using Survival Analysis. Survival and hazard functions were constructed for the cohort of players who were first drafted between years 1995 and 2008. A Cox regression was run to determine how career length is affected by draft rank. Consequences for drafting strategy and possibilities for future research are discussed.

Keywords: Survival analysis, Kaplan-Meier estimator, Cox regression, Australian Rules football

1. Introduction

Most Australian Rules football players begin their career at the highest level of the game via the AFL national draft. The draft is an annual event, held since 1986, which was brought in to promote an even competition by rewarding lower performing teams with high selections. Generally, picks are granted in reverse order of the ladder, with the last team receiving the first selection, the second last team receives the second selection through to the premier receiving the sixteenth selection. Several rounds of drafting may occur, with poor teams receiving priority selections. A discussion of the draft, and its flaws, can be found in [1].

There is debate over how to measure drafted player performance over their careers. Ratings systems designed to measure a player's output in each game they play and perhaps sum the results over their career. Output might be measured by the number of disposals or other measures commonly recorded by teams and media. However, assuming that the clubs act rationally in their player management, one good measure of the career of a player is the number of games they play over their entire career. Adopting this simple measure, a 300-game player is deemed to have a much better career than a player who has only played 10 games. Finding a relationship between the number of games played and expected draft position initially may seem an easy one. However, the question arises; how do we handle players who are still playing? In some cases, player careers may span more than a decade, and for the sake of relevancy, it is preferable to use data from the most recent drafts.

Consider a player that has played 100 games to the end of the current season and is required to play on. In any traditional analysis, including this player will bias the analysis, as we know the player will be playing beyond 100 games. Omitting the player from the analysis would also bias the results. In effect, if the length of career is the variable of interest, this observation is said to be "right-censored". If a player has 'retired' then his career is at an end, either through injuries, old age, or no longer being required by his club. For a retired player, we have observed his entire career, therefore the observation of career length is uncensored.

The analysis of censored data lies in the realm of Survival Analysis. This field deals largely with the analysis of time-to-event data, such as the lifetimes of patients in clinical trials (for example a trial to test a new cancer drug). If the event of interest is the time to death, once the end of the study is reached patients who are still alive are censored.

Given the similarities in such data with this problem, we propose to model the career span of players by survival analysis techniques.

2. Data

To test the applicability of survival analysis to the careers of AFL players, data from the national drafts conducted between 1995 and 2008 were collected. Also collected was the number of games played by each

player to the end of the 2009 season. Trades and players who have previously played AFL matches were excluded from analysis. In total, 970 players remained. Of these, 455 players were censored (still playing but not yet retired) at the end of the 2009 season. The remaining 515 players were no longer playing; therefore we have observed their entire careers within our analysis period.

Data on player's careers and the AFL draft is freely available online through a number of sources. Data for this study was obtained through the official AFL website (www.afl.com.au).

Within the cohort of 970 players, 777 have played at least one game while 193 draftees have not played a single game. The greatest number of games played within the sample is 279 by Brent Harvey who was drafted in 1995.

3. Estimates of survival and hazard curves

Defining the survival function, $S(t)$, for a player's career as the probability of a player having a career length T of more than t games we have

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t) \qquad (1)$$

where $F(t)$ is the cumulative density function for that player. The hazard function, $h(t)$, is the instantaneous risk of failure for a player at time t. This gives

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t . S(t)} = \frac{f(t)}{S(t)} \qquad (2)$$

Note that player lifetimes are modelled here as a continuous variable. While the number of games played is an integer - therefore discrete – the range of observations is large enough for the continuous model to be a valid approximation.

The Kaplan-Meier estimate of the survival function for our data is shown in Figure 1. This estimates the survival curve accounting for any censoring in the data. The empirical survival curve in this chart shows the probability that a player will play more than t games based on the collected draft data. Vertical hashes on the curve denote censored observations and 95% confidence limits for the survival function are also shown. The median career length for any drafted player is 70 games. This value can be deduced graphically from the survival curve by projecting a horizontal line from y=0.5 and deriving the x value corresponding to the point of intersection with the curve. As it is imprecise to measure this from the survival curves, Table 1 exhibits the survival values for a selection of games played.

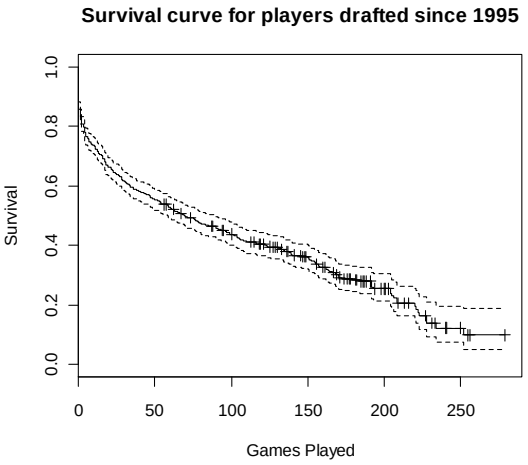

Figure 1. Survival curve for players drafted since 1995.

Games Played, t	Probability of survival, $S(t)$
0	0.86
10	0.73
50	0.55
100	0.44
150	0.36
200	0.24

Table 2. Probability of survival for AFL footballers.

A smoothed estimate of the hazard function for the player data is shown in Figure 2.The smoothing uses the kernel-based methods described in [2]. Recall that the hazard curve tracks the failure rate, where in this case failure denotes the end of a player's career. The hazard curve for AFL players exhibits the classic 'bathtub' shape, with a high failure rate at the start of a player's career, this being most likely due to the difficulty experienced by many new players in breaking into an established side. However, once established, there is a low, constant, hazard rate, which increases from 100 games. As we move along the right tail of the hazard curve, the rate of failure increases further. This is due to the age effect: older players are more likely to retire than their younger colleagues.

But the most striking feature of the estimate survival curve is the difficulty of a drafted player playing his first game. In fact from Table 1, 14% of players will not play a single game. The hazard at this point is at its greatest for any time in a player's career.

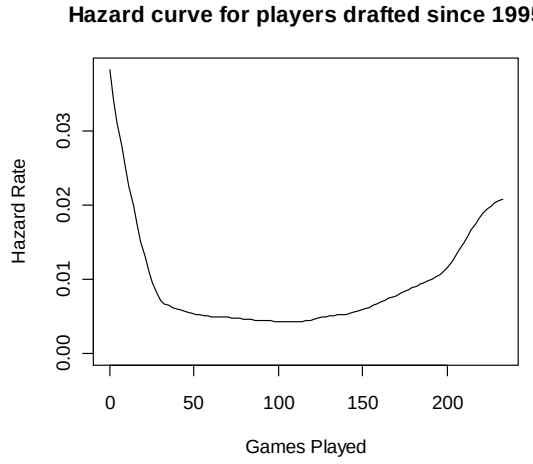

Figure 2. Hazard curve for players drafted since 1995.

4. The relationship between draft selection rank and career length

To examine the relationship between selection number and career length a Cox regression was utilised. Cox regression [3] is classically used to examine relationship of a survival variable to some predictors. A typical model which estimates the hazard of the *i*th observation is given by

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \qquad (3)$$

for *k* predictors. The $h_o(t)$ term is the baseline hazard. It does not need to be specified to fit the parameters, hence the model is semi-parametric. It assumes that the shape of the hazard for each player is the same, but is shifted up or down by the predictor variables. For further discussion of Cox regression and survival analysis see [4] or any text on survival analysis.

In this case there is only one predictor variable: each player's selection rank in the draft. Therefore, the hazard for the *i*th player becomes

$$h_i(t) = h_0(t)\exp(\beta \cdot Selection_i) \qquad (4)$$

where β is the parameter to be estimated. Various transformations of the Selection variable were trialled to improve the fit, but none significantly improved upon (4).

Using the draft data from 1995-2008, and the games played through to the end of 2009, we estimate β = 0.0204 with a *z*-value of 10.7. The *r*-squared for this model is 11.1%. These figures provide evidence of the effect of draft position on the length of a player's career although the low r-squared value does suggest that there may be other factors not captured by the model that influence career length. However, those who are deemed to be better prospects, on average, do have longer careers than those further down the pecking order.

To illustrate this, one can estimate the median number of games in a player's career as a function of the selection rank, such as that given in Table 2.

Selection rank	Estimated Median Career Games
1	204
10	170
20	139
30	98
40	54
50	32

Table 2. The relationship between Selection rank and games played.

The number one pick in a given year is estimated to have a median career length of 204 games. As we move down the ranks, the findings for the players become grimmer. For example, the fiftieth ranked selection has only a 50% chance of playing more than 32 games according to this model. The estimated survival curves for the #1, #10, and #50 ranked selections are shown in Figure 3. While there is not a great difference between the #1 and #10 curves, a player drafted at selection #50 has substantially poorer prospects of a long career.

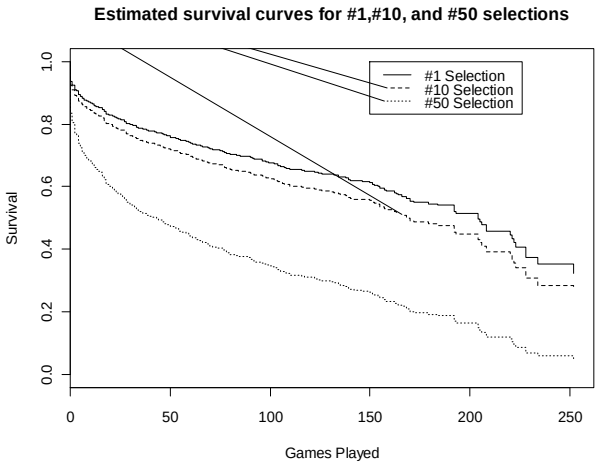

Figure 3. Estimated survival curves for #1, #10, and #50 selections.

5. Discussion

The simple survival analysis model can be used to describe the progress of the players' careers. However, care must be taken in interpreting any such model as the

draft process has been distorted over the years by artificial constructs designed to influence the competition. As new teams have entered the competition they have received extra early draft selections. Furthermore, poorly performing teams receive priority selections and the father-son rule allows some players to be selected out of their natural position.

The findings from this research have a number of implications for the drafting process. Notably, it is a first step in accurately quantifying the value of a draft selection. This is useful for an AFL club wishing to evaluate trades of established players for draft picks. By estimating the number of games that an established player has left in his career, clubs can compare this to the potential (estimated) number of games for any draft pick. Currently there is little analytical use of schemes by the clubs in this regard.

6. Conclusion and further work

In this paper we have estimated the number of games expected by players using simple survival analysis. While results are encouraging, further work can be undertaken to improve the player career modelling.

One improvement will be to add to the number of predictor variables. Each year, prospective players go to "draft camp" where their physical attributes and capabilities are measured. Endurance, speed, power, jumping ability and agility are just some of the variables measured. Adding these into the Cox regression model would allow one to assess which of these tests have a predictive effect on the careers of the players. Another use of this finding would be the analysis of trading picks for players during the allocated trade weeks. This would provide interesting comparison of deals completed and deals proposed.

Further work may be done examining the predictors of a player playing at least one game, as playing the first game appears to be the biggest hurdle a player has in his career. One way of modelling this might be a Tobit model [5], or perhaps by combining two or more models, such as a logistic regression modelling the chance of playing that first match followed by a survival analysis model for the rest of the career.

Quantifying the level of a player's performance during their career may be another way to approach modelling their career. Data are collected by the football clubs and the media on all manner of individual player statistics for every match played. Kicks, marks and handballs are common statistics that are readily available. By finding the performance measures that predict future success, one might be able to make better estimates of the length of a career while the career is still in progress.

References

[1] Bedford, A. Schembri, A (2006). A probability based approach for the allocation of player draft selections in Australian rules football, *Journal of Sports Science and Medicine*, 5, no.4, 509-516.

[2] Mueller, H.G. and Wang J.L. (1994). Hazard Rates Estimation Under Random Censoring with Varying Kernels and Bandwidths, *Biometrics* 50, 61-76,

[3] Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society. Series B (Methodological) , 34, No. 2*, 187-220.

[4] Collett, D. (1993). *Modelling Survival Data in Medical Research.* Boca Raton: Chapman & Hall/CRC.

[5] Tobin, James (1958), Estimation for relationships with limited dependent variables, *Econometrica* 26 (1): 24–36