CEP 13-01

# Youth Training Programs and their Impact on Career and Spell Duration of Professional Soccer Players

Mihailo Radoman
Carleton University

Marcel-Cristian Voia
Carleton University

January 8, 2013

# CARLETON ECONOMIC PAPERS

Carleton UNIVERSITY
**Department of Economics**
1125 Colonel By Drive
Ottawa, Ontario, Canada
K1S 5B6

# Youth Training Programs and their Impact on Career and Spell Duration of Professional Soccer Players

Mihailo Radoman
Department of Economics
Carleton University
Ottawa, Ontario
CANADA K1S 5B6
mradoman@connect.carleton.ca

Marcel C. Voia
Department of Economics
Carleton University
Ottawa, Ontario
CANADA K1S 5B6
mvoia@connect.carleton.ca

January 8, 2013

**Abstract**

A unique data set of post-war English trained soccer players that signed professionally with their parent club when they turned 18 is used to study the impact of their stay with the home team and their total career duration. The home team (first) spell and career durations of these soccer players in a top European leagues is modeled using robust hazard models. The results of the analysis show that players that start their professional careers after acquiring training in competitive youth academy/programs have different outcomes on their career and first spell duration depending on the clubs they start their training. The first spell duration analysis is performed to estimate the bond or loyalty factor established by clubs with their youth trainees. The spell analysis outlines the nature of the competitive environment in which smaller clubs have a chance to keep up with the larger ones in terms of producing and holding on to home-grown talent. This would be a necessary condition for them to remain competitive in light of their lagging financial resources that limit their activity and ability to attract top talent in the soccer transfer market. The analysis of career duration in the top European leagues will show the success of a specific academy's training programs in producing players competitive in top soccer leagues. Finally, the results of both analyses were tested for endogeneity bias using a split sample test.

**Key Words**: career duration of soccer players,youth training programs, duration models, model evaluation;

**JEL Classification**: C14, C41, C52, J24, J44.

# 1 Introduction

Linking education or specialized training with career success has been the subject of many economic studies in the past. Most articles focused on conventional types of education and schooling in assessing its return on investment. The sports industry has been growing rapidly worldwide, but the impact of education or training on professional success of players has been largely unexplored in economic literature. Youth System is a sports term that refers to a youth investment program within a particular team or league, that develops and nurtures young talent with the vision of exploiting the ones that show promise in the first team at a latter stage. Most sports in North America rely on the high school and collegiate system to develop young athletes, with the exception of hockey in Canada and baseball in USA that offer many junior and minor leagues for youth development throughout the respective countries. [1] On the contrary, most team sports in Europe concentrate and invest in youth systems and almost every country has youth level leagues in multiple disciplines for young athletes to compete and further develop. Soccer is the world's most popular sport, and the revenues earned by professional clubs and individual players are significant enough to warrant an independent study assessing the impact of education/training on career duration of individual players. Many European soccer clubs have youth programs or academies for training talented players under the age of 18, and this type of organizational structure paves the way for the analysis undertaken in this paper.

Youth soccer is an essential component in the development cycle of any aspiring young player. Those individuals selected at youth level are considered to have the necessary potential to become a professional soccer player. Most of these youth players are trained and nurtured at academies of licensed soccer clubs, with the hope of signing a professional contract with the club when they turn 18. This is the "grassroots" level of soccer that teaches young players all of the basics of becoming a soccer player. Soccer, much like any other team sport, is continuously changing, requiring grueling physical and mental effort by players. The basics provided at youth level are intended to equip the players with the necessary tools that help them establish foundations

[1] Considering the focus on youth development, it is not surprising that most of the talented players in the NHL are Canadian trained and Canada is a hockey powerhouse in world competitions

upon which to build on for the rest of their careers.

Clubs throughout Europe have invested a lot of time and money[2] in their youth academies, especially in recent decades, as the benefits of successful youth players have been realized. Clubs have numerous scouts and youth coaches that assess potential players at each step of their development process. The large investments made by clubs in youth academies are not surprising considering the size of the transfer fees[3] paid for soccer players throughout the world, and, particularly, in Europe. Developing home-grown talent is a viable long-term strategy for any club that wants to be successful, financially sound, and competitive in the long-run. Each soccer club faces an economic dilemma whether to invest in their youth academy or in the transfer market in search of new talented players. One can say that clubs investing more heavily in their youth academies have a long-run perspective of the future business conducted by the club. Producing high-performing players through the youth system can also ease the financial strains of a soccer club by avoiding large spending in the transfer market in order to stay competitive. Barcelona Football Club (Spain) is a perfect example of success brought on via the success of their youth academy (La Masia). They have won all major honors in domestic and international soccer competitions in recent years, and the majority of their first team are products of their youth academy. The success of their youth program has put less strain on their transfer market expenditures and allowed for more flexibility in their financial structure (they were one of the only teams in the world without an official sponsor on their shirts until 2010, showing the financial strength of the club). Not only does a successful youth academy ensure a continuous flow of home grown players into the first team, it can also serve as a profit driver for any soccer club. Among other channels, clubs generate revenue by selling players. The profit realized in each sale equals the club's revenue of selling a player minus the cost of acquiring that player's rights. Since the cost of acquiring a home-grown players are negligent or zero, the profit realized on the sale of these players is 100 percent. In addition to the financial benefits of running a successful youth program, there are intangible benefits in terms of bonds and

[2]One recent example is the Chelsea Football Club (England) who has invested millions of pounds into a new academy center.

[3]Transfer fees are paid in soccer when one club purchases a player under contract from another club. Generally, the transfer fees decline with the number of years a player has remaining on his contract with the current club. Transfer fees for top players are now approaching 100 Million euros.

loyalty created between the youth player and his home club that could last for the player's entire career. This is extremely beneficial to all soccer clubs, in particular to smaller clubs because it allows them to stay competitive with larger and financially superior clubs.

Specialized academies recruit youth players anywhere from 8 to 18 years of age. Generally, players do not engage in full pitch 11 on 11 games until the age of 12, and the most gifted ones are offered youth scholarships at the age of 16. Not very many of the remaining scholars are offered professional contracts at the age of 18, and very few of the players signed professionally at 18 are still playing professionally at the age of 21. The data set in this study consists of mostly those players that were trained at youth level and signed professionally by their parent club when they turned 18.

The empirical analysis in this paper relies on a uniquely collected data set of post-war (WWII) English soccer players trained at youth level by a certain club, having made at least one first team appearance for that club. These would be the players that passed the screening process at youth level and were signed under a professional contract upon graduation from the youth program or academy. In most cases, they were signed by the parent club that provided the youth training for these players in the first place. Surviving as a professional soccer player in the top European leagues is a daunting and challenging task. In addition to the competitive threat posed by upcoming youngsters and international players, soccer players in England (and other European countries as well) face the yearly threat imposed by the relegation system, whereby three (in the past it was two) bottom positioned clubs are demoted to a lower league. The aim of this research study is to identify the relative importance of different education/training programs, among other observed and unobserved factors, in explaining the labour market differences of individual soccer players measured by the length of their professional career and home-team spell. Given the scope and time covered by the data, and the lack of data on salaries and transfers for the players covered in this study, the economic returns to both clubs and players are estimated using players career and spell durations. Assuming that players earn income in each year of their professional career, their economic return will be greater the longer their career duration is. Clubs that hold on to their home-grown talent longer will tend to rely less on the transfer market for bringing in new players,

and will reduce the risk of new and unknown players not fitting into into their team system. From an economic perspective, this can represent a large cost saving and larger economic profit for any club.

The specific data on players spell duration allows for assessing the differences in career duration versus duration with the parent club that trained the player at youth level. Generating quality players through youth training is important, but it is just as important to establish which clubs have the ability and willingness to hold on to their youth products the longest. This analysis is testing whether certain clubs are more successful than others in creating loyalty or bonds with the players they brought up through their youth system. In other words, the spell duration analysis can shed some light on the return of investment achieved by the clubs through their youth academies. This analysis isolates the home-grown players that a club perceives as future first-team players, rather than profit generators in the short-term. Hence, this analysis evaluates differences between clubs based on the ability to assess their youth product in terms of incorporating them into their first team in future years. The implications of this analysis might be different for larger versus smaller clubs. The competition for first-team spots is more intense at larger clubs, and some home-grown players are forced to move to smaller teams in search of regular playing time and a longer career. The competitive first-team environment at larger clubs makes it much more difficult for youth players to establish themselves and have long spells with their parent club. Therefore, the results for larger clubs cannot be interpreted in the same fashion as the results for smaller clubs in the study.

For the career duration we are evaluating the differences of youth programs on a player's career duration, conditional on the player signing his first professional contract with his parent club. Therefore, we are indirectly also measuring the ability of a team [4] to evaluate their own youth products (by choosing who they offer a professional contract) in terms of a player's ability to compete at the top level of English football. Essentially, we are measuring the differences in the internal mechanism of the 16 clubs in assessing their home-grown talent at the final stage of youth development (when a player turns 18 and a decision has to be made whether to sign or release

[4]The teams selected for this study are the 16th best teams that survived in England's top soccer league for an extended period of time

him). Naturally, it is harder to make it professionally at larger clubs so players that get released by top clubs might make it at lower level clubs (this is the sample that we don't have). At the same time, larger clubs will sign players that they might profit from by selling them to other clubs, even if they don't envision them playing for their first team. Therefore, this study evaluates in addition to differences in the actual youth programs, the differences in the clubs internal mechanisms for home-grown player selection. From an economic perspective, clubs will sign home-grown players that they envision playing for their first-team or the ones they believe are talented enough to play in other clubs and leagues, thus profiting from their sale in the transfer market. Our analysis sheds some light on the differences in the business/economic decisions made by these 16 clubs in terms of the length of careers for their youth players that they decided to sign at the age of 18.

The unique structure of the data also permits multiple other questions to be broadly examined, including:

- The impact of a particular choice of position on the duration of a player's career. Players and coaches make choices early on in a player's career on the position to be assumed by that player, and that choice might impact the player's career duration because each position is demanding in its own way. More talented players are typically assigned midfield and forward roles on any given team. Certain positions, like full-backs and wingers are reliant on speed and stamina, so one would not expect players at these positions to play past a certain age when those required attributes start declining. In addition, certain positions in soccer are subject to more scrutiny than others because their performance is tied to easily identifiable statistics, like the number of appearances, goals scored, etc. For example, the productivity of forwards is generally measured by the number of goals they score. The same measure of productivity cannot be applied to the other positions in soccer. In addition, a player's adaptability to other positions is certainly valued by clubs as they can assume some less physically demanding positions as they age but still use their experience to help the team. [5] Therefore, it is important to determine whether the position of a particular player has any impact on the longevity of his professional career.

- Potential discrimination measured by the career duration in the top league of home grwon

[5] This study considers what position a player assumed for most of his career, which overlooks the adaptability factor for relevant players.

players against foreigners can also be tested in this study, however this test may not have sufficient power given that most players in the data set are from the United Kingdom.

- The competitive structure of English soccer has changed over the period of study. For example: (i) In 1973, the Football league announced that three teams (instead of two) would be relegated from the top two divisions; (ii) In 1981, a change was initiated so that three points were awarded for a win instead of two; (iii) In 1992, the Football Association created the FA Premier League replacing the Football League First Division allowing for 22 clubs to compete instead of 20[6]; (iv) In 1996, the Bosman ruling impacted all European competitions by increasing the number of foreign players allowed per team and allowing for free movement without any restrictions on clubs for players with passports from Union of European Football Associations (UEFA) countries. These institutional changes could have opposing effects on a player's survival in the top league; one can assume that increasing the number of teams in the top division will provide a player with a higher chance of retaining a spot in the top league, while increasing the number of relegated teams can have a negative effect on a player's chances of staying in the top competition. We hope to extrapolate the effect of some of these changes in the competitive structure of English soccer by testing the effect of the year of entry on a player's career duration.

Finally, using individual specific characteristics and productivity measure variables allows for the identification of conditional duration patterns for an individual player. Consequently, reduced form duration models are estimated to model a player's career duration, as well as his spell duration with the club who trained him at youth level. A benchmark model, that includes individual specific characteristics, is estimated first. Subsequently, productivity variables are included in the model, while the complete model will account for nonlinear functional form of relevant continuous covariates and unobserved heterogeneity, if applicable.

To the best of our knowledge, this paper is the first of its kind in economic literature, directly addressing the impact of different youth training (education) programs on the career success of professional players in any competitive sport. Frick, Pietzner and Prinz (2007) attempt to answer several questions simultaneously, mainly focusing on the effect of changes in the German soccer

[6]Later reduced back down to 20 teams in the Premier League.

institutional environment on individual player career duration. They estimate a Cox Proportional Hazard model to address their questions, based on data collected for all players appearing in the German First Division (Bundesliga) during a 40 year period. Their findings indicate that institutional changes have a strong effect on individual players' careers, career duration of goalkeepers is longer than for the other positions considered (goalkeepers, defenders, midfielders, and forwards), and that players from certain regions (i.e., Eastern Europe) might suffer from discrimination. Ohkusa (2001) examined the quit behaviour of Japanese baseball players, while Atkinson and Tschirhart (1986) studied the determinants of career length for NFL players. Others, like Groothuis and Hill (2004) have analyzed the exit discrimination of black players in the NBA.

The rest of the paper is organized as follows: Section 2 discusses the data set and provides some summary statistics; Section 3 discusses the methodology employed; Sections 4 and 5 discusses the results of the empirical analysis for career and home team spell duration; and Section 6 concludes the paper.

## 2 Data Set and Summary Statistics

The overreaching goal of this labour market study is to assess the role of training/education, among other factors, in the career duration of professional soccer players participating in the top Western European leagues.[7] Most European soccer clubs have youth programs or academies for children under the age of 18 that are responsible for training and developing young players, and equipping them with the right tools (beside natural talent and ability) to become professional players and enter the soccer labour market. Despite the abundance of soccer data available today, information pertaining to youth programs/schools of soccer clubs is not readily available, at least not publicly. England is home to the oldest soccer clubs in the world (dating from at least 1857), the world's oldest competition (the FA cup founded in 1871), and the first ever soccer league (1888). For these reasons England is considered the home of the modern game of soccer, and it is not surprising that some of the highest quality soccer statistics come from England.

[7]The top leagues considered for this study were from the following countries: England, Spain, Italy, Germany, France, and Portugal.

The empirical analysis of this paper is based on a data set collected from 16 English clubs[8] in the post-War period commencing in 1946 and ending with the 2010/2011 season. These clubs were selected based on data availability and their extended presence in the top tier of English soccer over the period of study. The data, as presented, will be less likely to induce selection bias in the results, as the selected clubs are among the most frequent participants in the top tier of English soccer.[9] In addition, a player's relative abundance of human capital[10] and natural talent should assure a presence in the top leagues, regardless of whether his home team, or any other team that he plays for, gets relegated from the top league. The data was gathered though a variety of Internet sources, the main source being http://www.neilbrown.newcastlefans.com/, a site providing A-Z player data for post-war English and Scottish Football League participating clubs. Every player that made at least one appearance for an English or Scottish club during this period is accounted for, some with more detailed information than others.[11] The data from this site contains statistics on the full universe of players, including international players, that appeared for the first team of any of these clubs. The English League player's statistics are drawn from Berry Hugman's series of books (1984-2005), "The Premier and Football League Player's Records". This site provides player information that is useful to the study undertaken in this paper, including: name, birth date, position, seasons played at British clubs in all leagues, source team and team the player next transferred to, league appearances and goals for each spell, international appearances (caps) and goals. The source data were of particular importance because they provide information about players appearing in the first team of a certain club and if the player was trained at youth level[12] by that particular club. This information can be used as a proxy for receiving some education/training prior to starting a professional career and define the initial conditions of the player-level data. Consequently, this proxy served as the backbone for the creation of the data set, and allowed for accurate tracking of

[8]The 16 clubs included in the study are: Arsenal, Aston Villa, Chelsea, Coventry City, Everton, Leeds United, Liverpool, Manchester United, Manchester City, Newcastle, Nottingham Forrest, Sheffield Wednesday, Southampton, Tottenham, West Bromwich, West Ham.

[9]The top tier of English soccer was called Division 1 until the Premier League was established in the 1992/1993 season.

[10]In soccer, human capital can be ascertained by a level of tactical knowledge and understanding that allows players to embrace tactical changes imposed by new managers or new systems of play.

[11]In fact, more detailed data is available for all players that have exited the market during the study period.

[12]Youth trainees, junior players and apprentices are all treated as players that received training by their parent clubs prior to signing a professional contract.

the careers of 1121 professional players trained by, and first appearing for one of the 16 clubs in this study during the post-war period.

Soccer data examined prior to the 90s contain information on spell level duration irrespective of the tier in which the club participated during the player's spell. Most of the clubs analyzed did not have a constant presence in the top English league and experienced some movement within the English League system through relegation and promotion.[13] Numerous other Internet sources, outlined in Appendix A, were engaged to gather the relevant statistics from the top European leagues for each player in the data set. Most players dropped from the risk pool were due to the inability of segmenting their club-specific spell data in the event the spell carried over to lower level leagues in any particular country in which the player appeared. The data track a player's spell duration for each club that he played for in the top leagues during his entire career, regardless of exit and re-entry as only the relevant seasons and statistics in the top tier are accounted for. For example, if a player had two distinct spells at a particular club included in this study, the data combines the two spells as one measure of duration at that club. Furthermore, if a club was relegated and promoted again, the data tracks a player only during his time in the top tier treating the lower league period as a temporary exit, unless a player never reappears in the top tier. Any final departure[14] from a top league of a Western European country, whatever the reason might be, was considered an exit from the top tier competitive soccer market. The players contained in the data set all enter and exit the top-level professional soccer market during the period of this study [15]. The unique feature of this data is that all of the observed players were trained at youth (unprofessional) levels by one of the 16 English clubs analyzed in this paper. The rules surrounding the recruitment of youth players in England have changed over the study period. Prior to youth academies and youth leagues, players belonged to junior squads of their parent clubs or were signed as an apprentice before they were offered professional contracts. Teams were not allowed to recruit players that

[13]The bottom two or three teams of every league, except the bottom league, got relegated to the lower league and the top 2 or 3 teams of every league, except the top league, got promoted to the higher league on a yearly basis. A similar system is in place for other Western European countries.

[14]A player may exit the top league for one of the four following reasons: movement to a lower level league, retirement, injury, or death.

[15]The aim of this study is to assess the differences of youth programs attended on a player's career and spell duration at the top professional level. This is why the data doesnt track players if they drop out of the top league, even though we are aware that they are still playing professionally in the lower leagues and earning income

were located more than a 90 minute drive away from the training ground. International/non-EU players under the age of 16 were/are not allowed to be signed. (Footnote: Teams, particularly the larger ones, got around these regulations by moving a player's family closer to their training location and by sending young international players to their affiliated feeder clubs in countries with lax regulation, like Belgium, so these players can gain EU citizenship) These rules have historically restricted smaller clubs from searching for talent outside of their area, but have made it more difficult (at least theoretically) for the larger clubs to take away their local talent. The clubs employed in this paper are large and historically consistent top league participants, and they were all facing the same rules and regulations and similar youth selection criteria. Furthermore, there is no evidence that the talent pools were significantly different in any one of the areas belonging to any clubs that are part of the study. Generally speaking, the athletic attributes of the preselected youths were not that different throughout the country and international (particularly non-EU) players do not represent a significant portion of the sample to distort the results. This assumption is more grounded in this study since the players included are youth trainees that sign professional contracts at the age of 18 and represent the cream of the crop from each academy (footnote: Less than 1

The data collected include detailed player level information and statistics that can be structured into two groups:

1. Variables with initial conditions and personal characteristics: player's name, age at first appearance in a top league, year in which the player first appeared in a top league, nationality of the player, youth/training program attended, and position played for the majority of his career.[16]

2. productivity measures: number of top league seasons played at each club in his career, total number of seasons played in top leagues, average (per season) first team top league appearances for each club in a player's career, average (per season) total career appearances in the top league, average goals scored per season, international caps, and international goals scored.

Basic summary statistics for the continuous variables are presented in Table 1. HT refers to

[16]Six positions were considered for the purposes of this paper: Forwards, Midfielders, Wingers, Central Defenders, Full Backs, and Goalies.

home team that trained a particular player at youth level. There are 13 players who were trained by a club but did not make a first-team appearance for that club in the top leagues.[17] The mean HT spell duration is not much lower than mean career duration, which suggests that a special bond exists between players and the clubs that provide youth training for them. Maximum career duration is 21 and maximum HT spell duration is 19. The youngest player to make a first-team appearance for a club in this study was 15, while the oldest was 29. The last observed entry was in 2003. Table 2 outlines the detailed information for all player exits at each observed time period for career duration and home team spell duration. Exits are a little more frequent in the early years for the spell duration versus career duration. Graphs of the empirical hazard and survivor functions for career and HT spell duration are presented in Figures 1-4. The empirical hazard graphs show a strong positive duration dependence for the players career and HT spell duration, indicated by the upward sloping curves. This is not surprising considering the age and physical limitations of individual players, as well as intense competition in the marketplace. The estimated survivor function for HT spell duration has a steeper slope than the same curve for career duration, particularly in the early years of duration time analyzed. This coincides with slightly higher hazard rates for those same duration times. The main observable difference between the career and HT spell duration empirical hazards is in the hazard levels of the two curves; The HT spell duration hazards are higher than career duration hazards in general. In addition, the career duration empirical hazard displays some variability in the upward sloping latter stages of the analysis (between duration times 17 and 20), whereas the empirical hazard for HT spell duration slopes upward sharply without notable variability after 15 years.

Visual and formal Wilcoxon and Log-rank tests[18] are performed to see if the observed subgroup differences in the survivor functions are significant. Test results indicate that subgroup differences are present for youth team attended and position played. Figures 5 and 6 illustrate sub-group differences at youth team level for career and HT spell duration, respectively.[19] Players arising

[17]This could be because the club was in the lower league at the time a player was there, or that the player was signed by another club professionally straight out of his youth program.

[18]These tests are non-parametric statistical hypothesis tests used when comparing two or more related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ.

[19]The figure displays 5 out of the 16 clubs that are part of this study. The graph with all 16 clubs is to condensed for interpretation.

from the Sheffield Wednesday youth program have the highest and most stable estimated hazard for career and HT spell duration analysis. Figure 7 illustrates the subgroup differences at the position level for career duration. This figure suggests that the estimated hazard rate for goal tenders is lower than for the other positions for the majority of the duration time analyzed. Figure 8 provides the same estimates for HT Spell Duration. The graph suggests that forwards have the highest hazard rate among the positions considered. Similar type of analysis resulted in the selection of Sheffield Wednesday as the appropriate youth team benchmark in career and HT spell duration analysis, since its estimated hazard rates appeared to be on the bottom end of all of the clubs considered. These results indicate that the different youth teams and positions each have a unique effect on the hazard rate, and that the objective of this paper can be explained by the data.

# 3 Methodology

Survival analysis is based on duration models that are used to estimate the hazard rate. The hazard rate, or the instantaneous probability of exit, is the probability of experiencing an event at time $t_i$ conditional of having survived to time $t_i$.

$$h(t) = \lim_{\triangle t \to 0} \frac{\Pr(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \tag{1}$$

The reduced form hazard models in this paper are conditioned on individual specific covariates, including dummies for the player's youth team, position played, and nationality. Nonparametric regression techniques for testing nonlinear functional forms of continuous covariates are applied to improve model specification. The effect of unobservables is reduced by accounting for individual-specific unobserved heterogeneity, where applicable. For robustness checks, competing parametric and semiparametric models are presented.

## 3.1 Parametric Models

One of the limitations of parametric models is that one has to make an assumption about the shape of the baseline hazard rate, i.e. the shape of the underlying duration dependence. Hence, parametric models can suffer from misspecification if the baseline hazard places overly strong restrictions (parametric assumptions) on the direction and shape of the duration dependence. On the other hand, if the characterization of the underlying duration dependence is accurate, the parametric models will provide more reliable and accurate estimates than semiparametric or nonparametric models. A large number of parametric models are considered, but results are reported only for the Gompertz model for career duration, whereas the Generalized Gamma model is considered for the spell duration analysis without reported results for reasons explained subsequently in the results section.

The Gompertz hazard is defined as:

$$h_i(t_i|x_i, v_i) = v_i \phi(x_i) \exp(\gamma t_i), \tag{2}$$

where $\phi(x_i) = \exp(x_i\beta)$ is a function of the observable time-invariant covariates, and $\exp(\gamma t_i)$ is the baseline hazard for individual player $i$.

The Generalized Gamma model nests several other parametric models and is useful in testing the appropriateness of these nested models:

$$h(t_i|x_i) = \frac{\lambda p(\lambda t_i)^{pk-1} \exp[-(\lambda t_i)^p]}{\Gamma(k)} \tag{3}$$

where

$$\lambda_i = \exp(-x_i\beta)$$

and $p$ and $k$ are two shape parameters, and $\Gamma(\cdot)$ is the gamma function.[20] The shape parameters allow for a flexible hazard rate, and by taking on certain values they can reproduce other parametric models in the following way:

[20] $\Gamma(k) = \int_0^\infty e^{-t} t^{k-1} dt$ which reduces to $\Gamma(k) = (k-1)!$ if $k$ is a positive integer.

- If $k = 1$, then the Weibull hazard is implied.
- If $k = p = 1$, then the Exponential hazard is implied.
- If $k = 0$, the Log-normal is implied.
- If $p = 1$, the gamma distribution is implied.

For non-nested parametric models, Akaike's Information Criterion (AIC) is used to distinguish between them. In fact, the AIC test is conducted on all parametric models, both nested and non-nested, to avoid over-reliance on the results from the Generalized Gamma model.

## 3.2 Semiparametric Models

Semiparametric models, like the Cox Proportional Hazards (PH) model, provide alternatives to parametric models. These models are more general and allow the estimation of slope parameters for the covariates irrespective of what the baseline hazard looks like. As such, the Cox PH model makes no assumptions about the distribution of survival times and is robust to misspecification of the baseline hazard. A typical semiparametric model is of the form:

$$h_i(t_i|x_i, v_i) = \phi(x_i)\lambda(t_i),$$

where $\phi(x_i) = \exp(x_i\beta)$ is a function of the observable time-invariant covariates, and $\lambda(t_i)$ is the nonparametric baseline hazard for individual player $i$.

The Cox model assumes that the covariates will have a proportional and constant effect that is invariant to time. According to Box-Steffensmeier and Jones (2004), testing the proportionality of hazards assumption is arguably the primary concern when fitting a Cox model. Residual-based tests using Schoenfeld residuals[21] are particularly useful when testing the proportional hazards assumption. The Therneau-Grambsch test (1994) is based on scaled Schoenfeld residuals and is

[21] If the residual exhibits a unsystematic pattern at each failure time, then this suggests that the covariate effect is not changing with time so the PH assumption holds. In this case, the plot of Schoenfeld residuals against time should display a zero slope.

applicable both globally and at specific covariate level.[22] One has to be careful when interpreting the results because the proportionality of hazards test can yield a false result of nonproportionality if the model is misspecified (Therneau and Grambsch 1994, 2000). Therefore, attempting to specify the model correctly should precede any testing of the PH assumption. If the PH assumption fails[23] the most common remedy is to interact the covariate that failed the PH test with the natural log of time and include both (the covariate and the time interaction term) in a new regression.

### 3.3 Functional Form of Covariates

Standard tests for nonproportionality of hazards are quite powerful, but they can also detect a variety of other specification errors, including the misspecification of the functional form for the covariates. This is particularly important for the Cox semiparametric model, but it shouldn't be overlooked for the parametric models as well. Testing for nonlinear functional forms of continuous covariates is a possible solution to improve model specification. This test is usually done by including fractional polynomials of the covariate, or by nonparametric methods such as splines. Keele (2010) notes that very few articles in the literature test for nonlinear functional form of continuous variables (linear form is taken as given), with no results of nonlinearity tests using the nonparametric spline technique. Keele[24] notes that the close relationship between functional form and the PH test lends extra urgency to modeling of any nonlinearity.

Productivity measures might have different implications in the career duration model and in the HT spell duration model. While we still expect a positive correlation between productivity and duration, HT spell duration can have opposite correlation with productivity depending on the team a player originates from. If a player originates from a smaller and historically less successful team

[22]The test statistic for the nonproportionality test is:

$$T_k = \frac{\{\sum(g_k - \bar{g})s_k^*\}^2}{dI_k \sum(g_k - \bar{g})^2}$$

where $I_k$ is the information matrix where elements for covariate $k$ and $d$ are the event times.

[23]The PH assumption needs to pass the test globally and at the individual covariate level for the researcher to be able to rely on the model's assumption.

[24]Keele's paper uses simulations and well-known empirical examples from the literature to demonstrate the importance of testing for nonlinear functional forms. He demonstrates that, if this issue is ignored, the analyst might might be mislead by incorrectly relying on results from the nonproportionality test.

in our sample, increased productivity might result in shorter spell duration (negative correlation) with his home team as he will attract the attention of larger and more successful clubs that try to engage his rights in the transfer market. For the larger and more successful clubs in our sample, we would expect a strong positive correlation between productivity and spell duration. This reasoning suggests that there might be a critical point (inflection) for productivity of players in small teams, where the correlation with spell duration changes from positive to negative as the player outgrows his current team. On the other hand, productivity can also be a factor differentiating career duration of players originating from larger and smaller clubs. A certain level of productivity in a smaller club might spell an end of a player's duration in the top league and make him more suitable for lower league competitions, while the same level of productivity at a larger club might result in a player's movement to smaller clubs in the top league that extends a player's career duration. Considering the above, it is not surprising that the most suitable career and HT spell duration models incorporate nonlinear effects for productivity covariates (appearances per season).

Allowing for nonlinearity in continuous covariates, particularly in productivity measure variables, will reduce the effect of any endogeneity on the model results. This paper will test the continuous covariates for nonlinearity using restricted cubic spline functions (Royston and Sauerbrei 2007) defined as follows:

$$\begin{aligned} s(x) &= \beta_{00} + \beta_{10}x + \sum_{j=1}^{m} \beta_j \{(x-k_j)_+^3 - \lambda_j (x-k_{\min})_+^3 - (1-\lambda_j)(x-k_{\max})_+^3\} \\ &= \phi_0 + \phi_1 x + \phi_2 v_1(x) + ... + \phi_{m+1} v_m(x) \end{aligned} \tag{4}$$

where $k_{\min}$ and $k_{\max}$ represent two boundary knots,[25] $\phi_0 = \beta_{00}$, $\phi_1 = \beta_{10}$ and $\phi_{j+1} = \beta_j, v_j(x) = (x-k_j)_+^3 - \lambda_j(x-k_{\min})_+^3 - (1-\lambda_j)(x-k_{\max})_+^3$ for $j = 1, ..., m$.

In particular for each continuous covariate, the most complex permitted regression spline model is chosen (determined by the degrees of freedom assigned to the function) that has m+1 degrees of freedom, where m is the maximum number of knots to be considered and m=0 represents a linear function. The most complex model, Mm, is compared to the null model (omitting x) using a chi

[25]The cubic regression spline is restricted to be linear beyond these boundary knots.

squared test with m+1 df. If the test is not significant at ($\alpha$) level, the procedure stops and the regressor x is eliminated. Otherwise, the fit of Mm is compared to M0, and if the difference is not significant at the ($\alpha$) level, M0 is chosen by default and the algorithm stops. The procedure continuous in this fashion until either a test is not significant or all the tests are significant, in which case Mm is the final choice. The Multivariate Restricted Spline algorithm considers each predictor in turn (in decreasing order of statistical significance in a full linear model), with the functions of all other covariates temporarily fixed. The algorithm cycles over each predictor repeatedly in the same order, changing the model according to the results of the tests of individual variables, and the process stops when there is no further change in the variables included in the model and in the knots chosen for each continuous variable.

## 3.4 State Dependence versus Unobserved Heterogeneity

There are two possible sources for observing duration dependence: state dependence or true state dependence (TSD) and unobserved heterogeneity or spurious state dependence (SSD). The notion of SSD amounts to observations being conditionally different (heterogenous) in terms of their hazard because individual specific unobserved characteristics. Models that don't incorporate unobserved heterogeneity assume that all observations with the same values for all covariates are identical. The amount of time these individuals spend in a certain state does not affect future probabilities of exiting or remaining in the state. On the other hand, under TSD, prior experience plays a role in determining the probability of remaining in a given state. Positive TSD implies that the longer an individual is in a given state, the more likely he is to stay there, which is equivalent to negative duration dependence as the hazard rate falls or remains flat over time. Negative TSD implies that the probability of exiting a given state increases with time spent in that state, which is equivalent to positive duration dependence. Applying these notions to career length of individual soccer players would suggest that individuals with longer survival times increase their exit probability due to physical limitations imposed by age (negative TSD), but professional experience gained at a young age can result in positive TSD until a player reaches a certain age. The empirical hazard for career duration is relatively flat until season 11 of a player's career, after which it starts sloping upwards

sharply. This suggests that prior experience is important (positive TSD) until year 11, and that the age factor dominates (negative TSD) after that time.

The presence of SSD in models that don't account for it will result in misspecification for the following results:

- The model without SSD will over-estimate (under-estimate) the degree of negative (positive) duration dependence in the baseline hazard.
- The proportionate effect of a regressor on the hazard rate in a continuous PH model would no longer be constant and independent of survival time.
- Estimates of positive (negative) regression coefficients will underestimate (overestimate) the "true" estimate.

Jenkins (Essex summer school course on survival analysis) notes that if a fully flexible specification for the baseline hazard is used (i.e. Cox model), then the magnitude of the biases in the non-frailty models are diminished. This implies that parametric models are more sensitive to model misspecification and, in particular, not accounting for unobserved heterogeneity when this is present will increase the bias in these models. According to Jenkins, conclusions about the empirical relevance of unobserved heterogeneity is likely to differ from application to application. It is important to note that, if the frailty effect is real, the PH model loses its normal proportional hazards property because the hazard ratios are now conditional on the unobserved frailty.[26] This implies a link between the nonproportionality of hazards test and unobserved heterogeneity tests.

Unobserved heterogeneity can take two forms: at the individual observation level and shared by a group of observations with similar characteristics. This paper introduces individual-level unobserved heterogeneity multiplicatively (in parametric models) relying on random effects models.[27] Conditional on choosing the proper covariates and SSD distribution, the random effects models (frailty models) generate estimates positively correlated to the empirical hazard and correctly predicting its slope. The preferred model, when controlling for heterogeneity, should be the one with

[26] As the variance of unobservables ($\theta$) approaches 0, the proportional hazards property returns.

[27] Appendix B outlines the frailty models considered and the assumed distributions for unobserved heterogeneity (Gamma vs. Inverse Gaussian).

least variation in the unobservables. Most scholars suggest that interpreting the sign and significance of the coefficients should be the limit for substantive interpretation of frailty models.

# 4 Career Duration Results

This section discusses the results from the analysis of individual player career duration and the determinants of their survival in the top leagues of Western European countries. All model specifications include player-specific variables as regressors that are divided into two groups: (i) player's initial conditions including personal characteristics, youth team that trained him, position played, and nationality, and (ii) player's productivity measures. The main objective in all specifications is to assess the effect of youth training on the duration of a player's professional career in the top leagues.

Since all of the models presented in the paper are proportional hazard (PH) models, the results are displayed as hazard ratios instead of the actual covariate coefficients. Hazard ratios are very useful for interpretation purposes.

## 4.1 Parametric Model

The Generalized Gamma model was estimated to test whether one of the nested models is appropriate, and the results indicate that none of the nested parametric models are satisfactory. Notwithstanding these results, AIC was used to adjudicate between nested and non-nested parametric models. The Gompertz parametric specification, which is not nested in the Generalized Gamma model, had the most desirable AIC score. The AIC scores for all parametric models considered are presented in table 3.[28]

Altogether there are five different Gompertz specifications considered for career duration:

- M1: initial conditions only.
- M2: initial conditions and productivity measure variables.

[28] AIC scores are provided for M2 and M3 specifications only. The other specifications considered render the same results.

- M3: initial conditions and productivity measures, with the addition of nonlinear functional forms.
- M4: initial conditions and productivity measures with heterogeneity.
- M5: full model including initial conditions, productivity measures, nonlinear functional forms, and heterogeneity.

The results of these Gompertz specifications are presented in Table 5. The specification with initial conditions (M1) serves as the benchmark model in this paper. It is evident that the youth team variables of primary interest become more significant as the model specification improves in terms of log-likelihood. For example, the M3 specification results in all youth team variable coefficients being significant at the given levels (with all but one significant at the 0.05 level), and the M5 specification suggests that all but one of these youth team coefficients is significant. This finding indicates that the choice of club at youth level is an important determinant in the career duration of professional soccer players. Taking one of the most renowned English clubs in history, Manchester United (Man U), the results suggest that attending their youth program would decrease the exit probability of players by anywhere from 31% to 43%, depending on the specification, compared to the players attending the Sheffield Wednesday youth program (benchmark youth team in this paper). Perhaps surprisingly, some of the historical powerhouses of English soccer (Man U, Chelsea, Everton, and Liverpool) lag behind some of the clubs (i.e. Leeds U and West Ham) with a lesser historical reputation in the top tier of English soccer. We can see that clubs like Arsenal, West Ham, and Leeds United adhere to the sound reputation of their youth programs based on the results from this analysis. Players that attended Leeds United youth training, as opposed to Sheffield Wednesday, decrease their exit probability from the top leagues by anywhere from 49% to 64%. Therefore, quality of youth training varies by club and a player's choice of club (if he has a choice) at youth level can have a significant effect on his career length as a professional. The importance of club choice at youth level is emphasized even when individual-level heterogeneity is accounted for, which should capture the unobserved differential in talent among the players in the data set.

Other than in the benchmark model, the position dummies are significant (at 0.01 level of significance) in all specifications indicating that the position played has an impact on career duration of professional soccer players. It is not surprising that the career duration of goal tenders is significantly longer than for any other position, and these results are consistent with the Frick, Pietzner, and Prinz's (2007) findings for German Bundesliga players. It is quite common for goalkeepers to play well into their late thirties and early forties, which is a rarity for players in other positions. Goalkeepers experience less intense training and physical strain in games, which allows them to preserve their bodies and extend their career. Forwards tend to have the shortest career duration, which is also not surprising considering the scrutiny they face and attention they receive due to the fact their productivity is easily measured and monitored via productivity measures like goals scored.

Discrimination based on nationality does not seem to play a role in most of the specifications. M3 and M5 specifications indicate that Irish nationals have a lower expected career duration (at 0.1 level of confidence) compared to the UK nationals, but there is no strong and consistent evidence of discrimination. International appearances are significant in all specifications, indicating that making one additional international appearance reduces a player's exit probability by around 2%. This result makes sense because players making international appearances are usually the cream of the crop from their respective countries and are most likely to find first team opportunities in the top leagues across Europe. Granted, this might not always be the case since playing on the national teams of Northern Ireland or Wales is much easier to accomplish than playing for England. In general, players making international appearances for any country are consistent top league participants. Average first-team appearances per season also has a predictable and highly significant negative effect on the hazard rate in all specifications. Average goals scored per season is weakly significant for all specifications and the impact on the hazard rate is negative as one would expect. This productivity measure is mostly applicable to forwards, so it is not surprising that it's not significant for all the specified levels of significance (0.05 and 0.01). Age of entry has a strong positive impact on the hazard rate ranging from an increase in exit probability of 15% to 31% for an additional year added to the age of entry compared to the average player. Year of entry,

a variable that can indirectly capture institutional changes in English soccer, has a negative effect on the probability of exit for each additional year in M2, M3, and M5. However, the statistical significance (two out of the four spline terms are not significant) in M3 and M5 is not sufficient to make any reliable inference about the effect of the covariate. The results from M2 suggest a slight negative effect on the hazard rate meaning that players entering during the latter years in this study have longer expected career duration. This result sounds reasonable since many of the institutional changes (i.e. league changes that increased the number of substitutions allowed in a game and increased the number of teams in the top league) in England over the years increased the number of places and possibilities for players in the top tier.

Model specification is improved by allowing for nonlinear functional forms for the average appearances and the year of entry variables. Comparing specifications M2 and M3 indicates that allowing for nonlinear functional form of the above variables has a significant effect on the magnitude of youth team coefficient estimates (and the resulting hazard ratios) for most clubs, with West Brom and Southampton being least effected. A similar observation can be made when comparing the results from the models with unobserved heterogeneity, M4 and M5, by looking at the position dummies. The productivity measure covariates seem to be most robust across specifications.

In terms of log-likelihood, model specification is greatly improved by adding productivity variables and nonlinear functional forms to the benchmark model.

Accounting for unobserved heterogeneity to M2 and M3 specifications also results in an improvement, but not to as great an extent as the previous addition of productivity variables.[29] It is interesting that the M3 specification without heterogeneity outperforms the M4 frailty specification in terms of log-likelihood, suggesting that modeling nonlinearity is more important than accounting for unobserved heterogeneity. The Gompertz distribution shape parameter $\gamma$ is larger in the frailty models (M4 & M5) than in the reference models (M2 & M3), meaning that the baseline hazard slopes upwards to a greater extent in models with unobserved heterogeneity. The estimated parameter $\theta$,[30] a measure of heterogeneity, is 0.34238 and 0.21685 for the M4 and M5 specifications, respectively, and statistically significant at the 0.01 level. Less unobserved heterogeneity in a

[29] This is especially the case for the M3 and M5 models where the difference in log-likelihood is almost negligible.
[30] $\theta$ is defined as the variance of the gamma distribution for unobserved heterogeneity.

model is preferred (as is the case for M5), which is additional evidence that the models allowing for nonlinear functional forms of covariates result in an improvement in specification and are preferred to the models that have linear functional form for the relevant covariates (M4). Additionally, the less heterogeneity in a model allows for more appropriate interpretation of any observed duration dependence. One has to be careful in the presence of frailty, as the interpretation in terms of hazard ratios is lost because the proportional effect of a given regressor is no longer constant and independent of survival time. Typically, most researchers limit their analysis to interpreting the signs of coefficients for frailty models. Therefore, the M3 specification serves as the best model for interpretation of the estimated hazard ratios considering the almost negligible difference in log-likelihood from M5.

Figure 9 graphically displays the estimated hazards of the various models versus the empirical hazard. M2 and M3 slightly overpredict the empirical hazard for the most part but they model the slope closely, especially M3. Models with unobserved heterogeneity, M4 and M5, scale up the estimated reference models (M2 and M3) resulting in significant overprediction, especially M4. The baseline hazards for the Gompertz specifications are presented in Figure 10. The shape of the baseline hazards for the parametric specifications accurately model duration dependence, suggesting there is positive state dependence. There is no indication that the Gompertz model is misspecified. The underlying time dependency seems to be properly characterized, which will generally result in more precise estimates than the ones produced by semi-parametric and nonparametric models where the underlying time-dependency is left unspecified. This hypothesis will be tested by choosing a Cox model with a flexible baseline in the following sub-section.

#### 4.1.1 Semiparametric Model

The Cox proportional hazard semiparametric model is estimated as an alternative to the Gompertz parametric model. There will be three different specifications for the Cox PH model:

- M1: initial conditions only.
- M2: initial conditions and productivity measure variables.

- M3: full model including initial conditions, productivity measures and nonlinear functional form for certain covariates.

The results for the semiparametric model are presented in table 6. The signs and significance of the estimated coefficients are not much different from the Gompertz parametric estimates. There is a difference in the magnitude of the hazard ratio estimates, especially for the position dummies. The covariate measuring international appearances is insignificant at all reasonable levels for M2, which is not the expected result.[31] The importance of youth training is compounded in the semiparametric specifications. The magnitude of the hazard ratios for M3 is a little more amplified in the parametric model than in the semiparametric model. The main difference lies in the M2 specification, where the semiparametric model results in five youth team regressors being insignificant at any conventional levels, and three youthteam regressors significant only at the 0.1 level. This is a stark difference from the M2 parametric specification where only one youth team regressor is not significant at the 0.05 level. This is an indication that M2 is misspecified in favor of M3 in the semiparametric model setting. Looking at M3 results, the solid youth training reputation of clubs like Arsenal, West Ham, and Leeds United is upheld in the Cox PH model. Only one youth team regressor (Manchester City) is not significant at the 0.05 level, and it happens to be the same insignificant regressor as in the Gompertz M3 parametric specification. The shared measure of heterogeneity was insignificant for the Cox PH model.

One of the main assumptions of the Cox semiparametric model is that the hazard rates for two observations are proportional to one another and that this proportionality is maintained over time. Testing this assumption is a key aspect of the analysis. The Therneau and Grambsch nonproportionality test suggest that the PH assumption is violated in M2 and M3, both globally and at individual covariate level. The PH assumption is violated globally for M2 at any conventional level of significance, while the M3 global test suggests that the assumption just fails at the 0.05 level of significance. This result outlines the interaction between model specification and the PH assumption. Keele (2010) argues that this test detects a number of specification errors besides

[31] The insignificance of this covariate is a result of introducing the interactive term of appearances per season and log-duration.

the tested objective. Even though the addition of nonlinear functional forms (for average season appearances and year of entry) in M3 does not result in the upholding of the PH assumption, the improvement in the global test suggests that model specification has an impact on this assumption. In terms of individual covariates, average appearances and international appearances fail the PH tests in M2 and M3, respectively. The common solution, implemented in M2 and M3, is to interact the covariates that show signs of nonproportional hazards with the natural log of time. These interactions are highly significant and are displayed at the bottom of the results table for M2 and M3.

In terms of log-likelihood, both M2 and M3 are an improvement to the benchmark specification M1. Surprisingly, M2 outperforms M3 in terms of log-likelihood. This is even more surprising considering the visual evidence for the baseline and estimated hazards of the three specifications. Figure 11 displays the estimated hazard for M3 against the empirical hazard, and the baseline hazards are presented in Figure 12. The estimated hazards for M1 and M2 are so far off base that they can't even be presented in the same graph with the estimated hazard for M3 and the empirical hazard. This is not surprising considering their baseline hazards provide a poor measure of duration dependence. M3 mimics the shape of the empirical hazard, but overpredicts it ever so slightly. It is evident that M3 models duration dependence far better than the other two specifications. Visual inspection undoubtedly concludes that M3 outperforms M1 and M2 for the semiparametric model.

In addition to the visual techniques above, Cox-Snell residuals are used to assess the goodness of fit for the semiparametric model. The idea is that if the Cox PH model fits the data, then these residuals should be distributed unit exponentially. Plotting the integrated hazard based on these residuals against the hazard rate estimates backed out of the Cox model should result in a 45 degree slope if the Cox specification is appropriate. Figure 13 displays the results of this test for M3.[32] The divergence from the 45 degree line midway suggests that the Cox model does not fit the data very well.

[32]Similar results are obtained for the other specifications but not presented here.

## 4.2 Model Evaluation

The violation of the PH assumption (even though it was addressed by including interactive terms between the offending covariates and the log of duration time) and the unsatisfactory fit of the Cox semiparametric model, coupled with the reasonable estimation of duration dependence by the Gompertz model, suggest that the parametric model provides a better specification and fit of the data in this analysis. In addition to visual inference about the slope and efficiency of the estimated hazards, Vuong's test for discriminating between rival non-nested models is performed.[33] Vuong's model selection criteria identifies the model that is closer to the true specification.

In selecting the most appropriate model preference is given to the slope of the estimated hazard because it may be used to identify true duration dependence. Figure 14 contains the estimated hazards for the M3 Cox and Gompertz specifications against the empirical hazard. Even though the estimated hazards for the two competing models are very close, the Gompertz model seems to be more efficient (varies less with the data) for the large portion of the examination period. With respect to duration dependence, the Gompertz model is the better predictor as it mimics the slope of the empirical hazard more closely than the Cox semiparametric model. Thus, the Gompertz parametric model is more reliable in explaining the career duration of soccer players in this study.

Vuong's criteria implies that one model should be selected over another if the average log-likelihood of that model is significantly greater than the average log-likelihood of the rival model.[34] The Vuong test statistic is simply the normalized ratio of the rival models' average log-likelihoods. The details of the Vuong test are explained in the appendix C. The calculated Vuong statistic is sufficiently large and negative[35] indicating that the Gompertz parametric model is closer to the true specification than the Cox model. This test confirms the intuition from the above results and

[33]Other tests, like the Cox test, were considered but the Cox test may reject both models, whereas Vuong's test selects the model that is closer to the true specification even if both models are far from the true specification. Preference is given here to Vuong's relative test that evaluates the models against the data and each other versus Cox's absolute test that only evaluates the models against the data.

[34]Comparing two distributions $H_c$ and $H_g$, there are three possible outcomes:

1. The two distributions are equal.
2. $H_c$ is better than $H_g$.
3. $H_c$ is worse than $H_g$.

[35]The calculated Vuong statistic = -718 and is statistically significant at all conventional critical values for the normal distribution (i.e. -1.96 = 0.05 level of significance). The high negative value suggests that the alternative hypothesis (look in the appendix) is true.

leads me to rely on the Gompertz model to make inference on the effect of youth training on career duration of professional soccer players.

# 5 Home-Team Spell Duration Results

This section discusses the results from the analysis of individual player spell duration at his home club (the club that trained the player). All model specifications include player-specific variables as regressors that are divided into two groups: (i) player's initial conditions including personal characteristics, youth team that trained him, position played, and nationality, and (ii) player's productivity measures. The main objective here is to assess which clubs tend to rely on their youth systems more heavily in terms of their first team composition. It is one thing for a club to produce quality players in general, but it is just as important to hold on to these players for as long as it makes sense from a competitive and financial perspective. Allowing more first team opportunities to players from their youth academies can serve as a signal to players at a young age in terms of choosing the youth program to join.

## 5.1 Parametric Models Considered

The AIC scores for all parametric models considered are presented in Table 4. The Generalized Gamma model had the most desirable AIC score out of the models considered, followed by some models that are nested within it like Weibull. Individual tests for the models nested within the Generalized Gamma model indicate that none of the nested models are satisfactory. None of the non-nested models are considered since their AIC scores indicate the fit of those models is even less satisfactory than the nested models. Despite the indications that a parametric model might not be appropriate here, the Generalized Gamma model is estimated using the following three specifications:

- M1: initial conditions only.
- M2: initial conditions and productivity measure variables.

- M3: full model including initial conditions, productivity measures and nonlinear functional form for certain covariates.

Figure 15 outlines the baseline hazards for the estimated Gamma specifications. The Gamma baselines suggest that duration dependence is poorly estimated. The characterization of the underlying time-dependency is inaccurately measured by the Gamma model, which would generally result in imprecise estimates generated by this model. A similar analysis was performed for the Weibull (nested) and Gompertz (non-nested) models[36] resulting in similar conclusions that the models are not appropriate for this analysis. Therefore, the visual examination of the estimation results and the AIC score tests suggest that none of the parametric models considered are suitable for the analysis of HT spell duration. This hypothesis will be tested in the next subsection by choosing a semiparametric Cox model with a flexible baseline.

## 5.2 Semiparametric Model

Once again, three different specifications are estimated for the Cox PH spell duration model:

- M1: initial conditions only.
- M2: initial conditions and productivity measure variables.
- M3: full model including initial conditions, productivity measures and nonlinear functional form for certain covariates.

The results of the estimations are presented in Table 7. The results are explained in terms of M3 because, as it will be explained subsequently, this specification is clearly superior to the other two. Two thirds (10 out of 15) of the youth team regressors are statistically significant at conventional levels. The interpretation in this analysis can be linked to an element of loyalty or bond created between youth trainees and their home teams, as this analysis examines a player's duration at the

[36]The figures are not presented here. The Weibull baseline hazards do not characterize duration dependency accurately, while the Gompertz baselines are slightly better. The estimated hazards for both models are poor measures of the empirical hazard, indicating that the models do not provide for a reliable fit of the data. Controlling for unobserved heterogeneity does not results in a improvement.

home team that provided youth training for that particular player. Clubs like Arsenal, West Ham and Leeds United seem to maintain the strongest bond with the players they train at youth level. This kind of loyalty creation would certainly aid the success and financial stability of these clubs in the long-run, as they wouldn't be overly reliant on transfer market spending and taking risks on players from outside their program.

Arsenal has a history (particularly in the last 15-20 years) of producing and maintaining quality players from their youth program in their first team, and they seem to be one of the most competitive and financially stable clubs in England in the recent era.[37] On the other hand, a club like Everton, that has the longest history of appearances in the top tier of English soccer, seems to be on the low end of the statistically significant youth team regressors for this analysis. This suggests that they haven't been able to hold on to their youth products for a sufficiently long period of time compared to some of the other teams in the analysis. It is not surprising that the expected spell duration is larger for some of the smaller clubs (i.e. West Ham, Leeds United, and Southampton) in this analysis versus some larger clubs (i.e. Manchester United or Liverpool). A club like Manchester United has enjoyed a great deal of success in the past, both domestically and internationally, and has been financially superior to most other clubs in England allowing them to acquire players in the transfer market more freely. Naturally, it is much more difficult for youth trainees to establish themselves at larger clubs like Manchester United due to the more competitive first team environment. Smaller clubs that are not as financially endowed have to rely on domestic youth products in order to remain competitive. Therefore, loyalty is not the only factor to consider when interpreting the results of this analysis. Larger clubs might create the same or even higher degree of loyalty with their youth players, but their financial strength and ambitions make it much more difficult for youth players to break into the first team and remain there for an extended period of time. Hence, the difference may be that smaller clubs are more reliant on their youth system due to their financial inability to compete for top talent in the transfer market.

[37]The current Arsenal manager, Arsene Wenger, has developed a strong youth program at the club and has emphasized relying on home grown talent versus active purchasing in the transfer market. In fact, Arsenal were one of the few English clubs in recent years to report profits instead of losses.

Except for wing players, all of the position dummies are highly significant. The goal tenders tend to have the longest expected duration with their home team, while the forwards (benchmark), unsurprisingly, have the shortest expected home spell duration compared to the other positions. The nationality dummies are not significant at any conventional level indicating that discrimination is not present in the scope of this analysis. productivity measure variable,[38] average appearances per season, has a strong and expected negative effect on the hazard rate probability or positive impact on expected spell duration. Age of entry has a strong and expected positive effect on the hazard rate or a negative effect on expected spell duration.

Year of entry, which can be thought of as a variable capturing institutional changes in English soccer over time, has a statistically significant positive effect on the hazard rate probability or a negative effect on expected spell duration. This suggests that players entering the professional soccer market in the latter years of this study have a reduced probability of survival in the top leagues playing for their home team than players in the past. This seems like a plausible result considering that institutional changes in England and Europe have made the domestic soccer markets more integrated and competitive over the years of this study, making it that much harder for players to maintain their places in their respective teams. These changes would subject players to more movement in the market and, as a result, shorter expected spell durations during their career.

In terms of log-likelihood, M2 and M3 are significant improvements on the benchmark specification, M1. The proportionality of hazards assumption is violated for M2, both globally and locally for the appearances per season covariate. However, the PH assumption holds up with a high level of significance for M3, both globally and locally for each covariate. As mentioned earlier, the nonproportionality test detects a number of other specification errors and this appears to be evident here as M2 appears to be misspecified. M3 allows for nonlinear functional form for average appearances and year of entry covariates, and this specification clearly satisfies the PH assumption for the Cox model. Therefore, relying on the nonproportionality test results from M2 would erroneously conclude that the PH assumption is violated rather than that the model is misspecified. In addition, the strong evidence that the PH assumption holds indicates that the variance of unobservables is

[38] International appearances and average goals per season are counted for the player's entire career and are not available for this spell analysis.

negligent and that heterogeneity does not pose an empirical threat to the estimation results. This assumption is further solidified by the goodness of fit test results. Hence, there is no need to model unobserved heterogeneity independently considering the quality of the M3 specification. Figures 16 and 17 provide further evidence that M3 is superior to the other two specifications in terms of modeling duration dependence and model fit.

The goodness of fit test using Cox-Snell residuals for M3 is presented in figure 18. The 45 degree slope is closely matched by the integrated hazard based on these residuals. The test and the resulting graph verify the statements made above about the quality of Cox model specification for this spell duration analysis.

### 5.2.1 Endogeneity Testing

A split sample test is used to test if the model specifications for both career durations and home team spell duration suffers from endogeneity bias. The test follows the same steps as the one proposed by Huynh, Petrunia and Voia (2010):

1. Randomly split the sample of soccer players in two equal parts.

2. Estimate the career and home team duration models using the covariates ($x^{(1)}$) from the first sample and retrieve the estimated coefficients ($\hat{\beta}_1^{(1)}$).

3. Use the estimates from sample one, $\hat{\beta}_1^{(1)}$, and the sample two information ($x^{(2)}$) to create predicted durations, then with the data from sample two, generate a variable which is the difference between the actual and the predicted durations: $\log(t^{(2*)}) = \log(t^{(2)}) - x'^{(2)}\hat{\beta}_1^{(1)}$.

4. The new outcome variable $\log(t^{(2*)})$ is then regressed against the same covariates from the two models using the data from sample two: $\log(t^{(2*)}) = -x'^{(2)}\beta_1^{(2)} + \log(u^{(2)})$.

5. A $\chi^2(m)$ test is constructed, where $m$ is the number of variables, with the null hypothesis of no bias or $H_0 : \beta_1^{(2)} = 0$. If there is a systemic bias induced by the parametric unobserved heterogeneity assumption, the test statistic would be rejected.

More details about the motivation and constrained of this test statistic can be found in the appendix. Note that the same test is used for both duration problems as no formal test can be used for the different parametric and semiparametric duration models. This test however can give an idea of what variables are potential candidates to be endogenous given that the test is performed on an approximately equivalent linear duration model specification. This linear specification can be viewed as an upper bound for the endogeneity bias. As the two preferred models are nonlinear, the effect of the potential endogeneity bias will be lower than in the tested model. The results of the split sample tests are presented in Tables 9 and 10. The results of the test suggest that career duration model suffers from less bias than the home team duration model and that the sources of bias may be found in some of the home team dummies.

# 6 Conclusion

There have not been many previous studies in economic literature examining the career duration of professional athletes. Little is known about the impact of education/training at the early stages of an athlete's career on his survival as a professional. This paper emphasizes the importance of pre-professional training (within specialized soccer academies and youth programs) on a soccer player's survival probability in the top Western European leagues. A unique data set of post-war English trained soccer players is used to estimate their career and spell duration. One contribution of this paper is that information on youth training program participation is included in the duration models. The results of model estimation suggest that the participation in youth academy or program is an important determinant on a player's career and spell duration in a top European league. Moreover, the findings indicate that certain clubs, that are renowned for their youth academies, outperform other clubs in the study in terms of career duration for the players that come out of their system. The spell duration analysis outlines the nature of the competitive environment in which smaller clubs have a chance to keep up with the larger and financially superior clubs; the results suggest that some smaller clubs outperform larger clubs in terms of producing and holding on to home-grown talent, which would be a necessary condition for them to remain competitive in

light of their lagging financial resources that limit their transfer market activity.

In addition to youth training, position played is an important determinant of a player's career and spell duration. The findings indicate that forwards have the shortest spell and career duration among the positions considered, which is the expected result considering their productivity is easiest to measure by readily available statistics, like goals scored. Variables measuring productivity have a strong positive effect on a player's expected survival. The year a player enters the market, which serves as an indirect measure of institutional changes in English soccer over the time span of this study, has opposing effects for career and spell duration. Institutional changes have increased the competitive nature of the English and European soccer market. The year of entry effect is positive for a player's career duration and negative for his spell duration. This is not surprising considering that institutional changes increased the number of possibilities for players to remain in the top leagues, but at the same time the enhanced competitive nature of the market subjected the players to more movement within the market that reduces their spell duration. One has to be careful in interpreting this variable because the majority of the players in the data base were not affected by the changes arousing as a result of the Bosman ruling (1995/1996), which has significantly altered the European soccer environment. Understandingly, discrimination based on nationality does not seem to have an important impact, considering the majority of the players in this study are UK nationals. This would be a more interesting question to address in the post-Bosman environment where the presence of foreigners increased drastically.

The duration models considered provide sufficient evidence of nonlinear effects for some of the covariates in this study. Nonlinear functional form of covariates (productivity measures) improves model specification and fit significantly. The nonlinear effects of the productivity measures have similar impacts on the HT and Career duration. The presence of individual-level unobserved heterogeneity for career duration points to other factors that might be important for a player's survival. Some of these factors might be a player's wages and transfer fees paid during his career. However, individual talent and tactical knowledge of a particular player is difficult to measure and will always provide a source of unobserved heterogeneity. Interestingly, the results indicate that a parametric model provides for best fit of the career duration data, while a semiparametric model is most

suitable specification for spell duration data.

## References


[1] Ashworth, J., and Heyndels, B. (2007): "Selection Bias and Peer Effects in Team Sports: the Effect of Age Grouping on Earnings of German Soccer Players," *Journal of Sports Economics*, 8 (4), 355-377.

[2] Atkinson, S., and Tschirhart, J. (1986): "Flexible Modeling of Time to Failure in Risky Careers," *Review of Economics and Statistics*, November, 558-566.

[3] Dufour, J.-M., and J. Jasiak (2001): "Finite Sample Limited Information Inference Methods for Structural Equations and Models with Generated Regressors," *International Economic Review*, 42(3), 81543.

[4] Elbers, C., and Ridder, G. (1982): "True or Spurious Duration Dependence: The Indentifiability of the Proportional Hazard Model," *Review of Economic Studies*, 49, 402-409.

[5] Frick, B., Pietzner, G., and Prinz, J. (2007): "Career Duration in a Competitive Environment: The Labor Market for Soccer Players in Germany," *Eastern Economic Journal*, 33(3).

[6] Geyer, H. (2010): "Quit Behavior of Professional Tennis Players," *Journal of Sports Economics*, 11(1), 89-99.

[7] Gramsch, P.M., and Therneau, T.M. (1994): "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515-526.

[8] Gramsch, P.M., and Therneau, T.M. (2000): "Modeling Survival Data: Extending the Cox Model," New York, Springer-Verlag.

[9] Groothuis, P.A., and Hill, J.R. (2004): "Exit Discrimination in the NBA: A Duration Analysis of Career Length," *Economic Inquiry*, 341-349.

[10] Hastie, T.J., and Tibishirani, R.J. (1990): "Generalized Additive Models," Chapman and Hall, London.

[11] Heckman, J.J., and Singer, B. (1982): "The Identification Problem in Econometric Models for Duration Data," *Advances in Econometrics*, edited by W. Hildenbrand. Cambridge University Press.

[12] Huynh, K. P.,Petrunia, R.J. and Voia, M.C. (2010): The Impact of Initial Financial State on Firm Duration Across Entry Cohorts , *Journal of Industrial Economics*, Vol.LVIII, no.3.

[13] Jenkins, S.P. (2008): "Survival Analysis," Unpublished Manuscript, Institute for Social and Economic Research, University of Essex, Colchester.

[14] Keele, L. (2010): Proportionality Difficult: Testing for Nonproportional Hazards in Cox Models," *Political Analysis*, 18, 189-205.

[15] Noll, R.G (2002): "The Economics of Promotion and Relegation in Sports Leagues: The Case of English Football," *The Journal of Sports Economics*, May, 169-203.

[16] Okhusa, Y. (2001): "An Empirical Examination of Quit Behavior of Professional Baseball Players in Japan," *The Journal of Sports Economics*, May, 80-88.

[17] Royston, P., and Sauerbrei W. (2007): "Multivariable Modeling with Cubic Regression Splines: A Principled Approach," *The Stata Journal*, 7(1), 45-70.

[18] Spurr, S., and Barber, W. (1994): "The Effect of productivity on a Worker's Career: Evidence from Minor League Baseball," *Industrial and Labor Relations Review*, July, 692-708.

[19] Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

[20] Zorn, C.J. (2000): "Modeling Duration Dependence," *Political Analysis*, 8, 367-380.

Table 1: Summary Statistics

| **Variable** | **Mean** | **Std. Dev.** | **Min.** | **Max.** |
|---|---|---|---|---|
| HT duration | 4.517 | 3.592 | 0 | 19 |
| totaldur | 6.374 | 4.846 | 1 | 21 |
| HT apps | 82.108 | 112.166 | 0 | 613 |
| totalapps | 126.616 | 148.904 | 1 | 714 |
| intcaps | 4.921 | 14.804 | 0 | 108 |
| goals | 15.202 | 33.453 | 0 | 366 |
| ageentry | 18.602 | 1.751 | 15 | 29 |
| yearentry | 1974.084 | 13.056 | 1946 | 2003 |

Table 2: Hazard Summary for Players' Career and HT Spell Duration

| | **Career Duration** | | | **Home Team Spell Duration** | | |
|---|---|---|---|---|---|---|
| **time** | **n** | **exits** | **h** | **n** | **exits** | **h** |
| 1 | 1121 | 187 | 16.68% | 1098 | 227 | 20.67% |
| 2 | 934 | 116 | 12.42% | 871 | 140 | 16.07% |
| 3 | 818 | 122 | 14.91% | 731 | 164 | 22.44% |
| 4 | 696 | 93 | 13.36% | 567 | 120 | 21.16% |
| 5 | 603 | 85 | 14.10% | 447 | 119 | 26.62% |
| 6 | 518 | 70 | 13.51% | 328 | 70 | 21.34% |
| 7 | 448 | 52 | 11.61% | 258 | 53 | 20.54% |
| 8 | 396 | 55 | 13.89% | 205 | 53 | 25.85% |
| 9 | 341 | 44 | 12.90% | 152 | 35 | 23.03% |
| 10 | 297 | 50 | 16.84% | 117 | 30 | 25.64% |
| 11 | 247 | 40 | 16.19% | 87 | 19 | 21.84% |
| 12 | 207 | 45 | 21.74% | 68 | 19 | 27.94% |
| 13 | 162 | 34 | 20.99% | 49 | 13 | 26.53% |
| 14 | 128 | 34 | 26.56% | 36 | 13 | 36.11% |
| 15 | 94 | 31 | 32.98% | 23 | 8 | 34.78% |
| 16 | 63 | 26 | 41.27% | 15 | 7 | 46.67% |
| 17 | 37 | 19 | 51.35% | 8 | 5 | 62.50% |
| 18 | 18 | 7 | 38.89% | 3 | 2 | 66.67% |
| 19 | 11 | 7 | 63.64% | 1 | 1 | 100.00% |
| 20 | 4 | 2 | 50.00% | | | |
| 21 | 2 | 2 | 100.00% | | | |

Table 3: AIC Scores for Career Duration

| **AIC** | **M2** | **M3** |
|---|---|---|
| Exponential | 2609 | 2581 |
| Weibull | 1877 | 1738 |
| **Gompertz** | **1761** | **1687** |
| Lognormal | 1886 | 1765 |
| Loglogistic | 1902 | 1748 |
| Generalized Gamma | 1854 | 1721 |

Table 4: AIC Scores for HT Spell Duration

| **AIC** | **M2** | **M3** |
|---|---|---|
| Exponential | 2576 | 2552 |
| Weibull | 1945 | 1846 |
| Gompertz | 2061 | 2017 |
| Lognormal | 1909 | 1803 |
| Loglogistic | 1948 | 1823 |
| **Generalized Gamma** | **1895** | **1789** |

Table 5: Gompertz Model Career Duration Results

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| | | | **Initial Conditions** | | |
| *Personal Characteristics* | | | | | |
| ageentry | 1.1511*** | 1.1994*** | 1.2237*** | 1.31*** | 1.2918*** |
| yearentry | 1.0043* | .99372** | | 1.310 | |
| yearentry-0 | | | .95421 | | .94277 |
| yearentry-1 | | | 1.1221*** | | 1.1134*** |
| yearentry-2 | | | .92539** | | .94798 |
| yearentry-3 | | | 1.0833** | | 1.1061*** |
| *Youthteam* | | | | | |
| Arsenal | .54813*** | .43828*** | .42673*** | .39180*** | .39718*** |
| Chelsea | .70281** | .44580*** | .46533*** | .48916*** | .50127*** |
| Everton | .70911** | .55310*** | .48803*** | .51185*** | .47166*** |
| Liverpool | .56444*** | .50684*** | .45144*** | .42595*** | .40342*** |
| Man U | .68828** | .58928*** | .63445*** | .57312*** | .61141** |
| Man City | .78707 | .79508 | .73962* | .77219 | .73905 |
| Tottenham | .69011** | .41249*** | .42179*** | .39229*** | .39672*** |
| West Ham | .63429** | .47947*** | .44320*** | .42991*** | .41521*** |
| Aston Villa | .89648 | .60726*** | .63417** | .75109 | .70563* |
| Leeds U | .50532*** | .47420*** | .38232*** | .42310*** | .36186*** |
| Newcastle | 1.0839 | .57734*** | .51702*** | .56198** | .50228*** |
| Southampton | .67142** | .61765** | .61364** | .55923** | .56749** |
| Nottingham F | .66900** | .60185*** | .56457*** | .55730** | .54266*** |
| West Brom | .79899 | .55711*** | .55293*** | .54839** | .5450*** |
| Coventry City | .75384 | .61150*** | .56765** | .59185** | .54902*** |
| Sheffield Wed - Benchmark | | | | | |
| *Position* | | | | | |
| Forward | 1.1885 | 2.3143*** | 2.4878*** | 2.2484*** | 2.4176*** |
| Central Defender | 1.0348 | 2.0763*** | 2.1718*** | 2.0546*** | 2.1507*** |
| Midfielder | .90928 | 1.9476*** | 1.9423*** | 1.7277*** | 1.8195*** |
| Winger | 1.4194** | 2.3548*** | 2.3483*** | 2.1382*** | 1.7705*** |
| Full Back | 1.0017 | 1.6926*** | 1.8601*** | 1.6057** | 2.2565*** |
| Goalie - Benchmark | | | | | |
| *Nationality* | | | | | |
| Ireland | .65846** | 1.4210* | 1.4786* | 1.4334 | 1.4937* |
| Other | .62523 | 1.2069 | 1.1149 | .93559 | .99234 |
| UK - Benchmark | | | | | |
| | | | **productivity Measures** | | |
| intcaps | | .98736*** | .98378*** | .98025*** | .97947*** |
| avggoals | | .96706* | .96580* | .95637* | .95974* |
| avgapps | | .87768*** | | .84766*** | |
| avgapps-0 | | | .23585*** | | .20696*** |
| avgapps-1 | | | .82678*** | | .78366*** |
| Obs. | 1121 | 1121 | 1121 | 1121 | 1121 |
| $\gamma$ | .09355 | .28058 | .29614 | .3975 | .36817 |
| $\theta$ (variance of $\gamma$) | | | | .34238*** | .21685*** |
| $\Delta \log L$ (vs. M1) | 0 | 545 | 586 | 567 | 594 |

Note: *, **, and *** indicates statistical significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table 6: Cox Model Career Duration Results

| | M1 | M2 | M3 |
|---|---|---|---|
| | **Initial Conditions** | | |
| *Personal Characteristics* | | | |
| ageentry | 1.1340*** | 1.0812*** | 1.2004*** |
| yearentry | 1.00343 | .99046*** | |
| yearentry-0 | | | .95394 |
| yearentry-1 | | | 1.1019*** |
| yearentry-2 | | | .93755** |
| yearentry-3 | | | 1.0717** |
| *Youthteam* | | | |
| Arsenal | .56665*** | .58522*** | .47278*** |
| Chelsea | .70153** | .45799*** | .52591*** |
| Everton | .73372* | .69436** | .53494*** |
| Liverpool | .59832*** | .73998 | .53494*** |
| Man U | .69301** | .64518*** | .66539** |
| Man City | .81303 | .80889 | .74790 |
| Tottenham | .71118** | .55023*** | .46121*** |
| West Ham | .63538** | .64693** | .49416*** |
| Aston Villa | .89724 | .80199 | .69232** |
| Leeds U | .52293*** | .71645* | .41984*** |
| Newcastle | 1.0607 | .71170* | .56546*** |
| Southampton | .67954* | .76599 | .66030** |
| Nottingham F | .69793* | .69942* | .60343*** |
| West Brom | .77666 | .64168** | .58881*** |
| Coventry City | .76893 | .75977 | .61448*** |
| Sheffield Wed - Benchmark | | | |
| *Position* | | | |
| Forward | 1.1670 | 1.6565*** | 2.1073*** |
| Central Defender | 1.0335 | 1.7534*** | 1.9165*** |
| Midfielder | .91732 | 1.6255*** | 1.7404*** |
| Winger | 1.3753** | 1.7889*** | 1.9827*** |
| Full Back | 1.0239 | 1.4678*** | 1.6728*** |
| Goalie - Benchmark | | | |
| *Nationality* | | | |
| Ireland | .67423** | 1.2139 | 1.1662 |
| Other | .69349 | 1.3448 | .98457 |
| UK - Benchmark | | | |
| | **productivity Measures** | | |
| avgapps | | 1.13706*** | |
| avgapps-0 | | | .30838*** |
| avgapps-1 | | | .8010*** |
| intcaps | | .99768 | 1.0649*** |
| avggoals | | 1.0231 | .97120 |
| *Variable Interactions with Log(Duration)* | | | |
| time-intcaps | | | .97152*** |
| time-avgapps | | .88711*** | |
| Obs. | 1121 | 1121 | 1121 |
| $\Delta \log L$ (vs. M1) | 0 | 752 | 474 |

Note: *, **, and *** indicates statistical significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table 7: Cox Model HT Spell Duration Results

| | M1 | M2 | M3 |
|---|---|---|---|
| | **Initial Conditions** | | |
| *Personal Characteristics* | | | |
| ageentry | 1.17171*** | 1.1042*** | 1.2056*** |
| yearentry | 1.0145*** | 1.0063** | |
| yearentry-0 | | | 1.0966*** |
| yearentry-1 | | | 1.1272*** |
| *Youthteam* | | | |
| Arsenal | .67459** | .72631* | .55690*** |
| Chelsea | .92051 | .84145 | .77768 |
| Everton | .88245 | .84801 | .70759** |
| Liverpool | .67131** | .92752 | .60516** |
| Man U | .76975 | .76967 | .66248** |
| Man City | 1.1078 | 1.0088 | .90391 |
| Tottenham | .87777 | .79011 | .57194*** |
| West Ham | .75380 | .69575** | .56634*** |
| Aston Villa | 1.0014 | .88469 | .82026 |
| Leeds U | .67004** | .73692 | .49401*** |
| Newcastle | 1.2453 | .90165 | .73043* |
| Southampton | .65642** | .66727** | .59482*** |
| Nottingham F | .82391 | .85433 | .71133* |
| West Brom | 1.0089 | .84261 | .75521 |
| Coventry City | .94622 | .90667 | .80738 |
| Sheffield Wed - Benchmark | | | |
| *Position* | | | |
| Central Defender | .63422*** | .75201*** | .71533*** |
| Midfielder | .72128*** | .83183* | .78798** |
| Winger | .91498 | .83624 | .83118 |
| Full Back | .66365*** | .67430*** | .66966*** |
| Goalie | .71968** | .67173*** | .592071*** |
| Forward - Benchmark | | | |
| *Nationality* | | | |
| Ireland | .88905 | 1.3140 | 1.1285 |
| Other | .73293 | 1.1270 | 1.0949 |
| UK - Benchmark | | | |
| | **productivity Measures** | | |
| apps-seas | | 1.0673*** | |
| apps-seas-0 | | | .41988*** |
| apps-seas-1 | | | .81575*** |
| *Variable Interaction with Log-Duration* | | | |
| time-apps | | .93759*** | |
| Obs. | 1098 | 1098 | 1098 |
| $\Delta \log L$ (vs. M1) | 0 | 331 | 257 |

Note: *, **, and *** indicates statistical significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table 8: Split Sample Tests

| | home team | overall career |
|---|---|---|
| *Personal Characteristics* | | |
| ageentry | 0.009 | -0.015 |
| yearentry-0 | 0.047 | 0.014 |
| yearentry-1 | -0.015 | -0.007 |
| yearentry-2 | 0.0077 | 0.027 |
| yearentry-3 | -0.008 | -0.037 |
| *Youthteam* | | |
| Arsenal | -0.072 | 0.259** |
| Chelsea | -0.354** | 0.106 |
| Everton | -0.122 | -0.213 |
| Liverpool | -0.229*** | 0.261 * |
| Man U | -0.086 | 0.136 |
| Man City | -0.11 | 0.04 |
| Tottenham | -0.121 | -0.088 |
| West Ham | -0.213** | -0.005 |
| Aston Villa | -0.267 * | -0.025 |
| Leeds U | -0.126 | -0.039 |
| Newcastle | 0.082 | 0.095 |
| Southampton | -0.244 | 0.064 |
| Nottingham F | -0.214 | -0.001 |
| West Brom | -0.331 ** | 0.246** |
| Coventry City | -0.221 | 0.022 |
| Sheffield Wed - Benchmark | | |
| *Position* | | |
| Forward | 0.208 | 0.107 |
| Central Defender | 0.081 | 0.083 |
| Midfielder | 0.109 | -0.023 |
| Winger | 0.057 | 0.029 |
| Full Back | 0.251 * | 0.068 |
| Goalie - Benchmark | | |
| *Nationality* | | |
| Ireland | -0.183 | -0.23 |
| Other | -0.145 | -0.192 |
| UK - Benchmark | | |
| | **productivity Measures** | |
| avgapps-0 | 0.051 | 0.0007 |
| avgapps-1 | -0.013 | -0.022 |
| intcaps | 0.002 | -0.0005 |
| avggoals | -0.016 | -0.0006 |
| constant | -0.131 | 0.196 |
| Obs. | 1121 | 1121 |

Note: *, **, and *** indicates statistical significance based on p-values at the 0.1, 0.05, and 0.01 levels, respectively.

Table 9: Web Sites Used in Data Collection

| **Website** |
|---|
| *http://www.lerwill-life.org.uk/astonvilla/* |
| *http://www.astonvillaplayerdatabase.com* |
| *http://www.statbunker.com/* |
| *http://www.leeds-fans.org.uk/leeds/* |
| *http://www.lfchistory.net/Players/* |
| *http://www.mcfcstats.com/* |
| *http://www.westhamstats.info/westham.php?west=0* |
| *http://www.adrianbullock.com/swfc/stats/swfcarch.htm* |
| *http://www.toon1892.co.uk/* |
| *http://www.sporting-heroes.net/football/* |
| *http://stats.football365.com/hist/default.html* |
| *http://www.soccerbase.com/* |

Table 10: Survival and Hazard Rates for the Analyzed Parametric Models

| Model | $S(t)$ | $S_\theta(t)$ | $h(t)$ |
|---|---|---|---|
| Weibull PH (G) | $e^{-e^{x\beta}t^p}$ | $(1+\theta e^{x\beta}t^p)^{-\frac{1}{\theta}}$ | $e^{x\beta}pt^{p-1}$ |
| Weibull AFT (G) | $e^{-e^{-px\beta}t^p}$ | $(1+\theta e^{-px\beta}t^p)^{-\frac{1}{\theta}}$ | $e^{-px\beta}pt^{p-1}$ |
| Gompertz PH (G) | $e^{\frac{e^{-x\beta}}{\gamma}(1-e^{\gamma t})}$ | $(1-\theta\frac{e^{x\beta}}{\gamma}(1-e^{\gamma t}))^{-\frac{1}{\theta}}$ | $e^{x\beta+\gamma t}$ |
| Gompertz AFT (G) | $e^{\frac{e^{-x\beta}}{\gamma}(1-e^{\gamma t})}$ | $(1-\theta\frac{e^{-x\beta}}{\gamma}(1-e^{\gamma t}))^{-\frac{1}{\theta}}$ | $e^{-x\beta+\gamma t}$ |
| Cox PH (G) | $e^{-e^{x\beta}\Lambda_0(t)}$ | $(1+\theta e^{x\beta}\Lambda_0(t))^{-\frac{1}{\theta}}$ | $e^{x\beta}\lambda_0(t)$ |
| Weibull PH (IG) | $e^{-e^{x\beta}t^p}$ | $e^{\frac{1}{\theta}[1-(1+2\theta e^{x\beta}t^p)^{\frac{1}{2}}]}$ | $e^{x\beta}pt^{p-1}$ |
| Weibull AFT (IG) | $e^{-e^{-px\beta}t^p}$ | $e^{\frac{1}{\theta}[1-(1+2\theta e^{-px\beta}t^p)^{\frac{1}{2}}]}$ | $e^{-px\beta}pt^{p-1}$ |
| Gompertz PH (IG) | $e^{\frac{e^{x\beta}}{\gamma}(1-e^{\gamma t})}$ | $e^{\frac{1}{\theta}[1-(1-2\theta\frac{e^{x\beta}}{\gamma}(1-e^{\gamma t}))^{\frac{1}{2}}]}$ | $e^{x\beta+\gamma t}$ |
| Gompertz AFT (IG) | $e^{\frac{e^{-x\beta}}{\gamma}(1-e^{\gamma t})}$ | $e^{\frac{1}{\theta}[1-(1-2\theta\frac{e^{-x\beta}}{\gamma}(1-e^{\gamma t}))^{\frac{1}{2}}]}$ | $e^{-x\beta+\gamma t}$ |

Note: (G) denotes a Gamma distribution, $g(v)=\frac{v^{\frac{1}{\theta}-1}e^{\frac{-v}{\theta}}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}}$, with a mean of unity. The population hazard for this case is defined as $h_\theta(t)=h(t)(1-\theta ln(S(t)))^{-1}$. (IG) denotes an Inverse Gaussian distribution, $g(v)=\left[\frac{1}{2\pi\theta v^3}\right]^{\frac{1}{2}}e^{-\frac{1}{2\theta}(v-2+\frac{1}{v})}$, with a mean of unity. The population hazard for this case is defined as $h_\theta(t)=h(t)(1-2\theta ln(S(t)))^{-\frac{1}{2}}$. The unconditional density function in both cases is $f_\theta(t)=h_\theta(t)S_\theta(t)$.

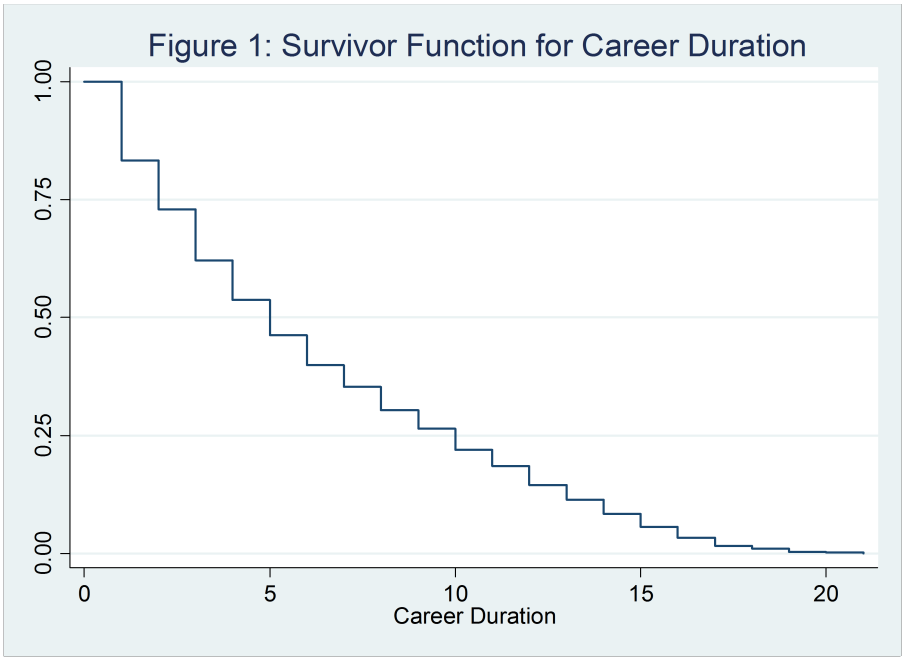
Figure 1: Survivor Function for Career Duration
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
Career Duration

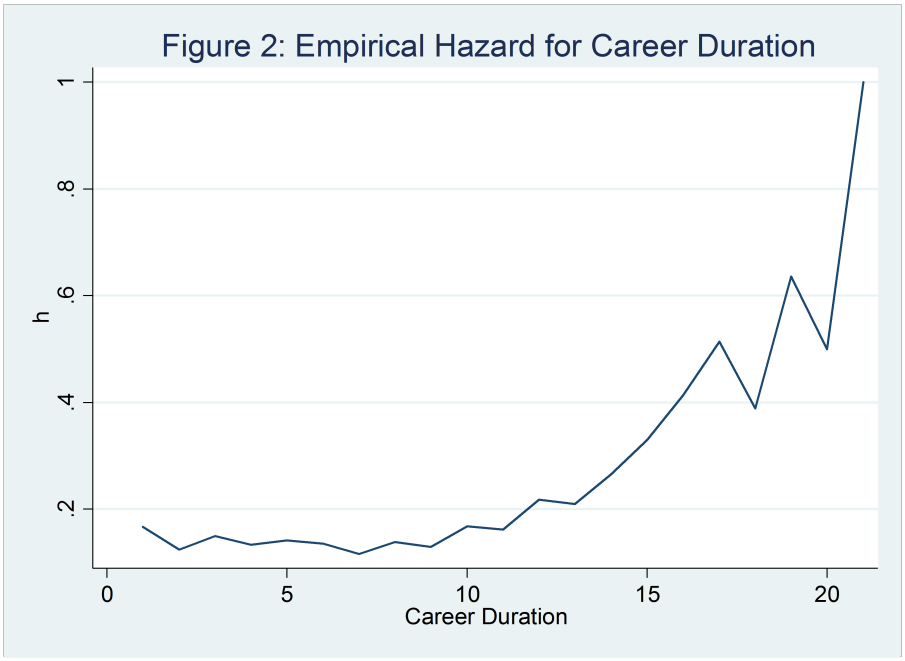
Figure 2: Empirical Hazard for Career Duration
1
.8
.6
.4
.2
h
0
5
10
15
20
Career Duration

Figure 1:

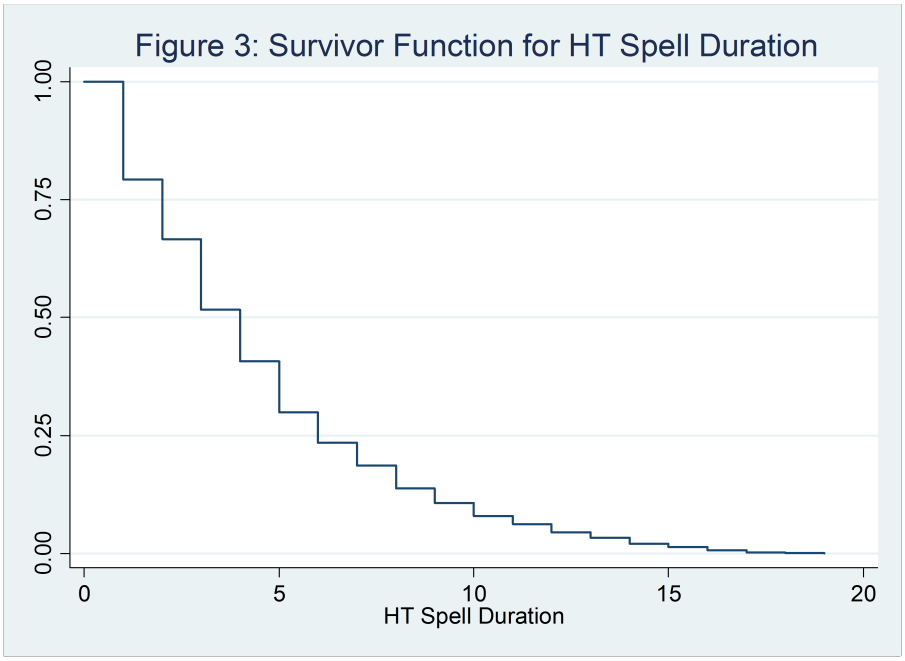
Figure 3: Survivor Function for HT Spell Duration
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
HT Spell Duration

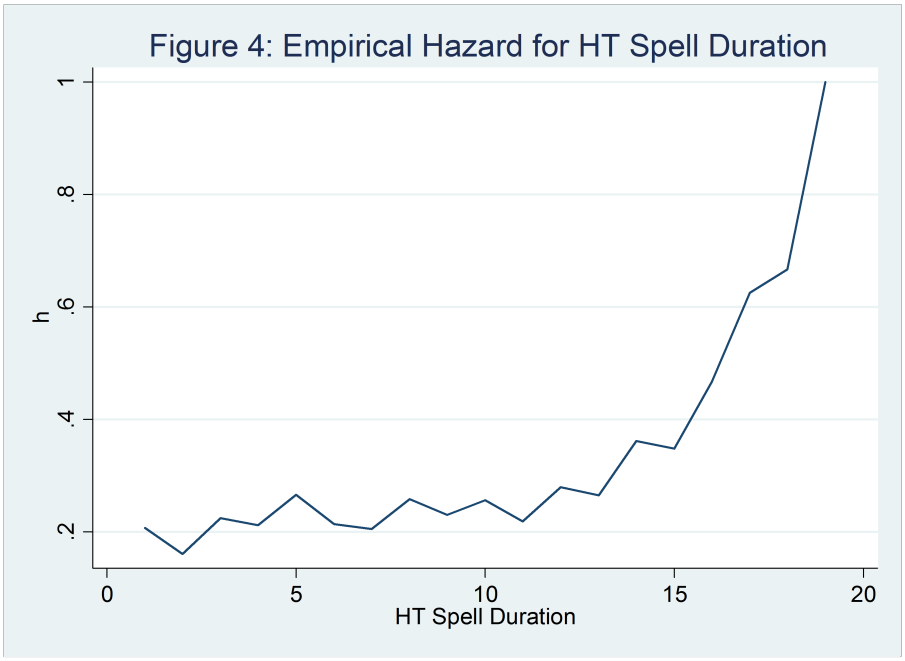
Figure 4: Empirical Hazard for HT Spell Duration
1
.8
.6
.4
.2
h
0
5
10
15
20
HT Spell Duration

Figure 2:

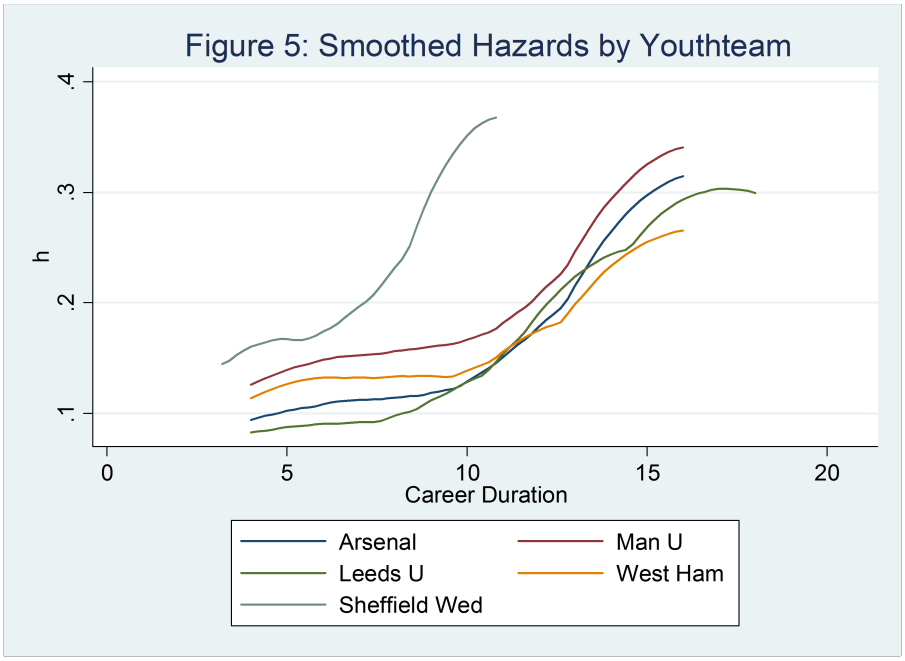
Figure 5: Smoothed Hazards by Youthteam
h
.1
.2
.3
.4
0
5
10
15
20
Career Duration
Arsenal
Man U
Leeds U
West Ham
Sheffield Wed

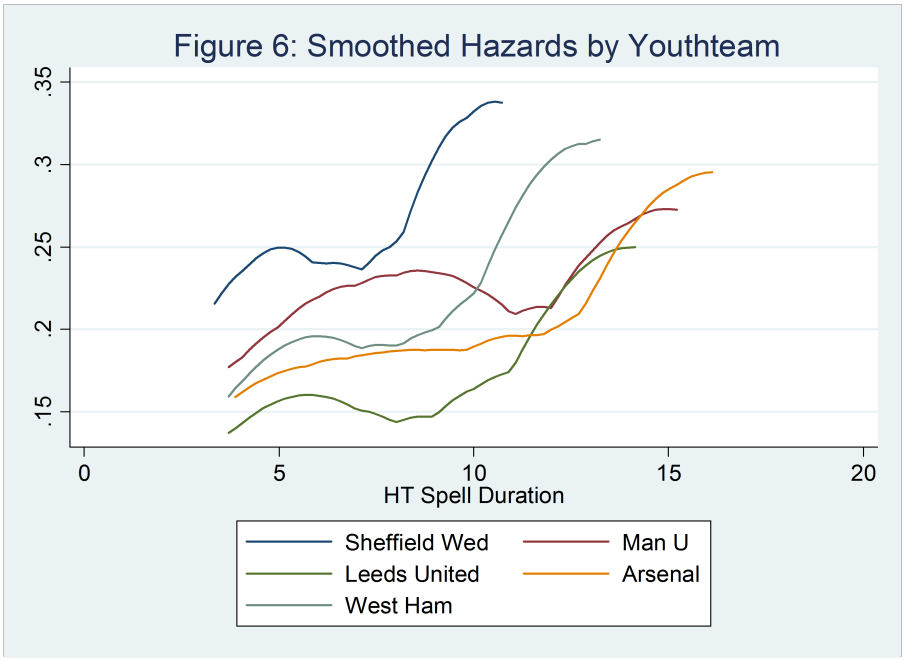
Figure 6: Smoothed Hazards by Youthteam
.15
.2
.25
.3
.35
0
5
10
15
20
HT Spell Duration
Sheffield Wed
Man U
Leeds United
Arsenal
West Ham

Figure 3:

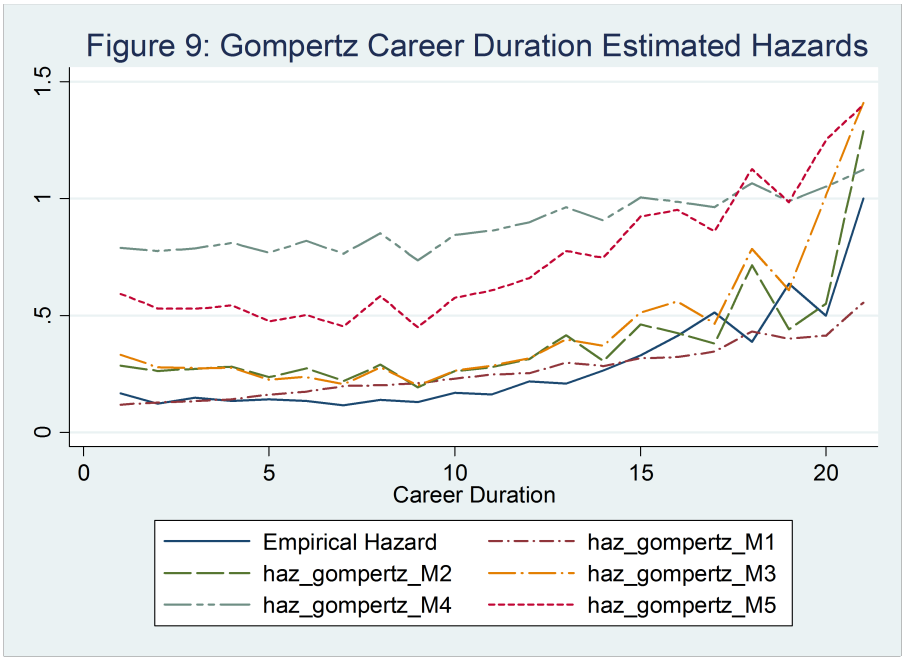
Figure 9: Gompertz Career Duration Estimated Hazards
0
.5
1
1.5
0
5
10
15
20
Career Duration
Empirical Hazard
haz_gompertz_M1
haz_gompertz_M2
haz_gompertz_M3
haz_gompertz_M4
haz_gompertz_M5

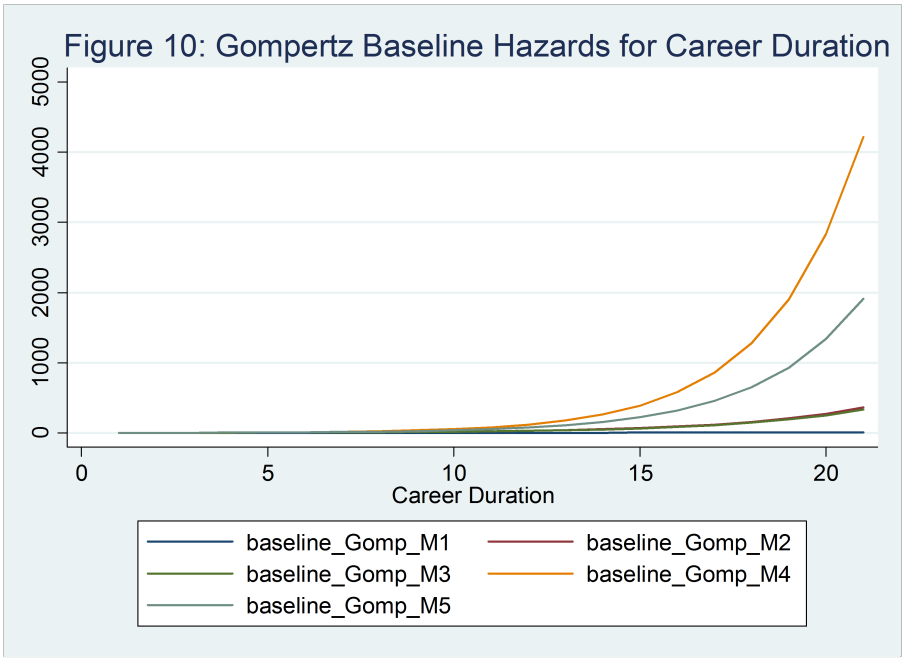
Figure 10: Gompertz Baseline Hazards for Career Duration
0
1000
2000
3000
4000
5000
0
5
10
15
20
Career Duration
baseline_Gomp_M1
baseline_Gomp_M2
baseline_Gomp_M3
baseline_Gomp_M4
baseline_Gomp_M5

Figure 4:

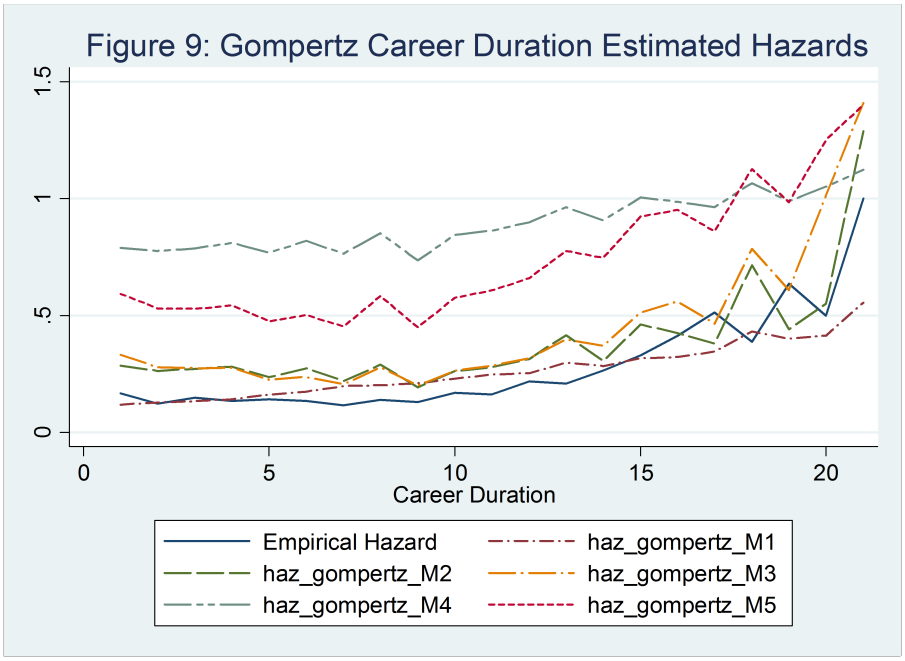
Figure 9: Gompertz Career Duration Estimated Hazards
0
.5
1
1.5
0
5
10
15
20
Career Duration
Empirical Hazard
haz_gompertz_M1
haz_gompertz_M2
haz_gompertz_M3
haz_gompertz_M4
haz_gompertz_M5

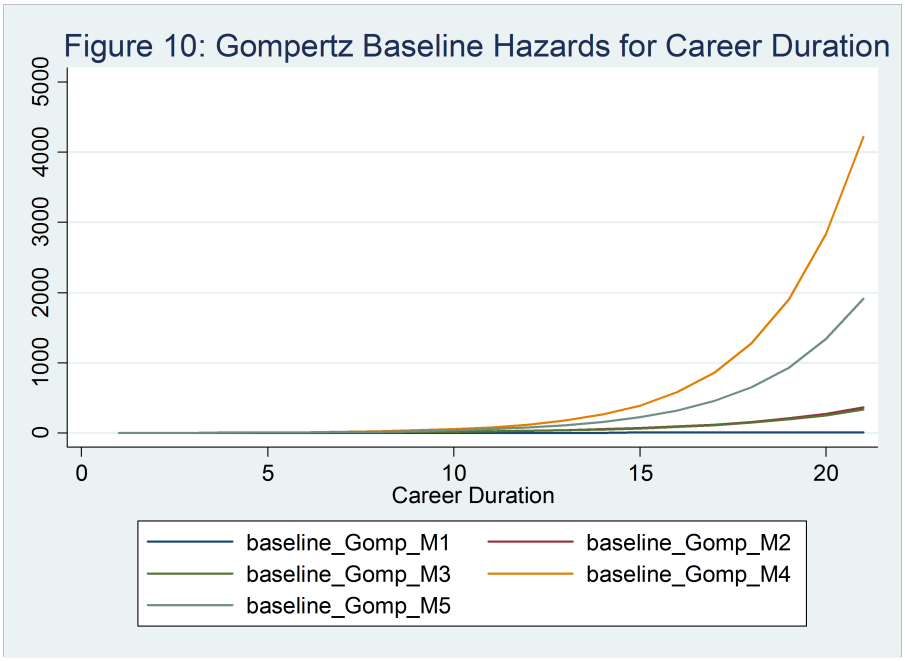
Figure 10: Gompertz Baseline Hazards for Career Duration
0
1000
2000
3000
4000
5000
0
5
10
15
20
Career Duration
baseline_Gomp_M1
baseline_Gomp_M2
baseline_Gomp_M3
baseline_Gomp_M4
baseline_Gomp_M5

Figure 5:

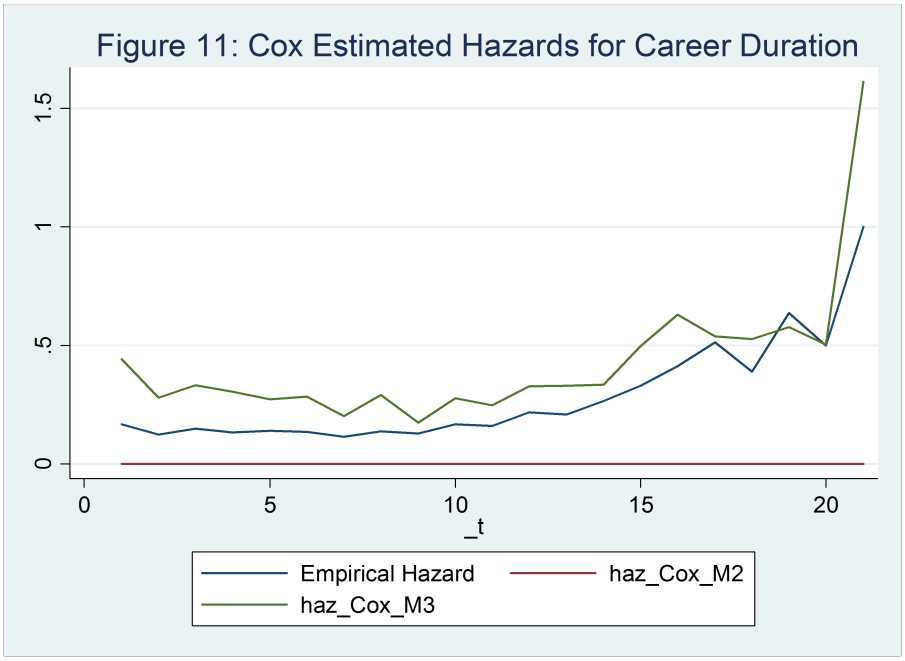
Figure 11: Cox Estimated Hazards for Career Duration
0
.5
1
1.5
0
5
10
15
20
_t
Empirical Hazard
haz_Cox_M2
haz_Cox_M3

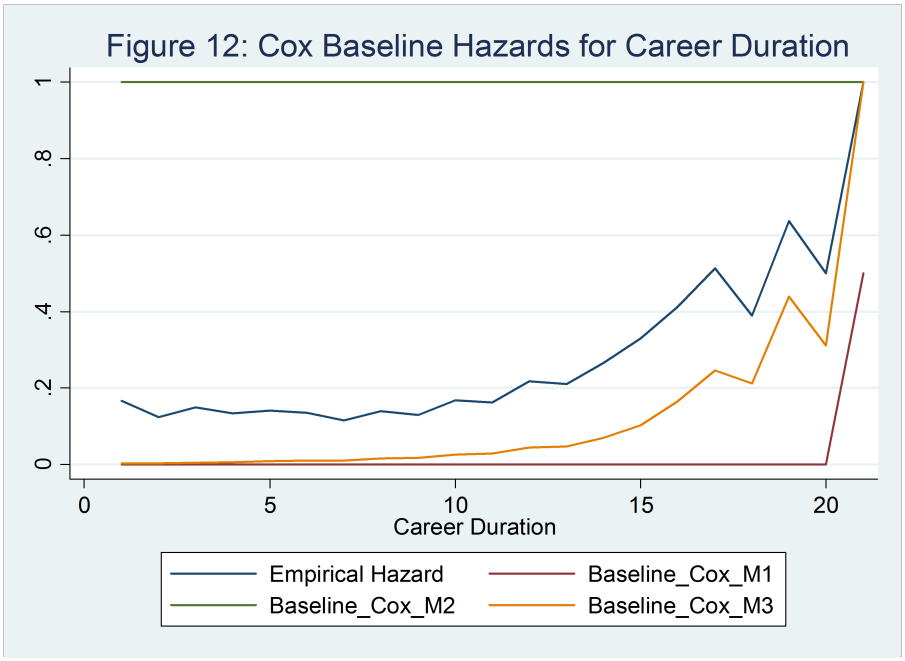
Figure 12: Cox Baseline Hazards for Career Duration
0
.2
.4
.6
.8
1
0
5
10
15
20
Career Duration
Empirical Hazard
Baseline_Cox_M1
Baseline_Cox_M2
Baseline_Cox_M3

Figure 6:

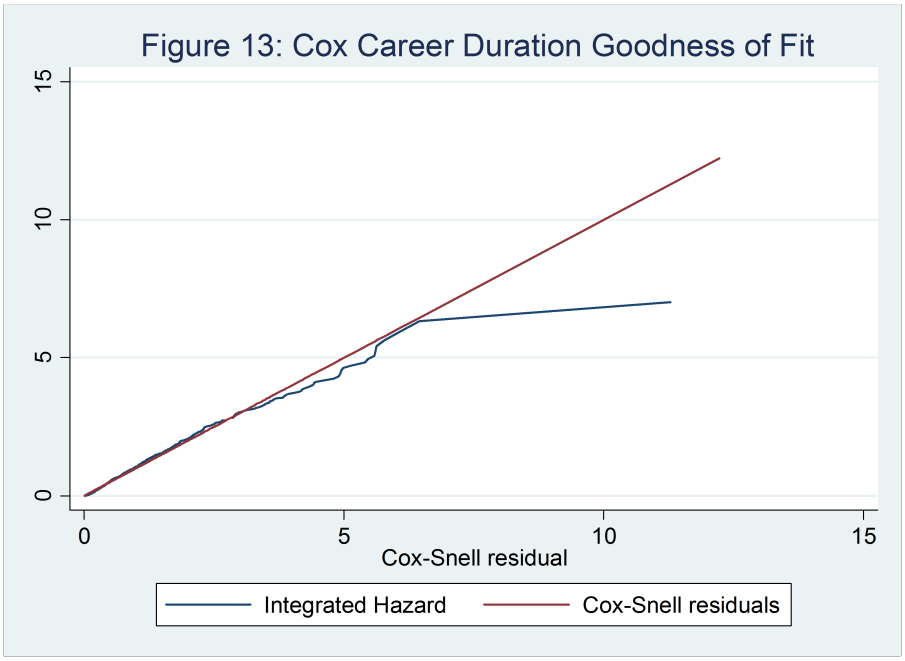
Figure 13: Cox Career Duration Goodness of Fit
0
5
10
15
Cox-Snell residual
Integrated Hazard
Cox-Snell residuals

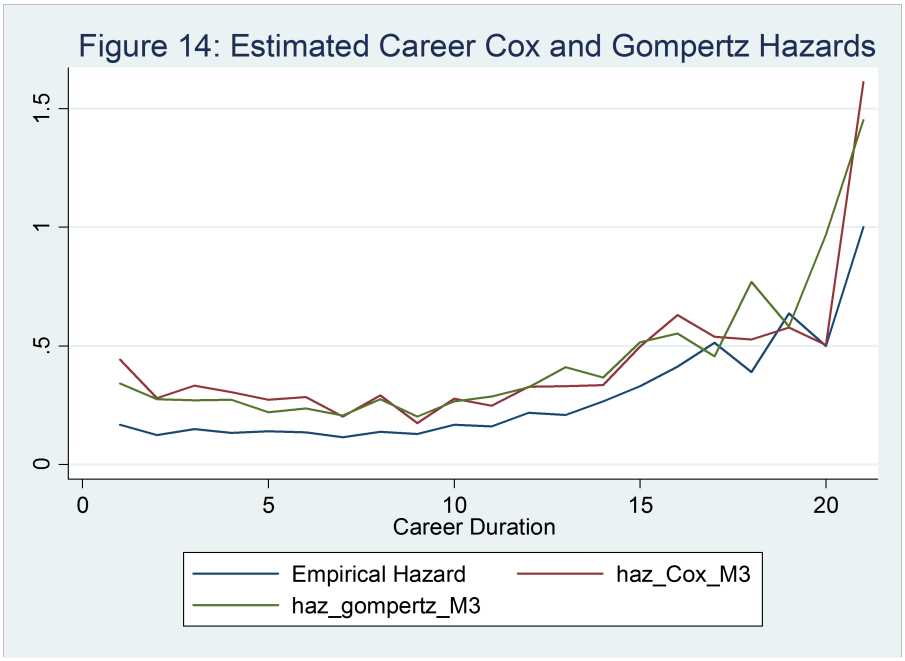
Figure 14: Estimated Career Cox and Gompertz Hazards
0
.5
1
1.5
0
5
10
15
20
Career Duration
Empirical Hazard
haz_Cox_M3
haz_gompertz_M3

Figure 7:

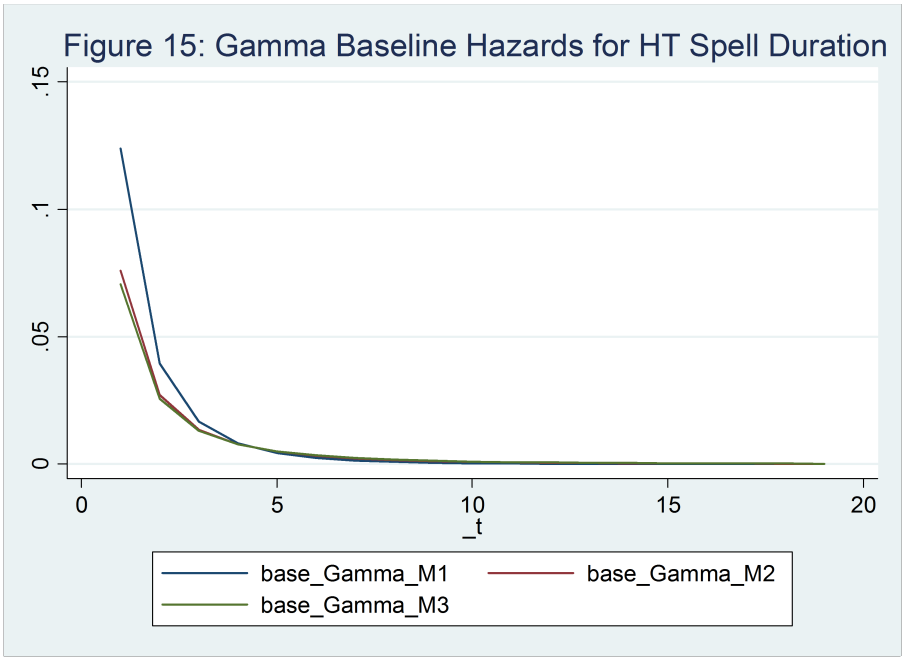
Figure 15: Gamma Baseline Hazards for HT Spell Duration
0
.05
.1
.15
0
5
10
15
20
_t
base_Gamma_M1
base_Gamma_M2
base_Gamma_M3

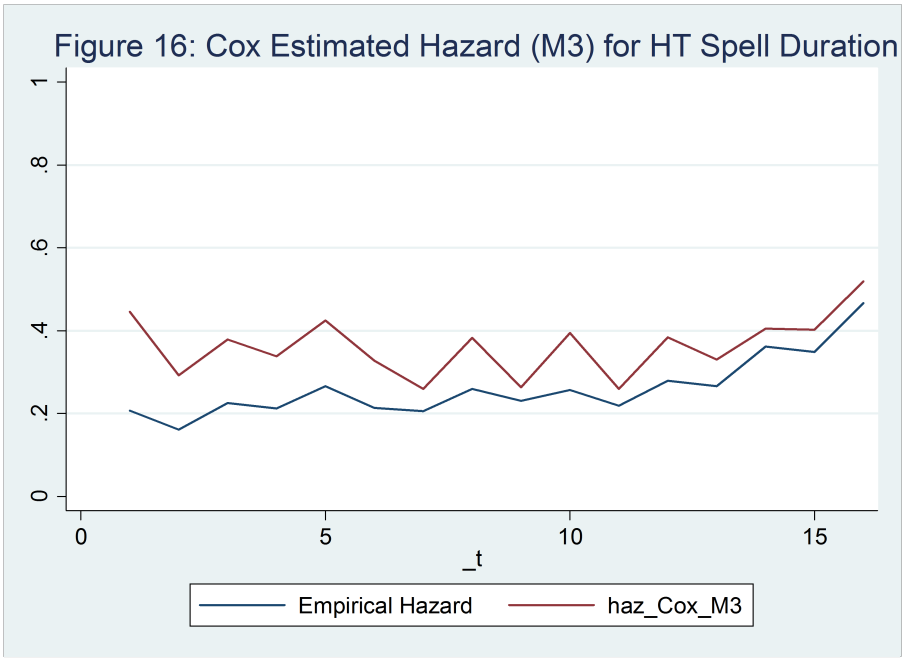
Figure 16: Cox Estimated Hazard (M3) for HT Spell Duration
0
.2
.4
.6
.8
1
0
5
10
15
_t
Empirical Hazard
haz_Cox_M3

Figure 8:

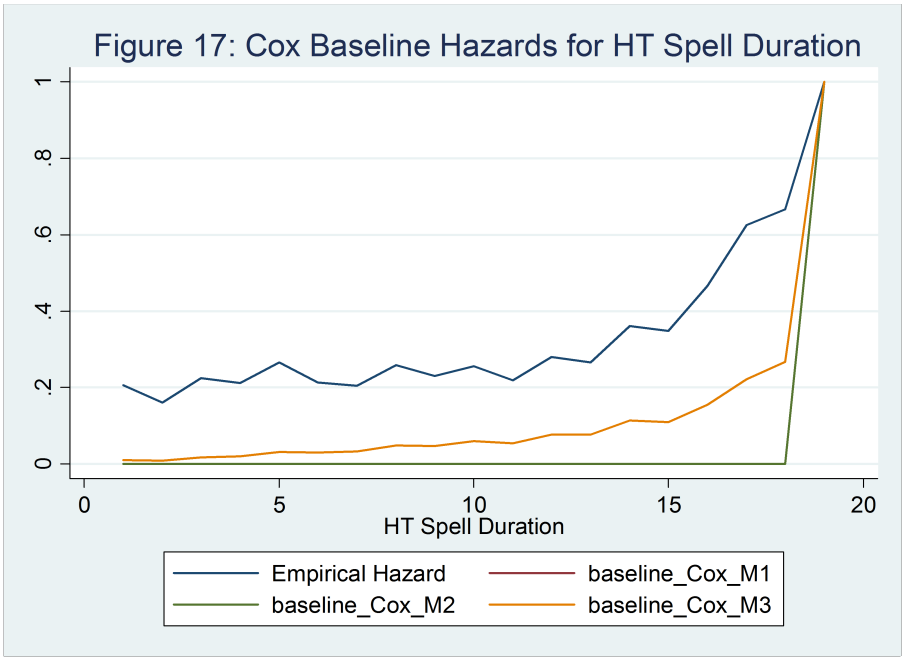
Figure 17: Cox Baseline Hazards for HT Spell Duration
0
.2
.4
.6
.8
1
0
5
10
15
20
HT Spell Duration
Empirical Hazard
baseline_Cox_M1
baseline_Cox_M2
baseline_Cox_M3

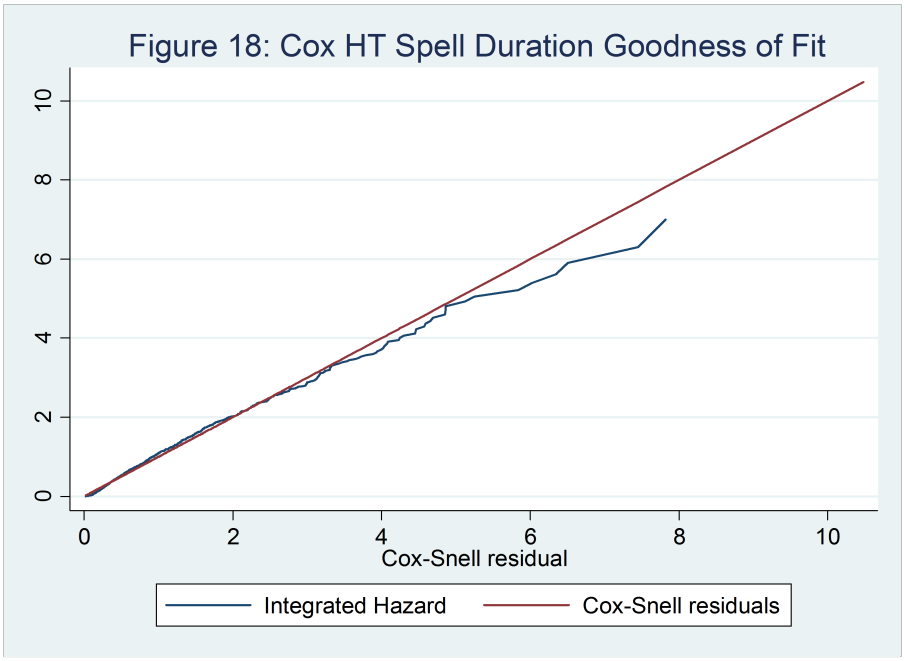
Figure 18: Cox HT Spell Duration Goodness of Fit
0
2
4
6
8
10
0
2
4
6
8
10
Cox-Snell residual
Integrated Hazard
Cox-Snell residuals

Figure 9: