

OPTIMISING PLAYER VALUATIONS IN THE INDIAN PREMIER LEAGUE

Ryan Stephen
29255199



Abstract

The Indian Premier League (IPL) is the highest-profile domestic cricket tournament in the sport of cricket. In fact, the IPL has even surpassed major international tournaments in terms of both viewership and gross earnings. As a result, player salaries have reached unprecedented levels, and are now at a point where they have completely dwarfed any earnings that a cricketer can earn in any other format of the game. This tournament has therefore made Twenty-twenty cricket the primary focus of a young Indian cricket player's dreams.

Unlike the English Premier League in Football, the IPL holds official auctions in which the competing franchises bid against each other for the diverse array of international players that have made themselves available for this tournament. This introduces a new dynamic into the prices that are offered for these sporting heroes, where sentiment, public perceptions and the heat of the moment could cause a franchise to bid sums of money in excess of \$ 2 million for a single player.

This project aims to analyse the prices and performances of different cricketers in order to quantify the factors that affect the final bidding prices, and how these factors contrast with those that are most likely to project a successful season ahead. If the findings of this assignment prove accurate, a similar approach can then be used by the various IPL franchises to save millions of dollars on misspent player salaries.



Table of Contents

Section	Page
1. Background.....	3
2. Commercial Basis for the Research.....	4
3. Factors that Affect Player Prices.....	6
4. Data Handling.....	9
5. Initial Model Design.....	17
6. Principal Component Analysis.....	20
7. Survival and Strike rate Indices.....	25
8. The Performance Price Index.....	36
9. Team Dynamics.....	40
10. Modelling the distribution of scores.....	44
Glossary.....	49
References.....	51

1 Background

The IPL has now progressed to its 6th year since its inception in January 2008, and the commercial success of this tournament has seen it grow into an internationally contested competition that now dominates the goals and aspirations of almost all young Indian cricketers. In a recent interview with former Indian batsman Rahul Dravid, it was stated that:

“We shouldn't be taken in by the politically correct noises players make about Test cricket being the ultimate challenge. The truth is, every young player wants to somehow land an IPL contract. That is what he is playing for.”¹

It is now all about the money for young Indian cricket players, and although the traditional cricket lovers will always favour the five-day version of the game, the IPL has become the commercial way of the future, even luring in those players who are not inclined to the shorter formats.

The IPL brand value was estimated to be around 2.99 Billion US Dollars.

For better or for worse, the IPL is probably renowned more for its commercial success than for the cricket played. If anything, some may feel that the less said about the cricket the better and that it should rather be viewed as a carnival of sorts. The opinion of the purists however, has never much bothered the wealthy Bollywood stars and business tycoons that own these franchises, perhaps it wouldn't bother anyone when the turnover figures begin to exceed tens of millions in US dollars.

What is more likely to bother them though, are the rising sums of money that are being spent in the player auctions. As several marquee players start to fetch prices in excess of \$2 million each, player costs are beginning to swallow up profits, especially when a single franchise owns up to 22 players at a time.

Rising player costs are introducing a new challenges for the various franchises of the IPL, whose aim has now increasingly tended towards that of bidding to reduce costs rather than to try and win the big names. In other words, buying reliable performers at a bargain is becoming the new fashion in the selection strategy of IPL franchises.² The dynamic is well captured by cricinfo columnist Amrit Mathur :

“In the early days of the IPL, franchises paid serious money to hire top players in order to attract attention towards their brands and engage with fans. Later, faced with financial losses, and alive to the danger of rising player cost that could damage the balance sheet and permanently cripple their business, teams got smarter in picking players. They are still a distance away from the Moneyball levels, but a clear understanding now exists that the player salary expense, unless capped, can wreak havoc.”¹

It is therefore imperative for IPL franchises that they find ways of cutting the costs of player salaries, and mark out definitive points at which they have to cap their expenses and stop bidding.³

2 Commercial Basis for the Research

Factors that determine bidding prices can be found by statistically modelling the various characteristics/numerical attributes of particular players against the final auction prices assigned to these players during past IPL seasons. It can then be determined how much weight or emphasis has been given to certain prediction factors such as a player's batting average. For example, it may be possible that the final bidding price of a player at an IPL auction can be modelled using a statistical regression model² similar to the following (very basic) example:

$$\text{Final Bidding Price} = A * (\text{T20 batting average}) + B * (\text{T20 strike rate}) + C * (\text{Matches played})^2$$

This regression model can be extended to more complex general linear models in order to predict as accurately as possible what the various weightings on certain factors should be to further minimise the R-squared value that emerges in our model of the price of a particular player.⁵

These results can then be used by an IPL franchise to forecast future prices of certain players before the auction begins, and thus determine in advance which players are becoming over-priced or under-priced by recent standards due to the exaggeration of their qualities in the public opinion. This may come as a result of various factors such as hype, irrational excitement, effects of emotions and ego during a competition with other bidders etc. All of which can occur during an auction.

The weightings of certain factors that influence prices can also be investigated for bias or inaccuracies by statistically modelling these same factors against performance statistics in order to identify which factors are most likely to indicate an increase in performance and which are most indicative of a price increase. The similarities and deficits of these results can be compared and contrasted in order to assess what kind of player would provide better value for money.

In other words, the investigation will attempt to find factors that have been incorrectly weighted in determining a player's final auction price compared to how it determines a player's future performance in the coming season. By using general linear and regression models to both find and prove these differences, a client franchise may gain a significant advantage over other bidders in assessing the true value of particular players.

In addition, similar models based on various performance indicators can be used to assess how well a player's prospective performances will fit within a specific team's needs. This is a key indicator in determining whether it is worth competing with other franchises in bidding for a particular player who may become too expensive relative to his ability to satisfy the specific team requirements of a client franchise. This may often occur as a result of the player's abilities to provide for the specific needs of one of the client's competitors.

Therefore, there are three main reasons why a player may be considered over-priced/under-priced due to the new information that may be obtained from these models:

1. It will be determined whether he fits well with another team's needs, but not necessarily with the client's team needs, thus causing other franchises to bid prices that are beyond the value of this particular player to the client. (Refer to Section 9: "Team Dynamics" for a full explanation on how this is done.)

2. If it is discovered through the pricing model that certain factors that determine auction prices are weighted too heavily or lightly in favour or against a player, the performance prediction models can be used to indicate the extent by which this factor is less or more important than the priorities that have been placed on them in past IPL auctions.
3. Bidders tend to get carried away in the heat of the moment in an auction, and could possibly make irrational decisions. It would be better to employ scientific methods based on statistics in order to discern a player's true value. This may provide the client franchise with the perspective that it needs in order to facilitate heat-of-the moment decision making as well as providing guidance in terms of when prices have started to display bubble characteristics through too much hype and competitive bidding.²

Thus if a client franchise were to use the models that will be investigated in this project, there will be three main areas where they may gain a competitive advantage over the other franchises during the bidding process.

One key factor that the models may not be able to ascertain is whether too much or too little emphasis is based on the iconic status of a player in determining a bidding price.⁴ In this matter, the client franchise will have to assess their own statistics related to advertising and commercial revenues in order to make a quantified assessment of the monetary value of these factors, which may form a possible extension to this project.^{2,4}

Note:

The analysis of players in the following assignment will focus only on Batsmen from section seven onwards. Most of the techniques applied to batsmen can be mirrored in their application to analyse bowlers. So for the sake of brevity, only batsmen have been analysed in most of the assignment that follows, with the direct implication that the same techniques can be applied to bowlers.

Words such as "T20", "ODI" and "Test" will be used as abbreviations throughout this assignment. For explanations of all terms used, refer to the glossary on page 49.

3 Factors that Affect Player Prices

Depken and Rajasekhar (2010) point out that “Although the IPL imposes a salary cap and other labour-market restrictions, it is anticipated that final bid prices reflect the aggregate value of player productivity statistics, potential leadership skills, and auction characteristics.”²

Therefore any model that seeks to predict player prices will have to contain covariate factors beyond that of merely explaining the performance records of a player, but also leadership, as well as the player’s ability to attract fans and TV audiences.⁴

The effect that auction characteristics should have on player prices is less simple to quantify. Parker and Burns (2010) do however agree that the auction rules seem to be designed for the purpose of minimising common value uncertainty.³ Common value uncertainty arises when the value of a player is approximately equal for all bidders, but the value itself is unknown. Bidders are therefore left to make estimates of the uncertain value with the winner being the most optimistic in their calculations (or even guesses).³

The problem with this situation in the perspective of the auctioneers is that bidders will begin to spot when this effect is taking place and will reduce their bids simultaneously, leading to lower player valuations.³

Depken and Rajasekhar (2010) also tested whether franchises had to pay different wages to similarly qualified players. In other words, is the fame and fortune of a particular franchise having an effect on bidding prices? Fortunately, however, their results were inconclusive and this definitely simplifies matters when modelling player prices on performance factors.

Another major factor that may have a significant effect on IPL prices are the IPL auction rules. These include a “rather modest salary cap which reduces the ability to pay exorbitant prices for players beyond those given the icon status.”² Those players who were given the icon status have to be paid 15% more than the next highest player in the team. Therefore factors indicating a player’s iconic status will have to be built into the model to accommodate this distortion in player prices. This does however, also have a balancing effect on prices that makes them independent of which franchise is bidding for the player, as Depken (2010) points out: “the IPL seems to have created a system in which it is difficult for one franchise to systematically spend more than other franchises on otherwise equally qualified players.”²

They even tested this conjecture by creating seven dummy variables for the eight franchises in a pricing model, (With RCB as the reference franchise) and indeed, their findings further support this view with empirical evidence.

See:

Depken, Rajasekhar (2010) “Open market Valuation of Player Performance in Cricket: Evidence from the Indian Premier League”, Social Science Research Network, page 11

With regards to the unknown extent by which the media affects a player’s value, Van den Berg (2011) seems to indicate that sports economists are yet to come to any common agreement on a single approach to quantifying these effects on player prices.⁴

He states that “a consensus still does not exist on whether financial success is a mere by-product of sporting success”⁴

For the purposes of maximising a player's value for money, the performance predicting attributes will be emphasised, and those unquantifiable factors that skew prices due to legendary status and other such biases will be captured in this assignment's model as the deficit between the price and the projected performance value.

Van den Berg (2011) further supports the proposed pricing model for this project by providing some thorough backing to the argument that performance indicators do indeed have the greatest effect on a player's value:

"It is extremely difficult to model and establish which of the two comes first. However, it is beyond doubt that a clear link exists."

Figure 4 (below) indicates that the high income teams are usually assured top spot in the league. Illustrating the strong correlation that exists between player price and performance in the football player market. ⁴

In referring to figures 5 and 6, Van den Berg (2011) states that these graphs "show that it is more logical to focus on personnel expenditure", and that "the strong correlation between personnel costs and on-pitch success suggests that titles can indeed be 'bought'" ⁴

It can surely be derived from the analysis of a much more mature player market (such as the European national football leagues) that the dynamics of player pricing in the IPL will soon follow the same trends, for which higher player prices correlate with winning matches. Hence the focus on performance predicting factors is a more important goal in player valuation, and it will ultimately draw the fans in a more sustainable way when compared to simply purchasing marquee players for the sake of short term fanfare without the long term substance.

The graphs below refer to the economic dynamics currently taking place in the more mature football player market. It is assumed that the IPL will follow a similar pattern in the years to come.

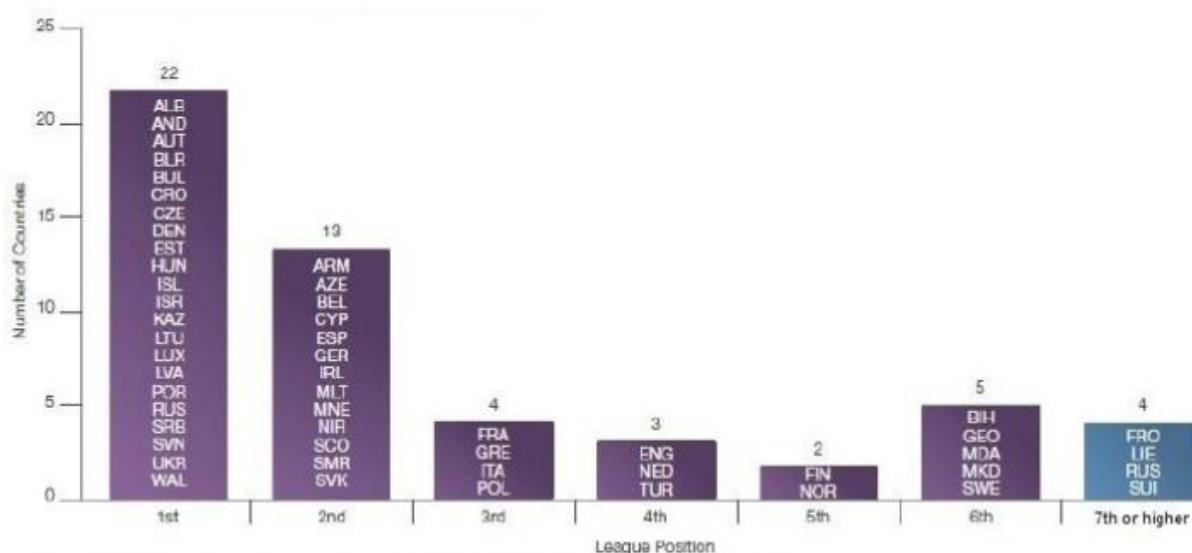


Figure 4: Link between sporting and financial success across Europe

This figure depicts the dominance of the teams that generate the most revenue in a league, by indicating their league finishes for the season 2008/2009. Across 53 European leagues, these teams finished in the top two positions an overwhelming 35 times (source: UEFA, 2011).

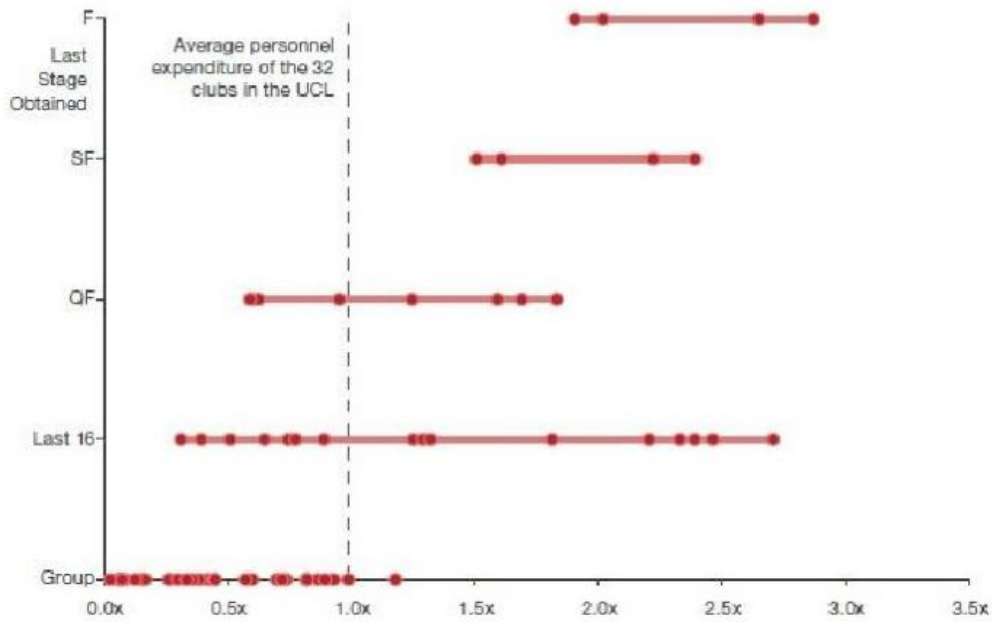


Figure 5: Relative personnel expenditure and Champions League success

This graph indicates the range of personnel expenditures (which are mainly expenditures on player salaries) by UEFA Champions League participants depicted against the stage at which they got eliminated from this prestigious competition. The scale of personnel expenditures appears to be constantly increasing, as clubs that spend more on player salaries advance further in the competition than less generous clubs. Source: UEFA

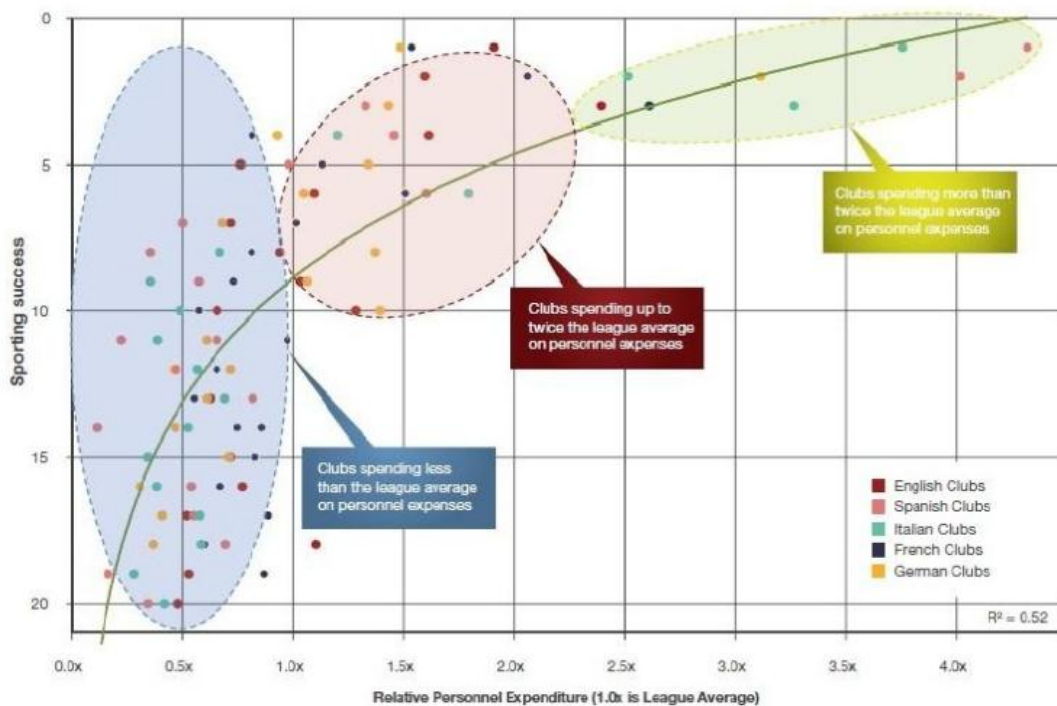


Figure 6: Relative personnel expenditure and league success

This graph depicts relative personnel expenditures (relative to league average) against league finish (x-axis) for the five major competitions (color markings are used to indicate the origin of clubs) for 2008/2009. The graph indicates a positive relation between relative spending and league finish, although interestingly the marginal benefit of allowing for a larger payroll appears to be decreasing, as clubs reach higher league positions (source: UEFA, 2011).

Source:

E. van den Berg (2011), "The Valuation of Human Capital in the Football Player Transfer Market", Erasmus School of Economics, pp 10 -16

4 Data Handling

4.1 Compiling the Database:

The purposes of the assignment, for which the database needs to support, would be to determine the factors which would both forecast and correlate with future player performances, as well as making comparisons with auction price correlations and forecasts of these prices. A consistent and comprehensive database would be required, and would have to be compiled in such a way that it combines all the facts, figures and characteristics of a cricketer in a single table of reference. This is the minimum requirement if all the data is going to be used simultaneously for linear regression modelling.

You could, of course, ensure that all of a player's statistics would be included adjacent to his name by searching out a complete player profile containing all the records for each individual cricketer, but this would have to be done one player at a time, which would take a considerable amount of time and effort, hence more efficient methods were sought out.

The immediate challenge would then be to gather separate statistics from various sources and list all such records in a single database entity adjacent to the name of the cricket player to whom these records belong. This may sound like a relatively simple task as it is often assumed that all such statistics can be obtained in a ready-made format for which records are consistent in both primary key⁶ (see appendix) and formatting. However, in reality, each list of publicly available cricketing records contains a list of player names alongside only a partial set of their individual statistics. For example, a list of batsmen with "most runs in test cricket" will provide only a subset of certain players, many of whom will not be relevant to an IPL database. It would also contain only their Test batting records, whereas other records and statistics for the same players will have to be sourced from other lists of cricketing records, such as "Most ODI centuries", which would hopefully provide the ODI statistics for the same player, but may not contain the same set of players for which the Test batting records were obtained. Therefore, several different player lists had to be combined and matched with the same primary key⁶, which would need to be unique as well as having a consistent format of name and initials for each cricketer.⁶ The challenges emerging from this task would be the inconsistent formats of the different sources of player statistics which were supposed to be merged. For example, the sources of Jacques Kallis' ODI batting record listed the primary key as "J.H Kallis", whereas the sources for his IPL bowling record listed his records under "Jacques Kallis".⁶

This made a mass merger of different statistics under a single, consistent primary key rather difficult. Other formatting inconsistencies included mismatching orders of player attributes, such a particular data source listing "wickets", "balls bowled", "wicket rate", and "five-fors" in a different order to the current data source which it has to be merged with. Not all cricketer's bowling statistics can be obtained from a single list of bowling records, as some lists are designated for bowlers with the "most wickets" and others are designated for "best economy rate" etc. Therefore, different bowling records for different players must be compiled and combined together from different lists of bowling statistics, which of course introduces the risk of listing the records of different players in an inconsistent format. This would make it impossible to perform queries and regression modelling on all the data simultaneously, which is one of the key objectives of the assignment. Therefore, it is imperative to ensure consistency.⁶

Sometimes the only solution to the above mentioned inconsistencies has been to simply access the complete universe of an individual cricketer's records one player at a time using the Cricinfo's "statsguru" query engine (address listed below). However, it has already taken a considerable amount of time to compile data from a relatively small sample of players in this manner, and it also requires the sifting of IPL and overall T20 records due to the fact that separate IPL records per individual are not available in this search engine.

For this assignment, not every player in the IPL has been included in the database, although the complete structure of the database is in place and capable of being extended to include all players. Therefore the models and the underlying database presented in this assignment may function as a template for the complete modelling and comparison of all players in terms of price and performance factors, whereas the analysis in this assignment will only be performed on a select group of IPL cricketers, if only as an example.

Appendix:

- The primary key of a database is a unique field in a table for which each row of a dataset can be uniquely identified without any probability of mismatching other rows of explanatory fields with the same identifier. This identifier is therefore a single cell in the dataset that does not have any repetitions across different rows. For the sake of referential integrity, rows from different tables of data that refer to the same entity must be linked to one another through a foreign key, which is a unique primary key of the other table which contains the data that refers to the same entity.⁶
- For this application, the table or dataset containing Jacques Kallis' ODI bowling figures would have to be linked to the table containing his T20I batting figures with mutually consistent fields that have equivalent values, and which function as a unique identifier for each individual player.

In other words, both tables denote him as Jacques Kallis in the primary key, rather than one table has a primary key of JH Kallis, another has J.H. Kallis etc.

4.2 Data Sources:

Different data sources were required, which had different strengths and short-comings as described below:

Cricinfo's general statistics page:

Strengths:

Simple, user friendly platform from which to search for general records of all players.

Shortcomings:

Lists of cricketers' records could only be accessed in the form of "most centuries" or "most wickets". This meant that a balanced list of players with both batting and bowling factors included adjacent to one another could only be compiled by means of accessing the complete individual record of one player at a time.

Lists of IPL results were also only available in the form of "most runs", "most wickets" etc.

Cricinfo's statsguru engine:

Strengths:

Most of the data for this assignment was obtained from the statsguru engine. It provides a comprehensive array of statistics for every player that has played the game. Along with every database filtering feature that any statistician could wish for.

Shortcomings:

Accessing a complete set of records for a set of individual players is not possible, the extensive filters will only provide certain statistics for a different set of players for each query. To access the complete records of a player, one would have to search for the records one player at a time.

Lists of IPL results were also only available in the form of "most runs", "most wickets" etc. This poses the difficulty that not all players in the required database are listed under these tables. Statsguru failed to provide lists of cricketers' IPL statistics per individual. Hence, a cricketer's Test, ODI, T20I and overall T20 overall records could be obtained for one player at a time, but the IPL statistics per individual players had to be sourced from the official IPL website. (Address listed below)

Official IPL website:

Strengths:




















Provided the complete as well as seasonal IPL records for each individual player.

Shortcomings:

Did not provide data for other formats.

Player attributes were listed in an order that is inconsistent with the statistics that were obtained from Cricinfo.

In order to get the largest (and hence most reliable) dataset of players, the analysis will only be focused on those players that have played the most IPL matches up until the end of the 2013 season. Unfortunately, the Official IPL website does not provide a list of players by the category of 'most matches played', so a little improvisation is necessary in that data can be collected from a category such as 'most fifties' or any similar category that would be likely to provide a list of the batsmen that have played the most matches up until the current 2013 season. The Data can then be collected and sorted on Excel in the order of matches played.

Most Fifties													
Pos	Player	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s
1	 Michael Hussey	17	17	3	733	95	52.35	566	129.50	0	6	81	17
2	 Virat Kohli	16	16	2	634	99	45.28	457	138.73	0	6	64	22
3	 Chris Gayle	16	16	4	708	175*	59.00	453	156.29	1	4	57	51
4	 Suresh Raina	18	17	4	548	100*	42.15	365	150.13	1	4	50	18
5	 Rohit Sharma	19	19	5	538	79*	38.42	409	131.54	0	4	35	28
6	 Ajinkya Rahane	18	18	4	488	68*	34.85	458	106.55	0	4	42	11
7	 Rahul Dravid	18	17	1	471	65	29.43	425	110.82	0	4	64	5
8	 MS Dhoni	18	16	5	461	67*	41.90	283	162.89	0	4	32	25
9	 Aaron Finch	14	14	0	456	67	32.57	336	135.71	0	4	54	16
10	 Dwayne Smith	13	13	0	418	68	32.15	341	122.58	0	4	40	19
Pos	Player	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s
11	 David Warner	16	16	3	410	77	31.53	323	126.93	0	4	41	14
12	 Gautam Gambhir	16	16	0	406	60	25.37	343	118.36	0	4	51	5
13	 Kieron Pollard	18	18	8	420	66*	42.00	281	149.46	0	3	27	29
14	 David Miller	12	12	5	418	101*	59.71	254	164.56	1	3	28	24
15	 Shikhar Dhawan	10	10	2	311	73*	38.87	253	122.92	0	3	37	5
16	 Shaun Marsh	8	8	0	300	77	37.50	249	120.48	0	3	39	4
17	 Shane Watson	16	16	2	543	101	38.78	380	142.89	1	2	59	22
18	 Dinesh Karthik	19	19	1	510	86	28.33	411	124.08	0	2	54	14
19	 Robin Uthappa	16	16	0	434	75	27.12	371	116.98	0	2	42	12

Data lists such as “Most fifties” (above) can be sourced from the official IPL website

Web addresses of Data sources:

Cricinfo's general statistics page :

<http://www.espncricinfo.com/ci/content/current/stats/index.html>

The Statsguru search engine on Cricinfo :

<http://stats.espncricinfo.com/ci/engine/stats/index.html>

Official IPL website : www.IPLT20.com

The Cricmetric website : www.cricmetric.com

4.3 Data Modifications:

A couple of new fields were added to the database as a calculated combination of other fields for the purpose of providing independent factors that complete and summarise the picture of a player's abilities (see section 6). These new fields were used in the modelling of a performance index.

As an example:

The survival rate, balls faced and strike rate are the three central indicators used to evaluate batsmen. However, data for the batsmen often does not come in the desired format and a couple of additions were required, as illustrated below.

	A	B	C	D	E	F
1	Player	Survival	Runs	Ave	BF	SR
2	SK Raina	=E2/(C2/D2)	4405	36.1	4806	91.65
3	RG Sharma	46.1986061	3049	36.73	3835	79.5
4	CH Gayle	44.72641427	8743	37.68	10378	84.24

Survival rates were derived from “Runs”, “Average” and “Balls faced”.

4.4 Redundant Data:

Some sections of the players' records were excluded from the dataset as a result of certain components being considered irrelevant in determining their value to the team. For example, it can be assumed with sufficient confidence that Morne Morkel's batting achievements will have little or no effect on either his auction price or his performance index.

	Test results				ODI results			T20 results				
	TBalls	TEcon	TWR	T5w	ODIBa	ODIEc	ODIWR	T20Ba	T20Ru	T20Ec	T205w	T20Wf
Z_Khan	17612	3.25	59.7	10	0	0	0	0	0	0	0	0
J_Faulkner	166	3.54	27.6	0	527	5.12	65.8	610	791	7.78	0	0
D_Steyn	14082	3.28	41.4	21	7032	4.74	44.7	1662	2075	7.49	0	29.7808
	4198	3.51	93.2	0	8054	5.1	52.2	93	123	7.93	0	0
	19962	2.82	69.3	5	10636	4.82	39.3	2138	2742	7.69	0	26.2742
	66	3.18	-	0	411	6.14	205.5	427	623	8.75	0	58.4427
	78	4.46	-	0	36	5.16	36	12	25	12.5	0	0
	3731	3.04	93.2	1	4392	5.26	45.7	346	478	8.28	0	22.6378
	120	1.95	120	0	186	5.48	46.5	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0

The data highlighted in yellow above was excluded

4.5 Dealing with Missing or Poor Quality Data:

Sampling errors would not occur in this project as the data has been sourced from a sound, international database that has been subject to a large number of data queries from around the world. Therefore, the issue of missing data is more likely to arise from a scenario in which players tend to specialise in particular formats, resulting in an excellent record in one format but little or no achievements in another. This situation may lead to erroneous projections of performances and prices due to the fact that any regression model would use average results achieved in a specific format would be used as an increasing factor in projecting the price or performance of a player. This would lead to trouble when a particular player may have an excellent T20 record but failed to attain selection into the ODI team, thus resulting in the dataset containing a "0" survival rate for this player in ODI matches. As a result, the model would present a biased reflection of the batsman's abilities, who may have achieved a decent record had he been given more opportunities to play in this format.

One method of dealing with this problem is to use "development factors" in the same way that they are employed in the basic chain ladder method.⁷ This method is often required for the completion of run-off triangles and is used by actuaries to project both outstanding and IBNR* claim values for future periods in which data has not yet been received.⁷

In applications of cricket player analysis, several issues have arisen of either missing data or data that lacks credibility, for which development factors will then be required in order to estimate hypothetical statistics that would have been achieved by the batsman or bowler had he played more matches in a specific format.

* IBNR stands for "Incurred but not yet reported".

Certain sections of the dataset had to be adjusted to contain the projected values that were obtained through the use of the development factors⁷ calculated below. An example of this can be found in James Faulkner's Test bowling records (highlighted in green). Comparative examination with other top bowlers indicate that Faulkner's wicket rate (27.6) is far superior to that of Dale Steyn (41.4)*. It would be absurd, however, to conclude that James Faulkner has greater wicket taking abilities than Dale Steyn, the reigning number one Test bowler in the world. This example illustrates the lack of credibility of certain key statistics, which can contain some ridiculous values whenever there is a small sample size of "balls faced" or "balls bowled".

	<u>Test results</u>			
	TBalls	TEcon	TWR	TSw
Z_Khan	17612	3.25	59.7	10
J_Faulkner	166	3.54	27.6	0
D_Steyn	14082	3.28	41.4	21

* Recall that "wicket rate" represents balls per wicket, therefore a lower wicket rate is superior.

4.6 Adaption of the Basic Chain Ladder Method:

The Development factors for these projections were based on the Basic Chain Ladder method. The use of the Basic Chain Ladder method allows future values to be projected by applying past development trends to current or emerging values, and thus assuming that future "developments" of a player's statistics will emerge again according to a similar pattern as they have done so in the past. The primary use of this method is for general or short term insurance reserving, where the assumption underlying the projections is that past patterns of claims will repeat themselves in future years, only this time in proportion to the size of recent claim developments.⁷ This method of "filling in the blanks" of future cash flows has been adapted to fill in the gaps of what could be called "expected data" for batsmen and bowlers who may have not yet attained records that are statistically reliable enough for modelling purposes. This mostly occurs due to inexperience, but on some occasions, this may happen when a certain top player has had reduced opportunities to attain sufficient records in a particular format.

For example, Sunil Narine is an outstanding Twenty-twenty bowler, but has had very little exposure to other formats of the game. If his statistics were left as they are, his projected results would be significantly reduced simply because of a lack of opportunities to play other formats.

Sunil Narine: Bowling averages

	Mat	Inns	Balls	Runs	Wkts	BBi	BBM	Ave	Econ	SR	4w
Tests	5	9	1299	721	15	5/132	8/223	48.06	3.33	86.6	0
ODIs	39	39	2123	1464	60	5/27	5/27	24.40	4.13	35.3	4
Twenty20	95	93	2161	1933	123	5/19	5/19	15.71	5.36	17.5	6



Below is an example of how the development factors used in this method were calculated in Excel:

5	Overall survival rates:		Development factors:	
6	Average survival rate for Test matches :	84.01	To find missing Test survival rate:	$2.3371648 \cdot D6 / ((D7 + D8) / 2) * ((ODI + T20) / 2)$
7	Average survival rate for ODI matches :	48.04929794	To find missing ODI survival rate:	$0.8702345 \cdot D7 / ((D8 + D6) / 2) * ((T20 + Test) / 2)$
8	Average survival rate for T20 matches :	25.02609081	To find missing T20 survival rate:	$0.3751218 \cdot D8 / ((D6 + D7) / 2) * ((Test + ODI) / 2)$
9				
10				
11				
12				
13				
14	Average strike rate for Test matches :	53.98638909	To find missing Test strike rate:	$0.4464778 \cdot D6 / ((D7 + D8) / 2) * ((ODI + T20) / 2)$
15	Average strike rate for ODI matches :	86.11444444	To find missing ODI strike rate:	$0.9573676 \cdot D7 / ((D8 + D6) / 2) * ((T20 + Test) / 2)$
16	Average strike rate for T20 matches :	131.3572222	To find missing T20 strike rate:	$1.9508428 \cdot D8 / ((D6 + D7) / 2) * ((Test + ODI) / 2)$

In the above calculations:

D6 =Average Test survival rate for all players

D7 =Average ODI survival rate for all players

D8 =Average T20 survival rate for all players

Test = The Test survival rate for the Individual player for whom the Development factors are being used to provide hypothetical statistics

ODI = The ODI survival rate for the Individual player

T20 = The T20 survival rate for the Individual player

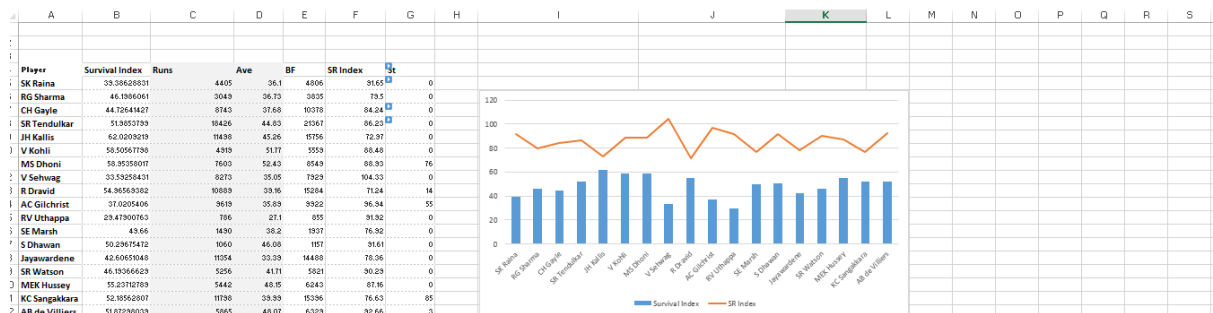
A particular player's lack of opportunity to play enough ODI matches resulted in a statistically unreliable sample size batting innings' for this format. This was easily adjusted for by projecting his hypothetical ODI statistics through the multiplication of his T20 results by development factors that were derived from the ratio of all batsmen's T20 records with respect to their combined ODI and Test records.

Of course, the most accurate way to fill in these gaps in a player's record would be to use a linear regression model. For example, the T20 averages of each batsman could be modelled against his Test and ODI records, with the resulting equation of $[T20ave = B_0 + B_1 \cdot ODlave + B_2 \cdot Testave]$ which minimises the squared errors between the model and the actual T20 results of all the players. This average could then be used to project hypothetical results for those players that have insignificant T20 figures by applying the derived model to their more reliable statistics obtained in other formats.

The accuracy of this method can be further extended by including all player factors and using diagnostics to systemically eliminate those factors that are either dependent or do not correlate with the T20 records. This will be explored later in the assignment.

The Final Dataset contains 63 description factors for each player, some of which are pictured below:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE																	
															II : Indian Idol										Use of reference parameter																						
															RA : Renowned Aussie										Yellow reference parameter																						
															GIS : Global Iconic Status										Yellow dot and blue colour based																						
															MIP : Mid-way Indian player (Future in National Team)																																
															SBR : Slow batsman Reputation																																
																									2013 IPL Results																						
Player	TSurvival	TRuns	Test results		TSR	TADO	Tst	ODI results					T20 results					IPLSurv					IPLRuns					IPLAve					IPLSR					IPL6s					II				
			Ave	TSR				ODISurv	ODIRuns	ODIAve	ODISR	ODISR	ODISR	ODISR	T20Surv	T20Z	T20M	T20F	T20SR	T20ZSR	IPLSurv	IPLRuns	IPLAve	IPLSR	IPL6s	I																					
Shane Warne	82.4202028	763	21.44	646	53.29	17	0	34.294203	0	0	0	0	0	24.826	4284	2647	2087	2408	2408	2408	4.777	2192	35.6	1962	1437	0	1																				
Shane Watson	82.0204047	0	0	0	46.1709011	0	0	0	0	34.949	2328	78.9	0	24.616	4282	2648	2083	2340	2340	2340	2	28.93242	2893	32.43	1920	128.94	0	1																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	46.1204662	1740	37.4	0	0	0	24.617	4782	2647	2143	2143	2143	2143	2	24.6489	2314	34.63	1962	108	0	1																			
Steve Smith	80.7302048	18937	59.31	20544	92.33	81	0	0	0	34.937974	16624	44.93	1021	84.23	27.078	2797	324	2316	1935	1935	0	24.60477	2334	34.63	1968	108.14	29	0																			
David Warner	82.0203035	12440	39.44	4662	46.62	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Virat Kohli	80.7302744	1706	41.94	2653	56.34	0	0	0	0	34.9566779	16624	44.93	1021	84.23	27.078	2797	324	2316	1935	1935	0	24.60477	2334	34.63	1968	108.14	29	0																			
AB de Villiers	82.0203035	4249	47.2	7699	76.99	0	0	0	0	34.9566779	16624	44.93	1021	84.23	27.078	2797	324	2316	1935	1935	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
David Warner	82.0203035	12440	39.44	4662	46.62	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29	0																			
Glenn Maxwell	79.9203036	4324	48.2	1073	1073	0	0	0	0	34.9209197	16862	44.62	9784	82.97	0	24.620	2897	2324	2082	2082	0	24.60477	2334	34.63	1968	108.14	29																				

[illegible][illegible]

5 Initial Model Design

5.1 General Linear Models:

This model will be expected to project player performances in the IPL using data collected from all the records of a player's performances to date. Response variables will be chosen from the data, which would typically be price, T20 Average or any factor for which a linear regression could be performed against the other factors or characteristics of a cricket player. Thus a general linear model will be employed.^{2,3}

General linear models are used to quantify the relationship between the predictors and the volatility of the response variable⁸. In other words, find the coefficients of the predictors (Dependent variables) that explain this relationship as efficiently as possible by minimising the squared errors between the model prediction and the actual value of the response variable.

If the response variable is denoted as y , and the model for the response variable is $(B_0 + B_1 x + B_2 x_2)$ containing covariates x_1 and x_2 , then the total sums of squares would be $\sum y^2$ whereas the corrected total shown below would be $\sum (y - \bar{y})^2$. This corrected total is equivalent to the sum of squares for the model (SSM) added to the sum of squares for the error (SSE). Where $SSM = \sum [(B_0 + B_1 x + B_2 x_2) - (B_0 + B_1 \bar{x} + B_2 \bar{x}_2)]^2$ and $SSE = \sum [y - (B_0 + B_1 x + B_2 x_2)]^2$ ⁸

It is this value of SSE that is sought to be minimised. In other words, the “unexplained variance”.⁵

It could also be said that the total corrected sums of squares is equivalent to the “explained variance” + “unexplained variance”.⁵

Example of SAS output for the sums of squares in a GLM procedure :

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4630.36091	926.07218	13.56	<.0001
Error	194	13248.51409	68.29131		
Corrected Total	199	17878.87500			

Later on, the distributions of scores for the different players will be looked into for the modelling and simulation of full team totals.

5.2 Structure of the Model:

The response variables would ideally reflect the sought after performance attributes of a cricket player in terms of what the team currently lacks, in order to test the likelihood of a player delivering such performances given his past record, which would make up the model's independent variables.

An example would be to use a player's ODI Strike rate, T20I Strike rate, Total ODI runs scored, T20I caps etc. as a set of independent variables which could be combined to form a linear predictor⁸ which is then used to forecast his expected IPL Strike rate for the season of

2013. Recalling the previous discussion regarding the dynamics of a Twenty-twenty cricket side, it may be found that the team are in need of a pinch hitter. The next stage of the process would then be to refer to the general linear model through which the factors in a player's characteristics that are most consistent in predicting an increased strike rate could be highlighted. These would be the factors that return the highest coefficients after running a "proc GLM" procedure in SAS, for which the IPL Strike rate achieved by a batsman is modelled against all other factors in his record.⁹

The task at this point would be to determine which records and statistics are the most relevant for use in a SAS GLM procedure. Perhaps a more systematic and thorough approach would be to include all the data that can be collected on a player's past performances, and model the dependant variable (i.e. the projected performance of some kind, such as strike rate) against all factors in a player's record. The output could then be used to refine the selection of independent variables for this model, excluding those which have a significant probability of not influencing the dependant variable.⁹ Refer to the example below:

SAS Output:

The GLM Procedure

Class Level Information

Class	Levels	Values
T20Survival	2	0 1
ODISurvival	3	1 2 3

Response variable: IPLSurvival

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4480.75918	926.07218	13.56	<.0001
Error	194	13668.76596	68.29131		
Corrected Total	199	17478.46500			

R-Square	Coeff Var	Root MSE	write Mean
0.258985	15.65866	8.263856	52.77500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
T20Survival	1	1141.853291	1261.853291	18.48	<.0001
ODISurvival	2	4264.350821	1637.175410	23.97	<.0001
T20Survival*ODISurvival	2	305.958889	162.979094	2.39	0.0946

The F Value highlighted above tests the null-hypothesis that the movement of one of the independent variables does not have a significant influence on the dependant variable, given the influence of other variables in the model. This F-Value is calculated as MSI / MSE , where I refers to the independent variable being tested. The F- Value follows an F- distribution where the numerator is the degrees of freedom of the variation caused by the component being tested, and the denominator is the degrees of freedom of the error, where "error" refers to the deficit between the dependant variable and the value predicted by the model.⁹

5.3 Interactions in the Model

Some unorthodox interactions may be required in the model as well, and will have to be built in manually in order to account for the following potential biases and discrepancies that may arise:

Certain players may have achieved an excellent strike rate or survival rate after only two or three performances. This could skew projections in favour of a player who may have achieved excellent results on debut, whether through luck or skill, it is hard to determine which is the case on this occasion. Therefore it may be necessary to set a certain minimum of innings' to be faced before a player's particular strike rate or survival rate should count in their favour. Likewise for bowlers, a certain minimum number of balls need to have been bowled in a particular format before the achieved strike (wicket taking) rate or economy rate is counted in their favour.

Perhaps a more comprehensive approach would be to create an interaction between "Number of Balls faced" and the two main performance indicators, i.e. "Strike rate" and "Survival rate". This interaction would account the contribution of the two main indicators to the overall price or performance value of a player as a function of the number of Balls faced.

For example:

Performance indicators could be expressed as a multiple of Balls faced:

Performance indicator = (Strike rate) * (Balls faced)

5.4 Weighting the factors:

The problem with the above approach is that the actual strike rate would be a far more influential indicator than the amount of time for which he attained this rate. So to simply multiply the two factors would be to give "Balls faced" an equal weighting in the model as strike rate, which would disproportionately favour the more experienced players.

It is therefore necessary to use the number of Balls faced as a smaller factor in proportion to the effect that the strike rate would have on a player's price and performance projections. This could be expressed as a function of Balls faced where it is used as an additional factor of benefit with only a ¼ size weighting. Two examples include:

Performance indicator = (Strike rate) * (1 + Balls faced / [4 * Average (Balls faced)])

Performance indicator =
(Strike rate) * (1 + ¼ (Balls faced - Average (Balls faced)) / Average(Balls faced))

Unfortunately there is still a residual inaccuracy within this method, as players with very little experience may still achieve a significantly high strike rate over one or two innings' where the reduced effect of "Balls faced" on the indicator will not be sufficient to make up for a lucky big performance on debut etc. This kind of outlier would have to be dealt with by building an algorithm into the model where "Balls faced" is used as both a qualifying factor as well as an additional factor of benefit, which would include a minimum number of Balls faced.

The algorithm below serves as an example:

If Balls faced > (lower bound) then:

(Strike rate) = (Strike rate) * (1 + $\frac{1}{4}$ *Balls faced / [Average (Balls faced)])

Else :

(Strike rate) = (Strike rate predicted by Development factors)⁷

6 Principal Component Analysis

The methods for player performance modelling introduced in the previous section will be based on the dataset constructed in section four.** This dataset contains no less than 63 descriptive factors for each player, and the process of modelling this sprawling array of data could prove an arduous task. It is therefore essential to summarise these factors into a parsimonious* model of a few linearly independent performance indicators.

*A parsimonious model has had the number of variables minimised without sacrificing too much accuracy.

**See page 16 of this assignment

The independence of these key indicators is very important as they will eventually be used as a basis for pricing* and ranking* players according to their performance value to the team. If the key performance indicators are in any way dependent on one another, the monetary value assigned to each unit measure of performance will overlap, in other words the value paid for a unit increase in performance index A will be paid in addition to a monetary value associated with an increase in performance factor B. If these two factors are dependent on each other there will be a double charge.

Most data available for batsmen would involve standard batting measures such as runs scored, strike rate, average, balls faced, number of hundreds, innings' faced etc. Rather than simply modelling all these variables in a GLM procedure, a pre-analysis of these variables could be carried out in order to "clean up" the data and summarise it all in a few key variables.

*(see page 37 -39)

PCA Overview

- Principal components analysis is concerned with explaining the variance-covariance structure of a set of variables.
- This explanation comes from a few linear combinations of the original variables.
- Generally speaking, PCA has two objectives:
 - ♦ "Data" reduction - moving from many original variables down to a few "composite" variables (linear combinations of the original variables).
 - ♦ Interpretation - which variables play a larger role in the explanation of total variance.
- Think of the new variables, called the principal components, as composite variables consisting of a mixture of the original variables.

6.1 Continuous, Semi-Continuous and Categorical Descriptive factors:

There are two kinds of descriptive factors in the dataset belonging to this assignment:

Continuous* DF, non-Continuous DF and Categorical DF.

(DF : *Descriptive factor*)

*These are, of course, not mathematically continuous variables. However, they are continuous from the perspective of linear regression modelling, which means “continuous in contrast to categorical”, rather than “continuous in contrast to discrete”.

For the purposes of simplifying the regression model, as many categorical variables as possible were turned either into dummy variables (1 or 0) or semi-dummy variables (1, 2 or 3). Therefore, the “transformed categorical” factors in the dataset would make up the non-continuous descriptive factors.

The list of continuous descriptive factors is as follows:

For Batsmen:

Runs, Dismissals, Balls faced, Strike rate, Average, Fifties, Hundreds, Sixes, Fours, Catches and Stumpings.

For Bowlers:

Wickets, Runs Conceded, Balls bowled, Economy rate, Average, Catches, Stumpings, Five-fors and Four-fors.

The techniques applied to both batsmen and bowlers will be almost symmetrical to each other, as reflected in the two sets of continuous descriptive factors above. However, for the sake of brevity, only batsmen will be analysed in the assignment from here onwards.

6.2 Distinguishing Reputation Factors from Performance Factors:

The factors that are categorical or semi-continuous are mostly those that affect a player's reputation (and hence, his price). Therefore, all these factors will be named “reputation factors” and are characterised as being more useful in determining a player's price than his performance value to the team.

The continuous factors are split between those that determine reputation, and those that determine performance. The reputation factors amongst the continuous factors are those that are associated with hype and fame such as “Hundreds”, “Sixes”, “Fours”, “Five-fors” and “Four-fors”. Catches and Stumpings are also continuous factors, but they refer to the fielding abilities of a player, and will only be used in an extension of this assignment.

“Fifties” may seem like a performance factor in which a large number of fifties may indicate consistency. However, it is actually just another reputation factor used to determine price rather than performance, despite the fact that it can be used quite efficiently to determine performance. This is because a batsman's “Runs” and “Average” can be assessed in combination with an Analyses of Variance (ANOVA) of his individual innings scores. This will

more than suffice as a combination of factors that quantify a player's ability to perform with consistency, and would do the job with a lot more precision than the use of a factor such as the number of fifties.

Thus reputation factors are not only those factors that are associated with hype and legend, but can also be identified as the factors that do in fact describe a batsman's performance attributes, but with less efficiency or precision than the actual performance factors. This renders them redundant from a performance measurement perspective, which leaves only the five performance factors remaining as those that will be used for the assessment of performance value to the team. The reputation factors will not be deleted though, as they will be useful in determining which aspects of a player inflates his price without "inflating" his performances on the field to the same extent.

So there are now two sub-categories of continuous factors, namely performance factors and reputation factors, whereas the remaining non-continuous factors are all of the reputation kind. This distinction is the key to determining whether a player is over or under priced with respect to the specific needs of the client franchise.

An example of how this is done would be to assess the regression coefficients of the dummy and semi-dummy variables when performing a linear regression of "Price" against all of the descriptive factors. The dummy and semi-dummy variables are all reputation factors such as "Iconic player", "Indian Idol" or "Aussie Legend". The regression model would contain all of these dummy variables as part of a long list of 63 descriptive factors* which all act as the covariates or linear predictors which are fitted together into a single linear predictor that models the price. (The question of dependence between these factors will be dealt with later)

The regression coefficients of these dummy variables would then indicate the average "premium" that is paid for a certain factor being present. Such as the "Aussie Legend" premium or the "Indian Idol" premium. These premiums can then be compared with the increase in earnings figures estimated from the commercial benefits of having such a player in your arsenal. This relates to all factors around fan-fare such as advertising endorsements and T-shirt sales. ^{2,4}

These could be named "Media factors". An analysis of Media factors could form part of a possible extension to this assignment.

6.3 The Key Factors:

There is yet another subset of descriptive factors, which will make up three out of the five performance factors. These three independent and all-encompassing descriptive factors will be named "Key factors". After a short analysis of the performance factors (which are a subset of continuous, descriptive factors) it becomes clear that there are now two subsets of three central attributes of a batsman or bowler that are seemingly independent of one another.

These two sets of three attributes are:

For Batsmen:

Runs, Dismissals, Balls faced.

Survival rate*, Strike rate and Balls faced

For Bowlers:

Wickets, Runs conceded, Balls bowled

Economy rate, Wicket rate** and Balls bowled

Hence, there are two groups (one for batsmen and one for bowlers) each of which consist of two subsets of seemingly independent factors. Neither of these two groups are mutually exclusive; in other words, both groups of two subsets include a common factor, namely balls faced for the batsmen and balls bowled for the bowlers.

*The Survival rate is a creation of this assignment, it will only be found in the dataset that was created for this assignment.

**Only this assignment refers to a bowler's "Strike rate" as "Wicket rate", (Regarding Balls per wicket). This was done to distinguish between a batsman's Strike rate and the Strike rate of a bowler.

6.3.1 Only Three Performance Attributes Required:

The entire scope of performance characteristics of a batsman could be captured and summarised in either of the two sets of three attributes. The same applies for bowlers. (All-rounders would have six Key attributes)

These three attributes provide the full spectrum of a player's performance factors, as every performance factor can be derived from a combination of these three attributes.

For example; when using Survival rate, Strike rate and Balls faced as the three key batting attributes:

$\text{Runs} = (\text{Strike rate}) * (\text{Balls faced})$

$\text{Average} = (\text{Strike rate}) * (\text{Survival rate})$

$\text{Dismissals} = (\text{Survival rate}) * (\text{Balls faced})$

Therefore, the other three performance attributes are simply derived from the key attributes in use, which indicates that only these three factors are needed to describe the full scope of a batsman's performance characteristics. It can be thought of as three dimensions, opening up a possible frame work for 3D-models that visually capture the performance characteristics of each player in a more tangible presentation, which can form part of an extension to this assignment.

6.3.2 Which Three Performance Attributes are Independent?

It has been demonstrated that only one of the above two subsets of performance factors are required for the assessment of batsmen, while it can be intuitively observed that the same situation applies to the bowlers as well. However, which of the two subsets should be selected? To answer this question it is important to recognise which aspects of a batsman's performance characteristics are reflected in a particular statistic.

Consider the first subset of performance factors:
Runs, Dismissals and Balls faced

“Runs” represents two central qualities:

- Ability to score a lot of runs per Balls faced , in other words, “Strike rate”
- Ability to face a lot of balls, in other words, “Survival”

This is because $\text{Runs scored} = \text{Strike rate} * \text{Balls faced}$.

It does not represent a pure, single quality of a player but it represents both his ability to remain at the crease, as well as his ability to score fast. In addition, it also represents his ability to play a lot of matches, in other words, his ability to please selectors and keep away from injuries.

“Dismissals” also represent two qualities of a batsman:

- How easily he goes out
- How many innings’ he has faced

The problem that arises here is that the longevity of a particular player is reflected in two different performance attributes simultaneously, which indicates that the first subset factors are actually not independent after all.

This situation is in contrast to what is found in selecting the other subset of performance factors:

Survival rate, Strike rate and Balls faced

Survival rate represents his ability to stay at the crease.

Strike rate represents his ability to score fast.

Balls faced represents how long he maintained this level of Survival and Strike rate.

It is clear therefore, that the second subset of performance factors do not present the same problem of a particular attribute being shared by two or more factors. Therefore, unlike the first subset, the second subset of factors are independent of each other. They represent clear cut qualities of a batsmen without blurring the lines.

It is important that the performance factors are completely independent for the purpose of regression modelling, so the first subset of factors will have to be rejected, leaving only three factors remaining. These three factors are the independent performance factors.

By the same logic, the three independent performance factors for bowlers would be the Wicket rate, Economy rate and Balls bowled.

Note: The remaining analysis of players in the assignment will only focus on Batsmen. As illustrated in the above section, most of the techniques applied to the batsmen can be mirrored in their application to bowlers.

7 Survival and Strike rate Indices

If the performance value of each player could be condensed into two simple indices, there would be a significant enhancement to the simplicity with which direct comparisons could be made between price and performance value. It would also enable both the isolation and the quantification of the cost of purchasing certain performance enhancements to your team, such as an increased team run rate or an increased survival rate of your batting line up.

The principal component analysis (above) has revealed that only three key attributes are required to summarise the performance value of batsmen and bowlers respectively, and that a set of three independent attributes (Key attributes) can be found to fit this purpose. Therefore, the survival and strike rate indices will be constructed from only these three independent factors in proportion to the reliability of the data from which the factors were derived.

For this purpose, credibility factors¹⁶ will be assigned to each data source in proportion to their relative reliability. The credibility factors will need to be based on the relative fit of the descriptive data to the actual performances attained by each player in the IPL.

The rest of section 7 will be based on finding these credibility factors from which the survival and strike rate index can be calculated.

7.1 Direct and Composite Sources of Data

The “direct” data that is used in assessing a player’s potential in the IPL would be his record of statistics that were achieved in the IPL itself, whereas the “composite” data would consist of a player’s overall T20, ODI and Test statistics.¹⁶ The reliability or “credibility” of the composite data will need to be assessed in comparison to that of the direct data. Therefore, credibility factors¹⁶ are assigned to each of these data sections, which are obtained through the testing of the correlations and sensitivities of the IPL records (direct data) in comparison to those of the Test, ODI and T20 records (composite data) with respect to the results attained in the IPL 2013 season. For example, the similarities between a high ODI strike rate and a good strike rate in the IPL 2013 season will be compared to the similarities that exist between a player’s historical IPL strike rate and the strike rate that was achieved in the 2013 season.

The resulting equation would be:

$$\text{Survival Index} = C [\text{IPL Survival}] + (1-C) [\text{T20 Survival} + \text{ODI Survival} + \text{Test Survival}]$$

Where C and (1-C) are obtained through three general comparisons:

1. A comparison between:
 - The correlations of the IPL 2013 results* with those of the complete IPL records of all the players (direct data)

- The correlations of the IPL 2013 results* with those that would have been predicted by the composite data. (T20, ODI and Test figures)
- 2. A comparison of the different R-square statistics that are obtained through a regression of the 2013 results against the key factors projected by the direct and composite data respectively.
- 3. A comparison of the regression coefficients obtained in a model of the 2013 results against the key factors that were obtained from the direct and composite data respectively

*The “results” of the IPL 2013 season refer to the key factors that were derived in the principal component analysis, namely the survival rate, strike rate and balls faced.

Thus the key factors of performance in the IPL 2013 season are compared to those key factors that were obtained from the full IPL records of all the players as well as those that were forecast from the composite data. (The process of which is explained below)

7.2 Forecasting Key Factors from the Composite Data:

In order to derive a survival rate from a dataset containing three formats, the respective weightings of the T20, ODI and Test statistics would need to be determined. In other words, credibility factors need to be assigned to the different components of the composite data.

This can be done by assessing the relationship between the Test, ODI and T20 results with the full IPL records of each player, which will provide more reliable results than if they were compared to only the 2013 results. (Due to a much larger sample size of data in the full IPL records)

For example, the relevance of a batsman’s Test survival rate will need to be assessed in terms of how well it indicates his likely survival rate in the 2014 IPL season. This needs to be compared with the relevance of his ODI survival rate for the same task. The assessment will be made by comparisons between the relationships of his Test and ODI survival rates with the IPL survival rate respectively.

Three credibility factors are thus obtained for the three formats which combine to make up the composite data.

7.3 Standardising the Dataset:

One issue that still needs to be resolved is the differences in means and variances of Test, ODI and T20 data. This could be solved by means of standardisation. To standardise the data means to subtract the mean of a sample set from each element and then divide the elements by their standard deviation. This will be done for all the data in the dataset used in the regression and correlation analysis. Therefore credibility factors will be obtained from the correlations of Test, ODI, T20 and IPL results according to the number of standard deviations from the mean that occurred. This eliminates all other noise from the data which

[illegible]

7.4 Overview:

1. A comparison between the correlations of the IPL results with the Test, ODI and T20 records of all the players respectively.
2. A comparison of the different R-square statistics that are obtained through a regression of the IPL results against the key factors projected by Test, ODI and T20 records respectively.
3. A comparison of the regression coefficients obtained in a model of the IPL results against the key factors that were obtained from the Test, ODI and T20 records respectively.

$$\text{Survival Index} = C [\text{Direct data}] + (1-C) [\text{Composite data}]$$

$$= C [\text{IPL Survival}] + (1-C) [C_{T20} (\text{T20 Survival}) + C_{ODI} (\text{ODI Survival}) + C_{\text{Test}} (\text{Test Survival})]$$

27

7.5 Using SAS to find the Credibility Factors:

In order to test the regressions of the T20, ODI and Test statistics with those that were achieved in the IPL, SAS will be used to run a general linear regression between the above-mentioned variables in the dataset after it has been standardised.

The regression coefficients of the records that were achieved in other formats will represent the sensitivities of the IPL statistics to these factors, which can in turn be used as basic credibility factors to be applied to the composite data. However, these regression coefficients cannot be blindly used as credibility factors without first testing other characteristics of these variables. The R-square statistic, F-Test statistic, SSE and MSE of each predictive factor in the composite data will need to be tested with respect to their relationship with the direct data (the player's full IPL record). Correlations between the covariates may also affect the results while the differences in the proportions of Test and ODI figures compared to the T20 figures will need to be accounted for.

Therefore the R- square statistic, regression coefficients, SSE, MSE and F-test results may all need to be derived from a regression analysis in SAS, and are then combined with the correlation data between these variables which could be derived in excel.

With reference to the principle component analysis in section six, only the survival rate, strike rate and balls faced will be considered for performance assessment. The other factors will come in handy when assessing the characteristics that affect a player's price.

General linear regression modelling in SAS is able to provide the majority of the statistics required to find the credibility factors. However, Excel will be required to perform an Analysis of Variance on the dataset in order to assess the different correlations between the predictive data and the response variables.

There are also two major concerns that need to be addressed before performing any regression analysis:

1. ODI Survival rates are larger than the T20 survival rate, and would also therefore have a larger variance proportional to the variance of T20 survival rates. This would skew the combined figure and make it a poor representative of IPL survival rates.
2. The two covariates may be highly correlated. The issue of correlation is relevant in that a regression analysis of the IPL survival rate is being performed against these three covariates simultaneously. Multiple regression modelling requires independence of the covariates in order to find the true sensitivities of the response variable to movements of each individual covariate.

The first issue above will be solved by means of standardising the variables. (see pages 26-27)

The second issue requires the batsmen's records from the different formats be modelled one variable at a time against the response variable*, rather than to model them simultaneously. This will provide regression coefficients that indicate the true sensitivity of the response variable to each covariate respectively.

However, for the purpose of assessing which variable to omit from the model due to irrelevance, the covariates will have to be modelled simultaneously to for comparisons of F-test and t-test results to be made. In this case it won't be necessary to worry about whether the covariates are correlated or not as an irrelevant variable that fails the F-tests can't possibly be correlated with the other variables anyway.

*The response variable is the key factor (survival, strike rate or balls faced) that is derived from the IPL records. This is modelled against the same key factor that is derived from each set of composite data respectively, say T20, Test or ODI records.

7.6 SAS Example:

The Section below serves as an example of how the above principles can be applied in SAS. The example below has only been applied to the application of finding the credibility factors of the composite data for survival rates. i.e. C_{T20} , C_{ODI} and C_{Test} .

7.6.1 Simultaneous modelling of variables to test relevance:

In order to make the regression models parsimonious*, it may be necessary to first check whether certain covariates are irrelevant in modelling the IPL results. This can be done by means of an F-Test** and a t-test, both of which are performed automatically in SAS.

*A parsimonious model has had the number of variables minimised without sacrificing too much accuracy.

** F-tests...

The following example illustrates the above process applied in SAS:

SAS Code:

```
proc glm data = D1;
model IPLSurvival = TSurvival ODISurvival T20Survival /ss3;
run;
```

SAS Output :

Dependent Variable: IPLSurvival

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2791.399539	930.466513	33.25	<.0001
Error	24	671.640747	27.985031		
Corrected Total	27	3463.040286			

R-Square	Coeff Var	Root MSE	IPLSurvival Mean
0.806055	30.09885	5.290088	17.57571

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TSurvival	1	0.8452942	0.8452942	0.03	0.8635
ODISurvival	1	131.2776899	131.2776899	4.69	0.0405
T20Survival	1	119.9685664	119.9685664	4.29	0.0493

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.0455660200	2.02135299	0.02	0.9822
TSurvival	-0.0117140096	0.06740072	-0.17	0.8635
ODISurvival	0.3081962804	0.14229670	2.17	0.0405
T20Survival	0.3843143142	0.18561613	2.07	0.0493

Both the F-test* and t-test* on these three covariates indicate that a batsman's ability to survive in Test matches is not very predictive of his IPL survival rate, so this covariate will be excluded and the model will be run again while the SSE* and MSE* of the model are monitored to see how this exclusion affects the model's accuracy.

The output produced the following:

Predictors	R-Square	SSE	MSE
Without Test survival rate	0.805811	672.486041	26.899442
With Test survival rate	0.806055	671.640747	27.985031

The new output therefore indicates that the exclusion had a negligible effect on model accuracy.

7.6.2 Standardising the variables:

The standardised IPL Survival rate was then modelled against the new standardised covariates: (with Test survival removed as explained earlier)

```
model IPLSurvival = ODISurvival T20Survival / ss3;
```

Selected Output:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ODISurvival	1	1.84021583	1.84021583	8.77	0.0066
T20Survival	1	0.93605433	0.93605433	4.46	0.0448

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.0000000001	0.08654891	0.00	1.0000
ODISurvival	0.5405546248	0.18249284	2.96	0.0066
T20Survival	0.3855278299	0.18249284	2.11	0.0448

By the F-tests, any hypothesis that either of these standardised variables lack predictive power would be rejected with a 95% level of confidence. What is interesting, however, is that the ODI Survival rate still has a larger effect on the IPL Survival rate, even after the standardisation of all variables.

7.6.3 Issues of Correlation:

What could make this result deceptive however, would be the possibility of a strong correlation between the two predictors. The covariance of the two variables is 0.844, which results in a correlation of 0.844 because the variables both have a variance of 1 after being standardised. This is a very high correlation and indicates that a simultaneous regression would not provide very accurate indication of the sensitivity of our response variable to either covariate. Therefore the IPL results will be modelled against these factors one at a time and the different sensitivities will be compared to see which covariate has a better predicting power:

```
model IPLSurvival = T20Survival / ss3;
```

```
model IPLSurvival = ODISurvival / ss3;
```

Predictors	Coefficient	R-Square	SSE	MSE
ODI Survival	0.8781387	0.771128	6.17955395	0.23767515
T20 Survival	0.8588598	0.73764	7.08371545	0.2724505

According to all of the above measures, the ODI Survival rate remains a superior predictor of a batsman's survival rate in the IPL.

It is important to note that this does not imply that the ODI Survival rate is more similar to the IPL Survival rate in terms of their respective values. Actually, the mean of the ODI Survival rate is approximately double that of the IPL, therefore the ODI survival rate would be very different from the IPL survival rates in the size and variance of its values*. However, the greater usefulness of this statistic in comparison to the T20 Survival rate is derived from the fact that the standardised variations in the ODI Survival rate have been proven to imitate the variations of the IPL results more closely than that of the T20 Survival rate. Therefore, batsmen that display a greater ODI survival rate are more likely to display a greater survival rate in the IPL.

* This is why the statistics are standardised, and only the standardised variations are used for projections.

This may be a counter-intuitive result, as many might have assumed that T20 records would always be a superior predictor of IPL results, which illustrates the need to test what may seem like an obvious conjecture in the cricket environment.

Given the high level of correlation between these two predictors, it may be superfluous to try and combine them in order to create a superior measure of survival rate. However, statistics are almost always more reliable when they incorporate more information, as long as the

information is relevant. Therefore, it may still be more beneficial to combine the two covariates in proportion to their relative regression coefficients with the IPL results, which will be used as part of their credibility factors.

7.6.4 General equation:

The following general equation for the composite survival index is thus obtained:

$$\text{(Standardised Survival)} = (\mathbf{C}_{\text{ODI}} * \text{Standardised ODI Survival}) + (\mathbf{C}_{\text{T20}} * \text{Standardised T20 Survival})$$

Where:

$\mathbf{C}_{\text{ODI}} = \mathbf{f}$ (Coefficient R-Square SSE MSE)

$= \mathbf{f}$ (0.8781387 0.771128 6.17955395 0.23767515)

$\mathbf{C}_{\text{T20}} = \mathbf{f}$ (Coefficient R-Square SSE MSE)

$= \mathbf{f}$ (0.8588598 0.73764 7.08371545 0.2724505)

Figures obtained from the following table derived from the SAS output :

<u>Predictors</u>	<u>Coefficient</u>	<u>R-Square</u>	<u>SSE</u>	<u>MSE</u>
ODI Survival	0.8781387	0.771128	6.17955395	0.23767515
T20 Survival	0.8588598	0.73764	7.08371545	0.2724505

NB: This is just the Survival rate obtained from the composite data. (Tests, ODI's and T20's)
This figure will still have to be combined with the direct data (IPL survival rate) by using the main credibility factor that will be calculated in section 7.7.

This figure must then be re-adjusted to be in proportion to the average sizes of T20 survival rates, which would be done by reversing the standardisation process and thereby returning it back to normal proportions.

Therefore:

$$\text{Survival} = \text{Standardised Survival} * \sqrt{\text{Var}(\text{T20 Survival rates})} + \text{average}(\text{T20 Survival rates}).$$

Completing the Survival index:

The accuracy of the modelling used for this assignment can always be enhanced through the further development of the above mentioned functions of the credibility factors, which in turn can be derived as a complex combination of the four parameters: (The Regression Coefficient, the R-Square test statistic, the SSE and MSE) However, for the purposes of this assignment, only an example of the derivation of credibility factors will be needed, thus a simple function of $(R^2_{\text{Statistic}}) * (\text{Regression Coefficient})$ will suffice.

Likewise, the same methods applied to finding the credibility factors for the composite data in the above SAS example will also be applied to derive the main credibility factor that can be applied to the direct and composite data in order to generate the final survival index.

7.7 Assessment of the Strike Rate Index:

It would normally be assumed that the entire process that was carried out above to find the survival index would then be repeated in order to obtain the strike index.

However, it would be a big waste of time if all the analysis above were to be repeated only to find out at the end that most of the data was not relevant or useful in describing the strike rate achieved by a batsman in the IPL. Therefore it is important to first check the relevance of the composite data before making use of it.

Recall that the dependence of covariates does not matter* when checking for which factors can be eliminated, therefore a multiple regression model can be used despite the fact that the three variables in the composite data are clearly not independent** of each other.

*Refer to section 7.6.3 (Issues of Correlation) on page 31

** The key factors (survival, strike rate and balls faced) were carefully selected to be independent of each other; however, the data for each factor will not be independent across the different formats (Test, ODI and T20) which make up the three covariates in the regression model. (which are also the three variables in the composite data)

SAS code and output for the multiple regression model :

```
model IPLSR = TSR T20SR ODISR/ss3;
```

Selected Output:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TSR	1	185.741736	185.741736	0.38	0.5454
T20SR	1	4836.047018	4836.047018	9.79	0.0046
ODISR	1	23.931090	23.931090	0.05	0.8276

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-2.642516534	8.91186989	-0.30	0.7694
TSR	-0.339565225	0.55364463	-0.61	0.5454
T20SR	1.031135509	0.32948309	3.13	0.0046
ODISR	0.161198142	0.73221974	0.22	0.8276

It appears that the ODI and Test strike rates are both poor predictors of a batsman's IPL strike rate, as indicated by the SAS output above.

Both F-tests on the Test and ODI Strike rates indicate that there are significant probabilities that these covariates have no predictive power at all for a batsman's IPL Strike rate. The small regression coefficients indicate a very low sensitivity of the response variable (IPL Strike rate) to both the Test and ODI Strike rates.

Hence, the above results indicate that only the T20 strike rate should be used to make up the composite data, resulting in no more requirements to calculate credibility factors for the composite data when forming a strike rate index.

7.8 Adjustment for Balls faced:

The calculation of a survival or strike rate index would not be complete without incorporating all three of the key factors that were identified in section six, where it was discovered that these three factors are both necessary and sufficient to determine the full spectrum of a batsmen's performance value to the team.

Therefore, the number of balls faced will have to be incorporated into both indices as a factor of improvement.*

* A factor of improvement refers to a factor that can either increase the index or do nothing, it cannot decrease the index.

With reference to pages 19 - 20 under the section "weighting the factors", the following algorithm was proposed:

If balls faced > (lower bound) then:

*(Strike rate) = (Strike rate) * (1 + ¼ *Balls faced / [Average (balls faced)])*

Else :

(Strike rate) = (Strike rate predicted by development factors)

(Of course, this algorithm could also be applied to Survival rate.)

It is important to notice that the “adjustment for Balls faced” only increases the affected statistic, in other words, a low number of balls faced will not reduce the Strike rate or Survival rate to which it is applied. The above algorithm adds experience as an additional benefit to the achieved strike rate but does not take anything away. This lack of experience only results in a smaller addition to the original rate achieved.

Hence, it allows the use of an estimated strike rate for those players with no experience, so that statistics projected from other formats can be used without having to be reduced in account of the lower number of balls faced*.

These estimates are obtained from other formats through the multiplication of the performance statistics by transition ratios called development factors** ⁷.

*If a low number of balls faced were to reduce the Strike rate or Survival rate, then the development factors would predict the expected Strike rate rather than a reduced Strike rate that should be used with the low number of balls faced.

**Development factors⁷ were calculated and applied to results in other formats in order to estimate a hypothetical strike rate that would most likely have been achieved had the player played more matches, (see pages 13-15 of the assignment)

Results from SAS:

For full SAS code and output related to the derivation of the credibility factors, see the “SAS appendix” for this assignment. This appendix was not printed but is available electronically.

(Refer to the glossary on pages 1 and 2 for an explanation of all the variable names that were used in the dataset)

8 A Performance Price Index

Now that two independent indices have been obtained which incorporate and summarise all the performance statistics of each batsman, the value and cost of these performances can be quantified. This will be done by measuring the aggregate change in price that accompanies a unit change in the relevant performance index, the dynamic of which is captured in the equations below:

$$\text{Unit cost of an increase in strike rate} = \frac{\Delta(\text{Price})}{\Delta(\text{SR Index})}$$

$$\text{Unit cost of an increase in survival rate} = \frac{\Delta(\text{Price})}{\Delta(\text{Survival Index})}$$

Thus providing the average cost of the performance gains that a particular player is expected to offer.

With this information, a performance price can be determined for each player based on the combinations of his respective performance indices.

For example, the performance price of player X is calculated as follows:

$$\begin{aligned} \text{Performance Price (X)} &= \left(\frac{\Delta(\text{Price})}{\Delta(\text{SR Index})} * \text{SR index (X)} \right) \\ &+ \left(\frac{\Delta(\text{Price})}{\Delta(\text{Survival Index})} * \text{Survival index (X)} \right) \end{aligned}$$

For full SAS code and output associated with the above derivations, see the “SAS appendix” for this assignment. This appendix was not printed but is available electronically.

There are several benefits to be gained through the use of the performance price index, as explained below:

1. Market related, yet reflective of changing value for money:

This performance price index can be advertised as being fully determined by the market, yet independent of changes in non-performance factors.

These performance prices will not be subject to criticisms of being derived merely by theoretical means, as they are based solely on the value that the market has placed on the respective measures of each player's performance index. It could be argued that the performance price index has been derived purely from market information as described by the relationship between the indices and prices above. This is because the market decides the price, while the performance price model simply measures the player's performance index and learns what price has been assigned to this level of a performance index, thus the

prices are purely decided by the market, yet the performance price is completely insensitive to changes in any player factors that do not form part of the three independent key factors (see section six). However, the average premium paid for all the reputation factors of all the players has been incorporated into the performance price, thus allowing a realistic reflection of the expected prices that is fully (decided by) and reflective of the current market, yet at the same time it fully reflects the changes in a player's value for money as actual bidding prices fluctuate with non-performance factors of certain players.

Therefore, if a particular player becomes over-priced due to an increase in his non-performance or reputation factors, this change in his marginal price per performance ratio will be reflected in the model because reputation factors will not change the performance price.

2. Budgeting performance requirements using performance costs:

The team can plan future player costs and budget performance needs according to the expected future price of these performances.

This can be done by assessing the relative increase in player prices with respect to an increase in the performance index of these players.

Thus the average price of gaining a greater run rate can be determined, or if a team needs more reliable batsmen, the required price of greater survival rates can be found by assessing the difference in player prices with respect to changes in the survival rate index.

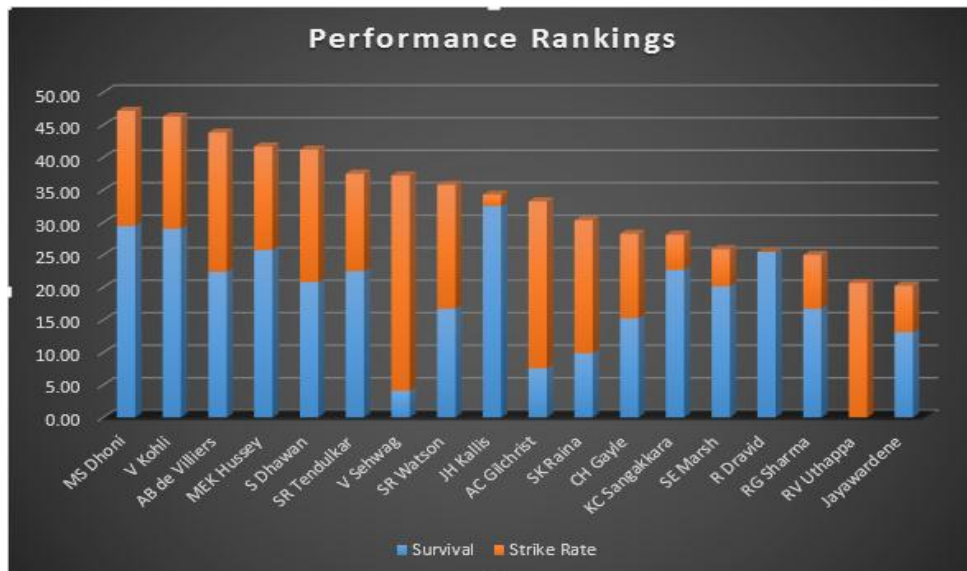
3. Provides an indication of which players or player factors are most likely to be over or under-priced:

The same approach can be used to compare the relative prices of key performance enhancements with the effect that the less relevant* player factors have had on player prices. This will indicate which players are most likely to be over or under priced with respect to their performance value based on what factors they

*Such as the reputation factors discussed in section six.

4. The performance price combines the indices into a single statistic that allows a simple ranking of players according to their total performance value

This can also be done using the "performance index", weighted according to team requirements. (See "A single performance index" on page 43)



The above graphs were generated from an example analysis of a few IPL “Big Names”. The accuracy of these results should not be very good as the analysis has only been applied to a small sample size, merely to illustrate the method used. A full and thorough analysis on a larger dataset will give much more accurate results.

These graphical results can be used to illustrate that AB de Villiers, Chris Gayle and Rohit Sharma are under-priced from a performance perspective, whereas Suresh Raina, Adam Gilchrist and Jacques Kallis are over-priced.

In summary:

1. The Performance Index provides:

- A ranking of players in terms of net performance value.
- A benchmark of the fair price of each player.

Actual price – Performance price = Reputation premium.

The Reputation premium of a player needs to be compared to the additional revenues that he is estimated to bring to the franchise through media publicity.

What the Performance price index does not provide:

The performance price index does not capture the changing needs of different IPL teams. What constitutes a good deal for one team may not be so for another team.

Therefore, team dynamics and team balance need to be measured in addition to the performance prices of each player.

9 Team Dynamics

A major factor to consider in pricing economics is the different relative value of players for different teams due to the changing needs of each franchise. Based on the principles of supply and demand, you would expect the price of a player to be equal to the “marginal price”, which is the maximum price that the “marginal consumer” is willing to pay for a given player.¹⁰ The marginal consumer is a buyer that is indifferent between either buying or not buying a particular player at a given price determined by global supply and demand forces. Thus he is the equilibrium player, who is “in the middle” in terms of being the median in the sample of all buyers listed in order of the price that they are willing to pay for a particular player, which is of course determined by the changing needs of each team. In an efficient market, bidders are assumed to all have “perfect information”, in other words, there is no advantage for some franchises over others in terms of knowing what the others don’t know about the respective values of the different players on the market.¹⁰ In such a market, players would be expected to be priced correctly according to the needs and demands of the franchises, and prices should be equal to the marginal price.*

All those who are not marginal buyers will know that the entity (or cricket player) is either under-priced or over-priced for their specific needs, even if this player is bought at the performance price. Thus franchises would buy and sell players according to their changing needs while prices should remain equal to the marginal price of the player (which is the value of the player to the marginal buyer).

This means that the price of a player will always be greater or less than their value to some teams even if they are priced correctly. Neither the marginal price nor the performance price will accurately reflect this value.

Therefore a need arises to develop a mechanism of selecting players based on their intrinsic value to the team, which depends on the nature of the players that the team already has at their disposal.

For example, if there is an excess of players with a higher than average strike rate, but a shortage of those batsmen with the ability to survive for a long enough time before going out, you may find that this imbalance in team selection results in too many wickets falling without anyone playing an “anchor” role. Of course, the reverse could be true and a team could select too many “survivors” and not enough “big hitters”, which would result in a team frequently reaching their full quota of twenty overs with only two or three wickets down, this is also not ideal as the team would end up wasting money on big strikers that never get a chance to bat. Albie Morkel could be considered as an example of this scenario, having one of the best strike rates (in excess of 160), yet often only comes in to bat when there are two or three balls remaining. A team should seek to maximise the use of their full arsenal of batting potential.

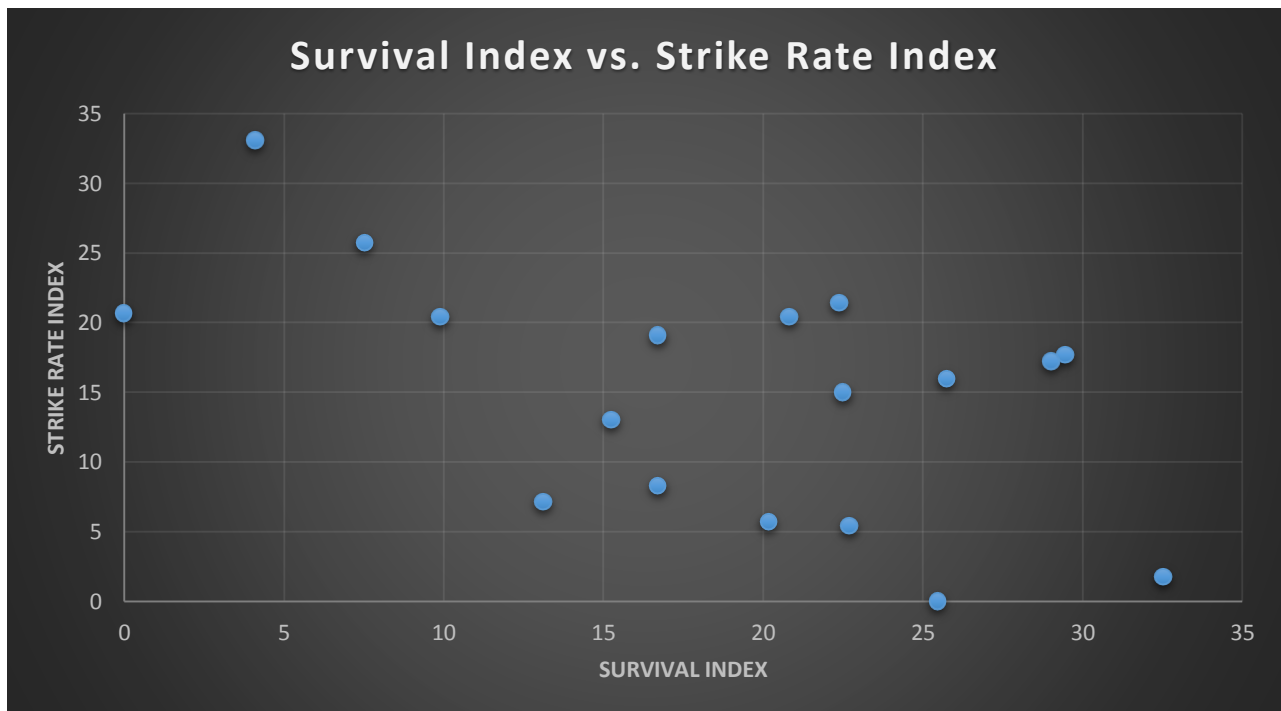
* An efficient market will not simply price all players with the performance price index, as this reflects the value that the market assigns to the performance value of a player, rather than reflecting the premium that the marginal buyer is willing to pay for other factors such as a player’s icon status.

Getting this balance right could be referred to as “team dynamics” which need to be accounted for in the auction bidding process. It’s no good to just go for the most under-priced players according to the performance price index, and then end up with five “survivor” batsmen like Kallis or Dravid.

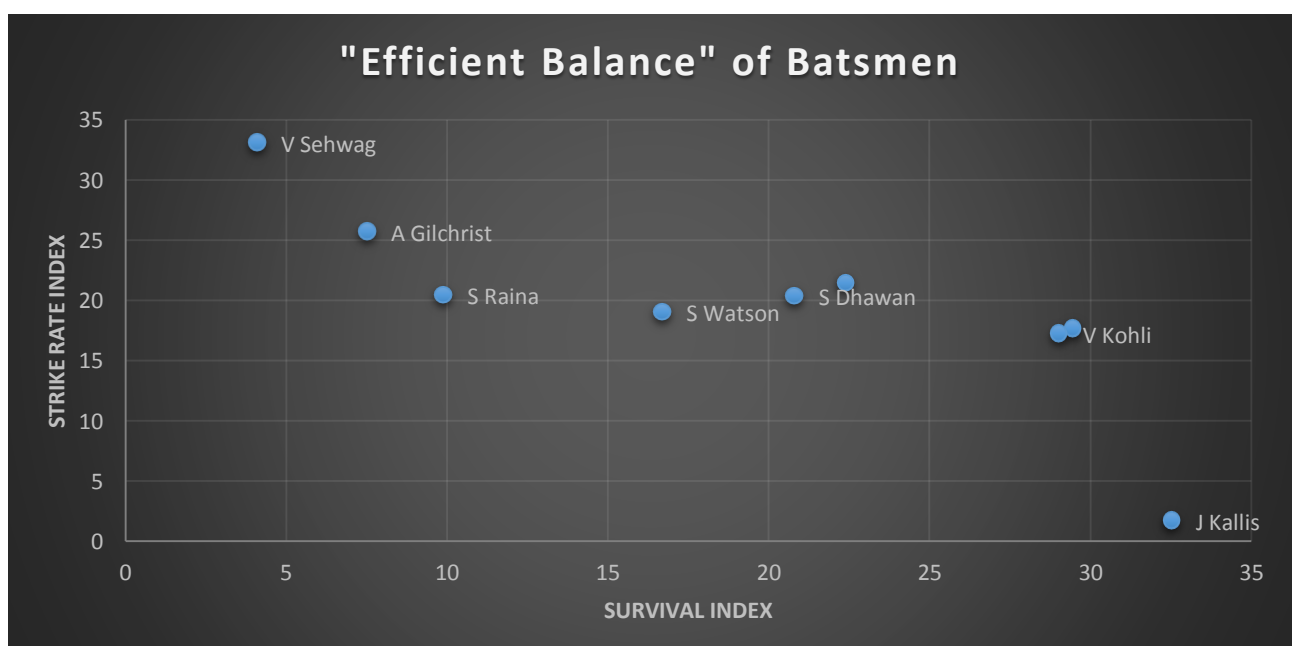
In order to simultaneously select players that are both balanced in attributes and “efficient” in their trade-off between performance and price, the concept of an efficient frontier from Markowitz’ Mean variance portfolio theory¹⁰ could also be applied to cricket player attributes.

“Efficient” players are those that maximise their survival index for a given strike rate index. This is measured relative to all other players that have achieved a similar strike rate index.

Efficient players can be selected graphically as illustrated below:



The “efficient balance” of batsmen has been selected from the above scatter plot

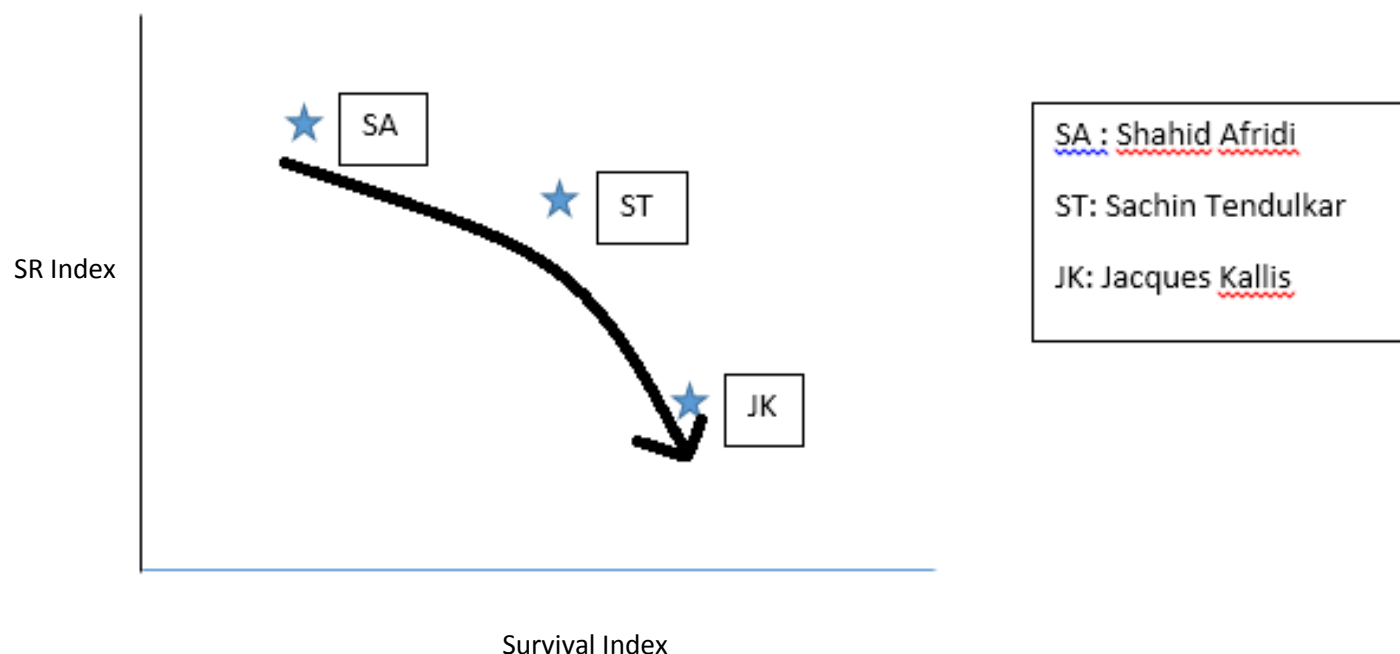


For each player represented in the above scatter plot, the minimum value of either set of indices were subtracted from the players' strike rate and survival indices respectively. This was done to improve graphical representation, and results in the batsman with the lowest value in the index being represented by a "0".

A balanced team dynamic would involve a diagonal line representing an inverse parabola graph such as $\{y = c - mx^2\}$

Where: y = Strike rate index

x = Survival index

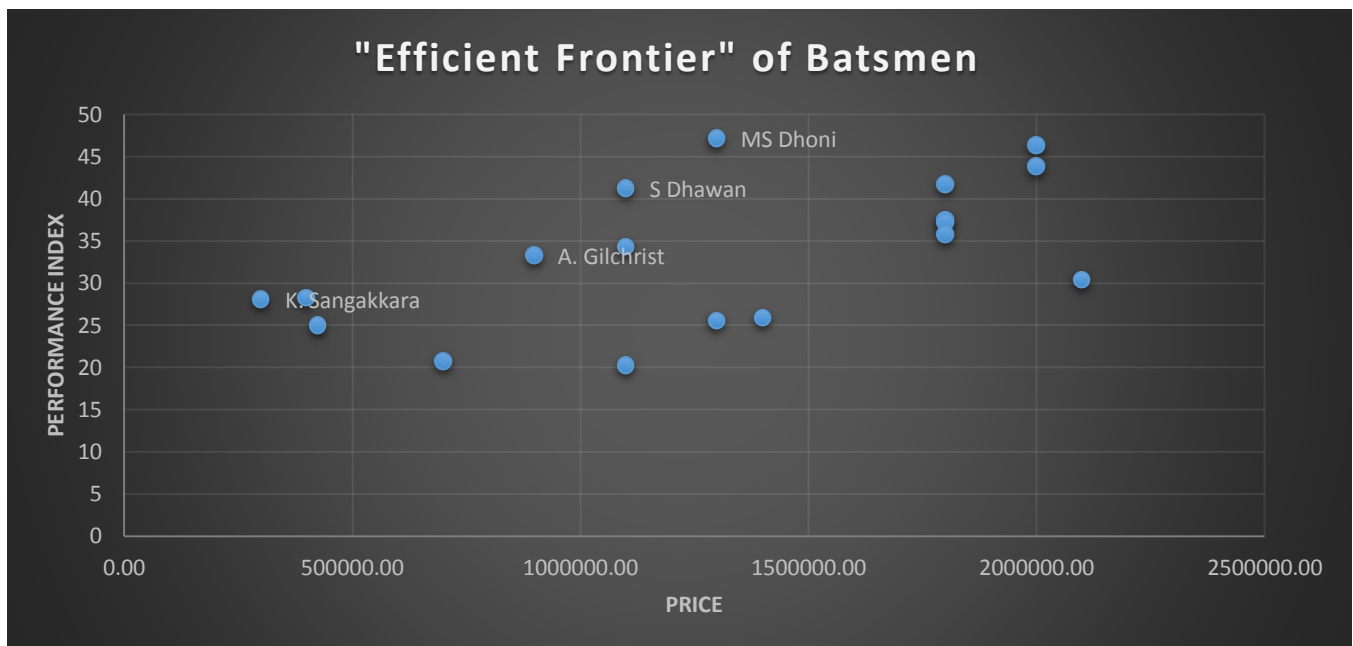


Ideal figure of "Team Balance"

It is important to notice that the intention here is not to plot the survival rate against the strike rate, but rather their respective indices as derived in section 7.

The Efficient Frontier of Batsmen:

Another method of graphical comparison between different players is to plot the total performance index of the sample of available batsmen against their prices, which then can be used to derive an "efficient frontier" of batsmen. As mentioned above, the method of using an efficient frontier has been adapted from the original model developed by Markowitz. If applied to cricketers (equities), the efficient frontier would be the selection of players (shares) where the price (risk) has been minimised with respect to all available options for a given level of performance (returns).¹⁰



The labelled players are those that lie on the efficient frontier

A Single Performance Index:

The “performance index” in the y-axis of the above scatter plot would be a combination of the strike rate and survival indices. This combination could be a simple addition of the two indices, but it would be better if team balance requirements could be incorporated into this performance index by assigning different weightings to the strike rate and survival indices in proportion to the priorities of the team.

General Strategy:

The process of selecting players for purchase can thus be summarised as follows:

1. Plot the entire batting line-up on the team balance graph.
2. Use the team balance graph to identify required performance factors from any imbalances in the line-up.
3. Plot the sample of available players for auction against each other on another balance graph. (Survival vs. Strike index)
4. Select players from this graph that best match the team’s required performance attributes
4. Plot the selected players against the efficient frontier of all available batsmen to assess value for money.
5. If value for money is poor, repeat the above process using other players that are closer to the efficient frontier curve.

10 Modelling the distribution of scores

10.1 Preventing Batting collapses:

In addition to modelling performances in order to find pricing anomalies, a franchise may want to assess the variations in player scores for purposes of team balance. For example, a batting line-up combines several different batsmen together in the top order. If all of these players have a high variance in individual scores, there would be an increased chance of a combined failure, leading to what is known as a “batting collapse”. An improved balance of batsmen with different levels of score variances may be desired in order to reduce the probability of this scenario. It is difficult, however, to quantify what the combined effect would be on the team total if several batsmen of different levels of score volatility would be playing together. Practical problems such as these are not always amenable to analytical solutions. As such, the simultaneous modelling of several batsmen’s scores in accordance with their respective score distributions can be performed using Monte Carlo simulation.

Advantages of modelling team scores:

- Capture the behaviour of different combinations of batsmen and bowlers.
- Determine optimal team balance with regards to strike rate and survivability.
- Determine optimal team balance with regards to frugal economy rate and wicket rate. (bowlers)
- Find a balance between batsman with a high variance of scores and those with a more steady performance.
- Find a balance between bowlers with volatile performances and the steady performers.
- Assess the volatility and of team totals and team performances.

10.2 Modelling Individual Scores:

In order to determine whether a team has obtained a desirable balance of different temperaments in batsmen, it may be useful to be able to stochastically model the projected future scores of each individual batsman in a team independently before assessing the combined effect on team total.

For this purpose each player has to be analysed for their individual score distribution. The only way to make this possible would be to collect a history of individual scores by each batsman for every innings that he has played. These scores can then be summarised and combined into a list of scores per player.

Below is an example of such data being summarised in excel from an “innings by innings” history of the Twenty-twenty internationals.

(This was only done for a small sample of batsmen as a preliminary example)

F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Innings	MS Dhoni	V Kohli	AB de Villiers	MEK Hussey	V Sehwag	SR Watson	JH Kallis	AC Gilchrist	SK Raina	CH Gayle	KC Sangakkari	SE Marsh	RG Sharma	RV Uthappa	M Jagadev
1	0	28	0	37	34	4	20	15	3	19	21	29	90	10	0
2	33	28	6	1	5	17	15	15	35	5	13	9	8	0	2
3	24	14	16	18	40	33	4	1	61	61	30	15	30	6	9
4	10	4	14	15	58	0	4	48	0	117	14	4	9	15	65
5	45	15	18	37	11	22	48	4	10	0	18	0	4	24	35
6	36	22	1	13	9	10	24	45	5	67	20	25	7	8	28
7	6	31	1	22	5	9	67	43	2	1	22	26	36	35	30
8	9	68	52	7	0	37	45	24	3	88	55	0	52	1	1
9	9	70	0	53	1	62	64	31	21	22	5	6	5	1	0
10	13	50	25	0	26	19	7	22	9	5	15	47	9	18	35
11	2	40	36	23	24	81	73	12	18	15	3	21	29		5
12	28	15	0	1	26	4	34	1	101	63	35	28	3		9
13	26	78	7	17	64	54	31	25	5	5	0		79		4
14	14	2	0	47	4	1	11		32	12	64	13			19
15	11	21	79	8	23	5	22		63	25	38		10		78
16	30	38	15	39	8	16	53		28	5	13		0		41
17	5	9	11	60	29	2	61		72	98	68		53		2
18	9	27	17	17	17	0	48		41	4	78		26		1
19	46	29	63	18		4	13		2	14	59		1		12
20	15		1	25		58	6		33	0	4		0		3
21	16		10	59		17	12		39	2	3		4		41
22	2		24	14		10	6		14	95	68		1		9
23	29		21	1		57	6		34	53	2		55		12
24	23		17	23		52			27	54	46		1		81
25	10		47	12		69			38	58	16		25		100
26	8		5	10		5			1	2	17		24		90
27	21		53	28		8			26	30	5		2		9
28	48		11	45		33			45	75	44		4		4
29	21		54			19			26	43	3		8		10
30	16		0	18		51			36	6	30		0		0
31	22		11			41			10	8	24		0		17
32	18		14			72			1	5	9		1		24
33	9		8			70			19	1	19		0		72
34	15		39			8					8		0		11
35	23		29			7					44		0		86
36	24		10			37					12		2		2
37	38		1			7					21		0		26
38	1		2			6					39		0		13
39	33		30								13				4
40	24		25								18				44
41			21								22				65
42			13								59				42
43			36								38				42
44			15												33
45			15												8
46			1												61
47															6
48															33

Although this data set only contains a history of IPL and T20I scores per batsman, the same analysis and methods could be applied to any format of the game. However, for the sake of time and relevance, the score simulations will only be based on this data.

The sets of scores for each batsman were then analysed to obtain the average, variance, skewness and kurtosis levels of the distribution of scores.

The results can be compared between the several players in the appendix.

10.3 Adjustments needed:

S. Dhawan, R. Dravid and S.R. Tendulkar will need to be removed from the data set as they share only four T20I matches amongst themselves. The small sample size for these players may produce statistical outliers which skew the results. For example, the record of Sachin Tendulkar, whom having played only a single T20I innings, would return a variance in scores of 0 if included, which provides an unrealistic depiction of what would have happened had he played more of these matches.

10.4 Variations in sample size:

It is notable that when measuring the likes of kurtosis and skewness, some of the variations in the sample sizes of the number of matches played by different players are quite significant and may influence the results.

For measurement of skewness there are several formulas available, however, given the variation of sample sizes, it would be more appropriate to adopt a measure which includes an adjustment for each sample size, such as the formula below:

Source: David P. Doane and Lori E. Seward (2011), "Measuring Skewness: A Forgotten Statistic?", Journal of Statistics Education, Volume 19, pp 7-8

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \left[\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}} \right].$$

According to Doane and Seward (2011) the formula denoted as G1 above would be the most simplistic measure of skewness to employ in a scenario where there are variations and limitations in the sample size.¹³ It is mentioned that after comparing the biases and mean squared error (MSE) of different measures of skewness, Joanes and Gill (1998) found that the above formula performed very well by obtaining a small MSE from variously skewed populations of varying sample sizes, which makes it the most appropriate measure for this purpose.¹⁴

Fortunately however, this formula is equivalent to the adjusted Fisher-Pearson standardised moment coefficient employed by both SAS and Excel,¹³ which indicates that under the use of these software packages, the results would not be biased when comparing the measurements of skewness taken from different players with varying sample sizes of matches played.

10.5 Incorporating Kurtosis into the model:

Horswell and Looney (1993) also noted that “The performance of skewness tests is shown to be very sensitive to the kurtosis of the underlying distribution.”¹⁵ In which case it is noted that most batsmen displayed a more leptokurtic (positive kurtosis) distribution of scores, and hence the measure of skewness will be expected to behave differently for those players with an unusually platycurtic score distribution.

Due to the large variation in kurtosis levels between each batsman, a simulation of the team total will produce the most realistic results if the different batsmen's scores are generated from a model of stochastic distributions, in other words, the random variables for each score would come from a variety of distributions, each possessing a new variation in skewness and kurtosis levels. However, it may be necessary for practical purposes to simulate each score with a single, constant distribution shape, while only the mean and variance are changed to fit the different records of each player.

This constant distribution shape will need to reflect the average levels of skewness and kurtosis across the different players' score distributions. Given that the majority of the batsmen produced a leptokurtic distribution of scores that were skewed to the right, an extreme value distribution¹⁷ possessing large tails may be the most appropriate for this purpose.

Neves C. and Alves I.F. (2010) point out that these distributions are far more accurate than normal statistical distributions when modelling and measuring events which occur with a small probability. It is therefore useful in simulating the scores of batsmen and bowlers that have a particularly volatile distribution of results. There are three types¹⁷ of extreme value distributions:

(i) *Gumbel (type I)*: $\Lambda(x) = \exp\{-\exp(-x)\}$, $x \in \mathbb{R}$;

(ii) *Fréchet (type II)*: $\Phi_{\alpha}(x) = \begin{cases} 0, & x \leq 0; \\ \exp\{-x^{-\alpha}\}, & x > 0, \alpha > 0; \end{cases}$

(iii) *Weibull (type III)*: $\Psi_{\alpha}(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & x \leq 0, \alpha > 0; \\ 1, & x > 0. \end{cases}$

Source :

Alves I. F. and Neves C. (2010) “Extreme Value Distributions”, Faculty of Sciences, University of Lisbon, pp 1-2

Due to its flexibility¹⁷, the Weibull distribution will be chosen to model the batsmen’s scores. The distribution’s pliability can be illustrated as follows:

If $\alpha = 1$, the Weibull distribution reduces to the Exponential model

If $\alpha = 2$, it becomes equivalent to the Reyleigh distribution.

If $\alpha = 3.5$, it resembles the Normal distribution.¹⁷

10.6 Monte Carlo Simulation:

The performances of the entire batting line-up can thus be modelled using Monte Carlo simulation of variables that contain a Weibull distribution. Each individual batsman’s score can be simulated by generating a Weibull random variable with the necessary parameters that will cause the variable to match the mean and variance of each batsman’s individual score distribution. This can be done by means of the inverse transform method.

Applying the Inverse transform method:

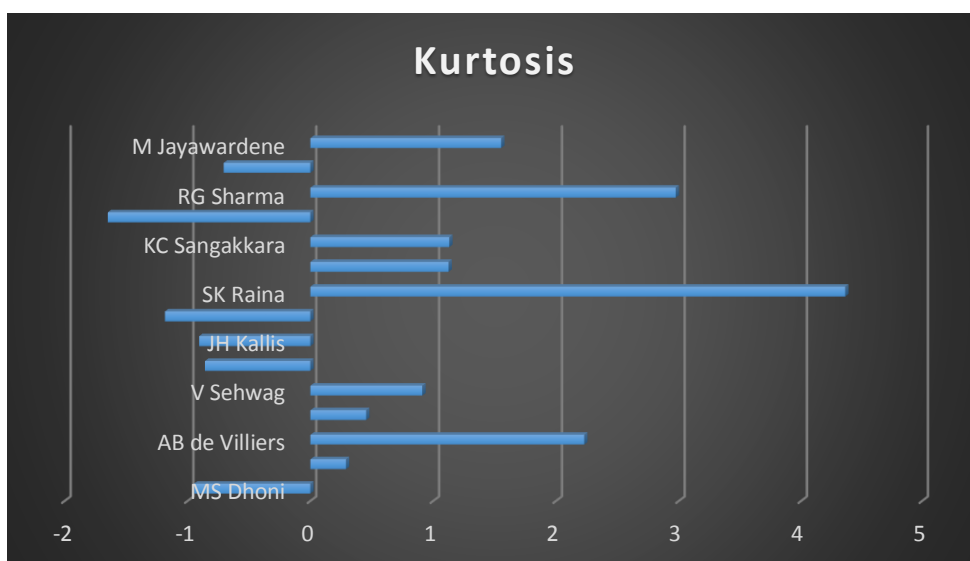
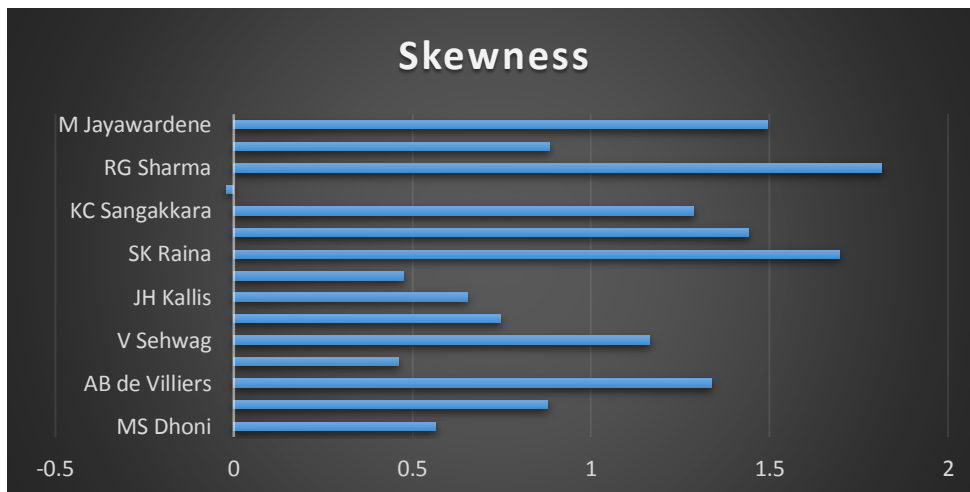
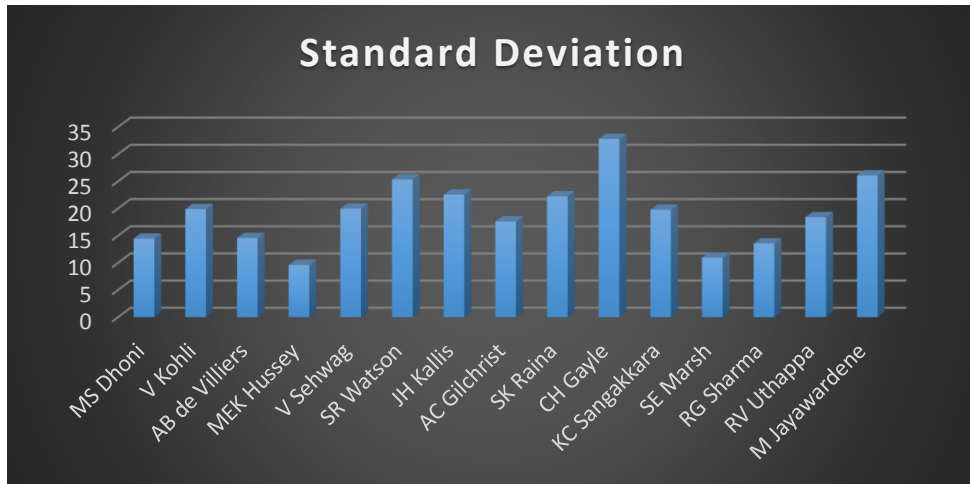
This method can be applied to simulate each batsman’s score by making use of a random number generator (provided by SAS) which can be used to generate a random variable with a “uniform (0,1)” distribution. This variable can be multiplied by the inverse of the Weibull cumulative density function in order to obtain a random Weibull distribution.

The parameters of the inverse cumulative distribution function will be chosen in order to generate a Weibull variable that matches the mean and variance of the batsman’s individual score distribution.

The continuous random Weibull variable can then be rounded to the nearest integer in order to simulate the discrete number of runs for the individual batsman.

This process can be repeated for the eleven batsmen in the batting line-up in order to simulate a team performance.

Appendix to Section 10:



Glossary

Cricket Terms:

The terms “Balls faced”, “Strike rate” and “Survival rate” will be used with capital letters to refer to these unique terms. i.e. “Balls faced” will be mentioned in place of “the number of balls faced” anywhere in a sentence.

Dot ball: A ball bowled for which no run was scored.

Five-for: A five wicket haul in a single innings.

Ten-for: A ten wicket haul collected across the two innings’ of a Test match.

Cricket Format References:

Format: When this word is used on its own, it will typically be referring to the “type of cricket that is being played” i.e. Test, ODI or T20 cricket.

Test: Used as a single word with a capital letter referring to an international Test match.

ODI: One Day International match

T20I: International Twenty - twenty match

T20: Any Twenty-twenty match, including IPL matches

IPL: Indian Premier League matches

Glossary of column names in the Database:

II : Indian Idol

RA : Renowned Aussie

GIS : Global Iconic Status

MIP : Mid-way Indian player (Future in National Team)

SBR : Slow Batsman Reputation

SR: Strike rate

= (runs scored / balls faced) * 100

Ave: Average referring to either batsmen or bowlers

= (runs scored / Wickets or dismissals)

WR: Wicket rate, also known as Strike rate for bowlers

= (Balls bowled / Wickets taken)

Runs: the number of runs score in total

BF: Balls faced

100: refers to number of centuries scored
50: refers to number of fifties scored
4w: refers to number of four-wicket hauls in a single innings
5w: refers to number of five-wicket hauls in a single innings
St: refers to number of stumpings
Ct: refers to number of catches

Statistical terms:

Leptokurtic:

Having positive kurtosis.

Leptokurtic distributions have higher peaks around the mean compared to normal distributions, as well as thicker tails on both extremes. The high peak is a result of the data being highly concentrated around the mean, whereas the thicker tails represent a more frequent occurrence of extreme values.

Platykurtic:

Having a negative kurtosis.

A platykurtic distribution appears flatter than a normal distribution, due to a more even spread of frequencies around the mean.

Heteroskedastic:

A sample of statistics is heteroskedastic if the variances are changing for different subsets of the sample. These results can cause errors in regression analysis.

References

1. Mathur A. (2013), "IPL-onomics"
2. Depken, Rajasekhar (2010) "Open market Valuation of Player Performance in Cricket: Evidence from the Indian Premier League", Social Science Research Network, pp 1, 6-8, 10-11
3. Parker, Burns, Natarajan (2010) "Player valuations in the Indian Premier league" Frontier Economics, pp 3-4, 7-8
4. E. van den Berg (2011), "The Valuation of Human Capital in the Football Player Transfer Market", Erasmus School of Economics, pp 10 -16
5. Ridder G. (2005), "Undergraduate econometrics", notes, lecture 2, USC Research Computing Facility
6. Joseph J. Adamski, Philip J. Pratt (2012) "Database Management Concepts" 7th edition, Chap 4, pp130-133
7. Chap 10 Actuarial notes summary, "The Teaching Network", Shanghai University of Finance, pp 8-12
8. Chap 11 Actuarial notes summary, "The Teaching Network", Shanghai University of Finance, pp 1 – 10
9. UCLA: Statistical Consulting Group (2013), "SAS Annotated Output: GLM" http://www.ats.ucla.edu/stat/sas/output/sas_glm_output.htm (accessed November 2013).
10. Luenberger D.G. (2009), "Investment Science", Oxford University Press, pp 155 – 159
11. Chap 10, CT4 core reading, "Binomial and Poisson models", The Actuarial Education Company, pp 5-9
12. Jay Rotella (2013), "Analysis of Population & Habitat Data, Binomial likelihood", Montana State University, pp 1-5
13. David P. Doane and Lori E. Seward (2011), "Measuring Skewness: A Forgotten Statistic?", Journal of Statistics Education, Volume 19, pp 7-8
14. Joanes, D. N. and Gill, C. A. (1998), "Comparing Measures of Sample Skewness and Kurtosis," The Statistician, 47, Part 1, pp. 183-189.
15. Horswell, R. L. and Looney, S. W. (1993), "Diagnostic Limitations of Skewness Coefficients in Assessing Departures from Univariate and Multivariate Normality," Communications in Statistics: Simulation and Computation, volume 22, pp 437

16. Chap 5, CT6 core reading, "Credibility factors", The Actuarial Education Company, pp 1-10
17. Alves I. F. and Neves C. (2010) "Extreme Value Distributions", Faculty of Sciences, University of Lisbon, pp 1-2