Analyzing Factors Contributing to the Success of a Team in Dota 2

JOHN ANDERSSON

OSKAR ALIN
Analys av faktorer som bidrar till ett Dota 2 lags framgång

JOHN ANDERSSON

OSKAR ALIN
Abstract

Competitive gaming or E-Sport is more popular than ever, this has resulted in an increase in the number of players and tournament prize pools. In traditional sports demographic factors have been shown to have high predictive power when it comes to determining a country’s success in the Olympic Games. Similar results have been found when it comes to E-sport which is why it is interesting to investigate whether there are any differences between regions in Dota 2 as well. The goal is to analyze factors that contribute to the success of a Dota 2 team by building a multiple-regression model. All data is collected from open sources and contains 55 active Dota 2 teams that have been playing between 2011 - 2018. The factors in the final model is the sum of the individual players estimated skill, the skill difference between the highest and lowest rated players on the team, the number of games the team has played and organization region. The result gives an insight in what a person or organization would want to look at when researching a team as well as a model that can be used to predict how good a team will perform.
Sammanfattning

# Contents

1 Introduction  
  1.1 Background .................................................. 1  
  1.2 Purpose ......................................................... 1  
  1.3 Research Question .............................................. 2  

2 Theory  
  2.1 Dota 2 .............................................................. 3  
  2.2 Elo ................................................................. 5  
  2.3 Previous Performance Research  
      2.3.1 Project Aristotle ............................................. 6  
      2.3.2 The Olympic Games ......................................... 8  
      2.3.3 E-Sport ....................................................... 9  
  2.4 Multiple Linear Regression  
      2.4.1 Regression Model ............................................ 9  
      2.4.2 Ordinary Least Squares ................................... 10  
      2.4.3 Assumptions .................................................. 10  
      2.4.4 Residuals ..................................................... 11  
  2.5 Hypothesis Test  
      2.5.1 Test for Significance of Regression ..................... 11  
      2.5.2 Test on Individual Regression Coefficients .......... 12  
  2.6 Model Selection  
      2.6.1 All Possible Regression .................................. 12  
      2.6.2 Akaike Information Criterion ......................... 13  
      2.6.3 Mallow’s C_p ............................................... 13  
  2.7 Model Errors ..................................................... 13  
  2.8 Model Validation  
      2.8.1 $R^2$ and $R^2_{\text{Adj}}$ .............................. 15  
      2.8.2 Cook’s Distance ........................................... 15  
      2.8.3 Variance Inflation Factor .............................. 16  
      2.8.4 Mean Square Error (MSE) .............................. 16  
      2.8.5 Leave-One-Out Cross-Validation ...................... 17  

3 Method  
  3.1 Demarcation ...................................................... 18  
  3.2 Variable Selection ............................................. 18  
  3.2.1 Response Variable .......................................... 18  
  3.2.2 Explanatory Variables (Naive Model) .................... 18  
  3.2.3 Explanatory Variables (Extended Model) .............. 19  
  3.3 Data Collection .................................................. 19  
  3.4 Model Selection .................................................. 20
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4</strong> Results</td>
<td>21</td>
</tr>
<tr>
<td>4.1 Naive Model</td>
<td>21</td>
</tr>
<tr>
<td>4.1.1 Validating Naive Initial Model</td>
<td>21</td>
</tr>
<tr>
<td>4.2 Naive Reduced Model</td>
<td>24</td>
</tr>
<tr>
<td>4.2.1 Validating Naive Reduced Model</td>
<td>24</td>
</tr>
<tr>
<td>4.3 Extended Initial Model</td>
<td>27</td>
</tr>
<tr>
<td>4.3.1 Validating Extended Initial Model</td>
<td>27</td>
</tr>
<tr>
<td>4.4 Extended Reduced Model</td>
<td>30</td>
</tr>
<tr>
<td>4.5 Validating Extended Reduced Model</td>
<td>31</td>
</tr>
<tr>
<td>4.6 Model Comparison and Analysis</td>
<td>34</td>
</tr>
<tr>
<td>5 Discussion</td>
<td>35</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>37</td>
</tr>
<tr>
<td>References</td>
<td>38</td>
</tr>
<tr>
<td>A Appendix</td>
<td>40</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Background

Competitive gaming known as Electronic sports (E-sport) has had a large increase in popularity since its creation in 1972. At that time around two dozen people gathered at Stanford university to compete in the game Spacewar! with a first prize consisting of a 1 year subscription to Rolling Stone magazine [4]. Since then E-sport has gone from something most people would only consider a hobby into a competitive sport with professional actors. It is predicted that the industry as a whole will have a revenue of 906 million dollars in 2018 and the estimated number of viewers will be 380 million [14]. E-sport is still young comparatively to traditional sports, and it has a lower barrier of entry to start a new team compared to starting a new team in a more established sport such as hockey or football.

Currently the largest game in E-sport seen from an economics perspective is Dota 2. On the list of the top 100 E-sport athletes by total prize earnings 77 are Dota 2 players [6]. Since the release of Dota 2 in 2011 its tournaments has paid out more then any other E-sport. By 2018 the total amount of prize money payed out from all tournaments combined exceeded 138 million dollars. This can be compared to the number two and three games Counter-Strike: Global Offensive and League of Legends whos tournaments have payed out approximately 52 million dollars each [7].

1.2 Purpose

The purpose is to get a better understanding of the underlying factors that determine a E-sport teams performance. This could be useful when it comes to building a team from the ground or in the event of making lineup changes in a existing team. This study will look at E-sport teams, specifically in Dota 2, but a study of this nature could be adapted to other team or group sports in a competitive environment such as two versus two tennis games. The project aims to analyze different factors that impacts the success and overall skill of a team. The Elo rating system will be used as a measure of the teams performance. The average Elo ratings of the teams is going to be the response variable in the regression models. The teams will be studied from two points of view and both group factors as well as individual factors will be investigated.
1.3 Research Question

The goal is to create a linear model that can estimate the skill level of a team. Two different models will be tested to find a satisfactory model.

**The Naive Model:** A model that only includes the individual players estimated skill and their position in the Dota 2 team.

**The Extended Model:** A model that includes the combined estimated skill of all the players but also takes into account the skill difference, the players nationalities, how many players have represented the team, the region and age of the team organization and the number of games played.
2 Theory

2.1 Dota 2

Dota 2 has a active competitive scene with players from all over the world competing in tournaments and leagues of different sizes. The major Dota 2 tournaments have the largest prize pools in all of E-sport and the largest so far was The International 2017 which had a prize pool of 24.8 million dollars [20]. Professional Dota 2 matches are streamed live on websites such as youtube and twitch.tv and the viewers can reach into the millions.

Dota 2 was released in beta form in 2011 and it is a free to play multiplayer online battle arena (MOBA) computer game developed by Valve corporation, an American game developer and digital distribution company. It is a sequel to the popular Defense of the Ancients (Dota) which was a community created mod for Blizzard's 2002 game Warcraft III [8].

In a game of Dota 2 two teams with five players each face each other with the objective to destroy the opponent teams “Ancient”, which is the central building in the teams base on the opposite corner of the playing field. Generally in professional Dota 2 a match between two teams is played as a best out of three games. In large tournament finals best out of 5 games is often the preferred format.

![Dota 2 playing field](image)

Figure 1: Dota 2 playing field

The playing field is divided into two areas. The lower left area is initially controlled by the team starting in the green base and the upper right area to the team that is starting in the red base. To accomplish the goal of destroying the opposite teams Ancient, the teams need to collect resources (experience points and gold) by defeating computer generated opponents called creeps. Creeps are automatically generated in the teams home base every thirty seconds and walks a fixed path, called lane, towards the enemy base. The lanes are shown as dashed lines in figure 1.
Each player controls a character that is referred to as a hero. As of 2018-04-12 there are 115 different heroes in the game all with unique abilities and play styles. Each hero starts at level 1 and has a maximum level of 25. Experience points allows them to level up and increase the power of the heroes abilities and gold is used to purchase items that increase the strength the hero. The heroes can be divided into two primary categories: core heroes and support heroes. Core heroes main purpose is to inflict as much damage as possible to the opposite team players and their structures. Core heroes generally start weak but scale up as they gather more gold and experience. Support heroes focus on assisting the core hero to increase in strength and to gather intelligence about the opposite team, for example monitor enemy movement on the map. Support heroes often does not need much gold or experience to be useful.

Just like many other team oriented sports Dota 2 players have different positions available to them. In soccer the players have different positions such as goalkeeper, forward and midfielder. In Dota 2 the positions are often referred to by a number which indicated that positions farm priority. Position 1-3 are often refereed to as core positions while 4 and 5 are the support heroes. Dota 2 doesn’t have any fixed positions and can be played many different ways, but a common lineup is detailed below:

1. Hard Carry: The hard carry is often focused on gaining resources, resulting in a hero that can inflict large amounts of damage as the game goes on. This position is often supported by the number 4 and 5 players early to ensure a strong late game.

2. Solo Mid-laner: The solo mid-laner is also focused on gaining gold and experience, but usually not to the extent of the hard carry. The mid-laner levels up faster than the hard carry (because the mid laner is alone and not sharing the lanes resources with the support) and is usually the main damage dealer until the hard carry is ready.

3. Solo Offlaner: The offlaners role early is to harass the enemy team’s hard carry and try to get as much resources as possible.

4. Roamer/Jungler/Secondary-Support: This position is flexible but usually it revolves around roaming the map and helping out whichever lane needs assistance.

5. Hard Support: Early in the game the main goal of the hard support is to help out the team’s hard carry. The support is also the one that provides vision of the map to his team buy purchasing and placing an in game item called wards.
2.2 Elo

The Elo rating system, named after its creator Arpad Elo, is a framework for comparing the relative skill between different players in zero-sum games. It was originally developed in order to rank chess players but it has since been implemented as a rating system for numerous other sports and games. In this system a player will be assigned a numerical rating based on his or her performance against other players. Winning matches increases a players rating and losses leads to a reduction in rating. In a match where one player has a higher rating than his opponent the rating system predicts that higher rated player is more likely to win. The larger this rating difference is the more favored the higher ranked player is in the matchup. This rating difference also dictates how many points the winner takes from the loser. If the difference is large and the higher rated player wins he will not gain many of his opponents points, but in the case of an upset the lower ranked player can gain many of his opponents points.

A central assumption in this model is that every players performance can be described as a normally distributed random variable whose mean represents that player’s true skill. For simplification reasons all players are assumed to have the same standard deviation for their performance random variable [9].

When analyzing paired comparison data, it makes little difference whether one assumes the logistic distribution or the normal distribution [18]. For this reason the logistic distribution is most often used when comparing the skill of two players since it is simpler. Using this logistic distribution and two players with rating \( R_{\text{Player1}} \) and \( R_{\text{Player2}} \) respectively the following expressions shows their expected score:

\[
E_{\text{Player1}} = \frac{1}{1 + 10^{(R_{\text{Player2}} - R_{\text{Player1}})/400}}
\]

\[
E_{\text{Player2}} = \frac{1}{1 + 10^{(R_{\text{Player1}} - R_{\text{Player2}})/400}}
\]

After a game has been played the players rating can be updated, in the case of Player1 the following formula shows the change in rating:

\[
R'_{\text{Player1}} = R_{\text{Player1}} + K(S_{\text{Player1}} - E_{\text{Player1}})
\]

Where \( R'_{\text{Player1}} \) is the new rating for Player1, \( S_{\text{Player1}} \) is the score of the player achieved in the previous game and K is the K-factor, a factor that determines the maximum possible adjustment in rating per game. This factor is often set at \( K = 16 \) for masters and \( K = 32 \) for newer players.

2.3 Previous Performance Research

Group performance and the factors that contribute to the success of a team is an area of research where many studies have been done. It is not only relevant when it comes to a group of athletes, businesses are often segmented and employees
work in smaller groups, students have study groups and researchers work in teams in order to make important findings. This section will look at the results from previous studies.

2.3.1 Project Aristotle

In 2012 Google performed a large study called Project Aristotle in order to learn what factors contribute to the success of a team. They were hoping to identify a perfect mix of individual traits and skills necessary to allow a team to perform at a high level. Over two years they conducted over 200 interviews with their employees and examined more than 250 attributes in more than 180 active Google teams. To their surprise they found that who is on the team matters less than how the team members interact with each other, how they structure their work and view their contributions. They managed to identify five key dynamics that set the successful teams apart from other teams at Google. The researchers also found that individual performance of team members were not significantly connected with team effectiveness. The important dynamics or factors are as follows in order of importance:

1. Psychological safety: Can we take risks on this team without feeling insecure or embarrassed?
2. Dependability: Can we count on each other to do high quality work on time?
3. Structure and clarity: Are goals, roles, and execution plans on our team clear?
4. Meaning of work: Are we working on something that is personally important for each of us?
5. Impact of work: Do we fundamentally believe that the work we’re doing matters?

**Psychological Safety** Psychological safety is a part of team culture or corporate culture where the individual members of the group feel like they can take risks and speak their mind without being seen as ignorant, incompetent, negative, or disruptive. In a team that has a high level of psychological safety, teammates feel safe to take risks around their team members. They feel confident that no one on the team will embarrass or punish anyone else for admitting a mistake, asking a question, or offering a new idea [10].

This aspect of corporate culture can have large affects on the performance of a team. A study of the February 1, 2003 Columbia space shuttle disaster found that Engineers had noticed that the shuttle was damaged during routine reviews of videos taken at the launch. However the camera angle was not the best to assess the damage and managers downplayed the threat, noting that foam strikes
had caused damage to shuttles in the past but had never resulted in a major accident. Some concerned engineers described the foam strike as “the largest ever” and asked that additional satellite images of the strike area be taken, but top managers rejected these requests. Many individuals at NASA reported that the group dynamics did not encourage a candid discussion of threats. Meeting transcripts revealed that managers did not actively seek dissenting views. Packed agendas inhibited thoughtful discussions of potential threats. Hierarchy and status differences made it difficult for lower level engineers to express their concerns [3].

**Dependability** Several of the advantages that derive from the use of organized or collective production, as opposed to individual, stem from the process of the division of labor and task specialization. Task specialization or the division of the production process into separate operations allows the human and physical resources used at each stage of the production process to become more skilled, specialized, and therefore more productive [11]. Task visibility refers to how well a task allows for the monitoring and evaluation of individual performances. When an individual work alone their output can often be measured easily and their task visibility is therefore high. In the case of a group working together on obscure task in for example a research and development laboratory the individuals contribution can be hard to measure and the task visibility is therefore low [11]. This can result in a situation where it is hard to see the individual workers marginal utility. The employees will have less incentives to work hard and no incentive to improve performance unless conditions allow employees to demonstrate their contributions and to obtain the rewards gained from increased performance [11]. For many productive activities, task specialization and joint specialization (when a group of two or more individuals work together) makes it impractical or too costly to monitor an individual’s performance or marginal productivity. This means that even though the division of labor is efficient on technical grounds, it can produce problems of control and coordination at the social system level [11]. The problem is that the same factor that increases the potential total output in team or collective production serves to decrease the average outputs of the individual team members. In essence, individuals lack the incentive to increase their performance when their performance cannot be measured or when rewards are not distributed on the basis of their marginal productivity [11].

**Structure, Clarity and the Meaning and Impact of Work** These aspects are all about making the workers feel fulfilled and help them see the larger picture and how their work relates to that. Structure and clarity are about the individuals understanding of the job expectation and the process of fulfilling these expectations. It is also about setting specific and attainable goals. Meaning helps the workers find a sense of purpose in either the work or the output and it can be an important tool in increasing the teams effectiveness. The meaning of the work can be personal and unique for all the individual workers,
the meaning can come from many different aspects from financial security or helping the team succeed. The impact refers to the workers own subjective view that their work is important and is making a difference for the team as a whole. Seeing that one’s work is contributing to the success of the organization’s goals can help reveal impact and keep the workers motivated [10].

Purpose and meaning can be very important tools to motivate a group of employees. There is an anecdote about when U.S President John F. Kennedy visited NASA headquarters for the first time in 1961. Supposedly when he toured the headquarters he met a janitor and asked him what he did. The janitor allegedly replied “I’m helping put a man on the moon!”. The story mediates the idea that a workforce motivated by a strong sense of higher purpose is essential to engagement [16].

The audit, tax, and advisory firm KPMG has focused a lot on enhancing their employees sense of purpose. To do this they encouraged everyone at the firm, from the Chairman all the way to the interns, to share their own stories about how their work is making a difference. They tried to reframe their roles within the firm and encouraged their employees to change their view of themselves. Instead of identifying themselves as professionals executing audit they wanted them to feel like members of a profession that helps millions of American families make better and informed decisions about investing their life savings. They wanted to get people talking about purpose to create a narrative that could connect their employees with the firm’s history of purposeful work and easier see their part of the bigger picture. To do so, they began collecting employee stories, highlighting the impactful work already being done, and teaching leaders how to talk about purpose with their people. To achieve their goals they started what they called the 10,000 Stories Challenge, asking their employees to share stories about the difference they make. They were hoping to get 10,000 stories but at the end of the challenge they had received approximately 42,000 stories. Surveys KPMG performed after this challenge showed a strong relationship between leaders who talk about the positive societal impact of their teams’ work and a variety of positive human resources and business indicators. Among the employees who said that their leaders discuss higher purpose, 94 % said KPMG is a great place to work, But among those employees whose leaders didn’t discuss purpose, the corresponding results was only 66 %. This group also reported they are three times more likely to think about looking for another job than those whose leaders did talk about purpose. Since then KPMG has incorporated purpose storytelling training into their leadership development programs [16].

2.3.2 The Olympic Games

In the world of sport team performance is also a subject of much research. The Olympics is a particularly researched subject since the nature of the event involves many different disciplines and athletes from all over the world. Many studies have been performed with the goal of trying to determine Olympic suc-
cess on a country level. One finding is that a nation’s GDP and population are key factors that can explain why countries perform the way they do at a given Olympics [1] [2] [5] [17]. This is not a surprising find since a larger population should lead to more Olympic caliber athletes available in the nation. A large GDP can indicate that the country has the available resources to support their athletes in the form of coaching, infrastructure and adequate training facilities. Additional factors such as host country advantage and the sport traditions in a country has also been shown as relevant to predict results [17]. Differences between Summer and Winter Olympics have been found. The effect of GDP is larger in the case of Winter Olympics while the effect of population is lower. The host country advantage has also been shown to be smaller in the Winter Olympics [1].

2.3.3 E-Sport

E-Sport is a new area that has not been the subject of much research. However one paper from 2016 aimed to examine whether similar country affects like those that was found in the Olympic can be found in E-Sport. Interestingly enough the findings showed that the factors GDP and population, the factors most important to predict Olympic success, was not significant when it came to E-Sport. Instead regional performance differences was attributed to other national factors such as high quality education and health care, long term orientation and masculine culture [15].

2.4 Multiple Linear Regression

A regression model is a model built from analyzing the relationship between the variables in a data set. A multiple regression model is a model that includes more than one factor, or covariate. All formulas and definitions are from [23] and [19].

2.4.1 Regression Model

A model with \( n \) observations has the following form:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \quad (4)
\]

\( y \) is called the response variable, \( x_j \) are the covariates, \( \beta_j \) are the covariate coefficients and \( \epsilon \) is the error term. This can be rewritten in the matrix notation

\[
y = X\beta + \epsilon \quad (5)
\]
where the different matrices are

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \] (6)

The covariate coefficients, \( \beta_i \), are unknown.

### 2.4.2 Ordinary Least Squares

To estimate the coefficients \( \beta_j \) one can use the method of least squares. The method is to minimize the least square function

\[
S(\beta) = \sum_{i=1}^{n} \epsilon^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta)
\] (7)

by taking its derivative with respect to \( \beta \) and set it equal to 0. Which results in the least-squares normal equations.

\[
X'X\hat{\beta} = X'y
\] (8)

The solution to the equations gives the least-square estimators of \( \beta \).

\[
\hat{\beta} = (X'X)^{-1}X'y
\] (9)

### 2.4.3 Assumptions

Five assumptions need to be fulfilled for a model to be stable [23].

1. There exists a linear relationship between the covariate and the response variable.
2. The error term \( \epsilon \) has a zero mean.
   \[
   E[\epsilon] = 0
   \] (10)
3. The error term \( \epsilon \) has constant variance \( \sigma^2 \)
   \[
   V[\epsilon] = \sigma^2
   \] (11)
4. The errors are uncorrelated. Which means that the value of one error does not depend on the value of any other error.
5. The errors are normally distributed.
2.4.4 Residuals

A residual is defined as the difference between the observed value and the fitted value

\[ e_i = y_i - \hat{y}_i, \quad i = 1, 2, ..., n \]  

The residual can be viewed as a measure of the variability of the response variable that the fitted model cannot explain. It is then convenient to think of the residual as the model errors. Meaning that the assumptions for errors must also hold for the residuals [23].

2.5 Hypothesis Test

A hypothesis is an educated guess or a proposed explanation to a phenomena. Hypothesis testing is used to conclude if the hypothesis proposed from the known data is a result by chance or not. One common hypothesis is if there exists a relationship between two variables [19]. To determine if there exists a relationship the two hypotheses are made, the null hypothesis and the alternative hypothesis.

\[ H_0: \quad \text{There is no relationship} \]
\[ H_1: \quad \text{There exists a relationship} \]

2.5.1 Test for Significance of Regression

To determine if there exists a linear relationship between the response variable and the covariates the test for significance of regression can be used [23]. The two hypothesis that are appropriate are

\[ H_0: \quad \beta_1 = \beta_2 = ... = \beta_k = 0 \]
\[ H_1: \quad \beta_j \neq 0 \quad \text{for at least one } j \]

If the null hypothesis is rejected it means that at least one of the covariates contributes significantly to the model. To test if the relationship exists the F-test can be used. The definition of an F statistic is

\[ F_0 = \frac{(SS_T - SS_{Res})/p}{SS_{Res}/(n - p - 1)} \]  

\[ p \] is the number of covariates and \( n \) is the number of observations. \( SS_R \) is the sum of squares due to regression and is defined as

\[ SS_T = y' y - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 \]
and $SS_{Res}$ is the sum of residual squares, defined as
\[ SS_{Res} = y'y - \hat{\beta}'X'y \] (17)

The $F$ statistic follows the $F_{k,n-k-1}$ distribution. If $F > F_{\alpha,k,n-k-1}$, where $\alpha$ is the level of significance, the null hypothesis is rejected.

2.5.2 Test on Individual Regression Coefficients

When it is determined that there exists a relationship between at least one of the covariates and the response variable it is interesting to know which covariates that is contributing. Adding covariates that does not seem to contribute may lead to a decrease in usefulness for the model. The hypotheses for this test are

\[ H_0 : \beta_j = 0 \]
\[ H_1 : \beta_j \neq 0 \] (18)

If the null hypothesis is not rejected the covariate $x_j$ can be deleted from the model. The test statistic for this hypothesis is
\[ t_o = \frac{\hat{B}_j}{\hat{\sigma}^2 C_{jj}} \quad \text{where} \quad \hat{\sigma}^2 = \frac{SS_{Res}}{n-2} \] (19)

$C_{jj}$ is the diagonal element of the matrix $(X'X)^{-1}$ corresponding to $\hat{\beta}_j$. To reject the null hypothesis the inequality $|t_o| > t_{\alpha/2,n-k-1}$ must be satisfied.

2.6 Model Selection

2.6.1 All Possible Regression

All possible regression is a method where all possible models are compared against each other [23]. In general with $n$ different covariates there are $2^n$ models. Since the number of models is exponentially related to the number of explanatory variables this method is only suitable when the number of covariates is relatively low. Lets assume a model with 3 covariates that are denoted $x_1, x_2, x_3$. Then there is a total of $2^3 = 8$ possible models. Let $y_n$ denote the the different models for $1 \leq n \leq 8$. The models can then be written in the following way:

\[ y_1 = \beta_0 + \epsilon \]
\[ y_2 = \beta_0 + \beta_1 x_1 + \epsilon \]
\[ y_3 = \beta_0 + \beta_2 x_2 + \epsilon \]
\[ y_4 = \beta_0 + \beta_3 x_3 + \epsilon \]
\[ y_5 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \]
\[ y_6 = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon \]
\[ y_7 = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \]
\[ y_8 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \] (20)
2.6.2 Akaike Information Criterion

The Akaike information criterion (AIC) is a model selection method for comparing different models against each other. AIC is used to determine whether or not a covariate should be included in the model. For ordinary least square regression the AIC is defined as

\[
AIC = n \ln \left( \frac{SS_{Res}}{n} \right) + 2p
\]  

(21)

Where \( n \) is the number of observations and \( p \) is the number of covariates. To compare models look at:

\[
\Delta AIC = AIC_{FULL} - AIC_{REDUCED}
\]  

(22)

If \( \Delta AIC > 0 \) the reduced model results in less information loss and is therefore preferred [13].

2.6.3 Mallow’s C\_p

Mallow’s C\_p is another model selection method used to assess the fit of a regression model. Mallow’s C\_p has been shown to be equivalent to Akaike information criterion in the case of Gaussian linear regression. It is defined as

\[
C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p
\]  

(23)

\( n \) is the number of observations and \( p \) is the number of covariates. \( SS_{Res}(p) \) is the residual sum of squares for a \( p \)-term subset model [23].

2.7 Model Errors

2.7.1 Heteroscedasticity

Non-constant variance of the error terms are called heteroscedasticity. It violates the import assumption that the errors have a constant variance. Which leads to having inconsistencies with the standard deviation of the estimators while the estimators stay the same.

To detect heteroscedasticity a scale-location plot can be used [12]. The plot illustrates the spread of the points across the predicted values. If the plot shows a horizontal band of points, the variance is constant indicating that there exists no heteroscedasticity. The model is then called homoscedastic. A non-straight band or a funnel indicates that the model is heteroscedastic. See example in figure 1.
Figure 2: Heteroscedasticity

One solution to heteroscedasticity is to transform the data with an appropriate transformation, for example $\log(y)$ or $\sqrt{y}$.

### 2.7.2 Non-Linearity of Data

One of the assumptions is that there exists a linear relationship between the response variable and the covariates. If the relationship is instead non-linear the model fit and ability to predict will be greatly reduced [19]. The residual plot is one way to control if there exists a linear relationship. If the plot shows horizontal band without any clear pattern of a bow, the relationship is linear [12].

If a non-linear pattern emerges one can apply a non-linear transformation to one or more covariate. For example $\log(x_j)$ and $\sqrt{x_j}$. 

![Scale-Location](image)
2.8 Model Validation

2.8.1 $R^2$ and $R^{2}_{\text{Adj}}$

A tool for comparing different models against each other is the coefficient of determination. Coefficient of determination, or $R^2$, is a statistic that tells the proportion of variance explained by the regressors [19]. It is defined as

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_T}$$ (24)

Since $0 \leq SS_{\text{Res}} \leq SS_T$, the value of $R^2$ is between 0 and 1. For example, $R^2 = 0.89$ means that 89% of the variance is explained by the evaluated model. Adding a covariate will increase $R^2$, resulting in an increase even if the covariate does not add any value to the model. This could lead to adding terms that are unnecessary for the model. To work around this the $R^{2}_{\text{Adj}}$ statistic will show if the additional covariate will improve the model.

$$R^{2}_{\text{Adj}} = 1 - \frac{SS_{\text{Res}}/(n - p - 1)}{SS_T/(n - 1)}$$ (25)

2.8.2 Cook’s Distance

Cook’s distance is a measurement commonly used when estimating the influence of individual data points when performing a regression analysis. It is used to find data points with a large influence on the overall model which can then be checked for validity. These data points, outliers and light leverage points can distort the regression model and its accuracy. It works by comparing the least...
square estimate $\hat{\beta}$ based on all data and the estimate $\hat{\beta}_i$ where observation $i$ has been deleted. Cook’s distance, denoted $D_i$, of observation $i$ is defined as the sum of all the changes in the regression model when the observation $i$ is removed from the dataset. If the model have $p$ covariates the Cook’s distance for observation $i$ would be given by:

$$D_i = \frac{\hat{y}_i - \hat{y}}{pMS_{Res}} \text{ where } MS_{Res} = \frac{SS_{Res}}{n-p}$$

(26)

Where $\hat{y}_{(i)}$ is the response value obtained when excluding observation $i$.

### 2.8.3 Variance Inflation Factor

A tool to identify multicollinearity in a model is the variance inflation factors (VIF). It can be written as

$$VIF = \frac{1}{1 - R^2_j}$$

(27)

$R^2_j$ is the coefficient of determination from the linear equation

$$x_1 = \alpha_0 + \alpha_2 x_2 + \ldots + \alpha_k x_k$$

(28)

If $x_j$ is nearly linearly dependent for some other covariates $R^2_j$ will approach 1 and VIF will increase. If the value of VIF is larger than 10 there exists multicollinearity [23].

### 2.8.4 Mean Square Error (MSE)

The mean squared error (MSE) can be used to measure the quality of a regression method [19]. It is defined as the the average of squared errors

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(29)

The value is always a non-negative number. If the MSE is close to zero the fitted values, $\hat{y}_i$, is close to the observed values.
2.8.5 Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) is a method to estimate the test error for a model by creating a training set and a test set. The training set is used to fit the regression model to and later compared against the test set to obtain the test error. The test set for LOOCV contains only one observation \((x_1, y_1)\). The observation is used to calculate the mean squared error, \(\text{MSE}_1\). Since the observation was not used when building the model the \(\text{MSE}_1\) is an approximately unbiased estimate for the test error. Repeating this process and going through all the observations gives a set of values \{\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_n\}. The LOOCV is then defined as the average of the MSE [19].

\[
\text{CV}_n = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i
\] (30)
3 Method

3.1 Demarcation

This study is primarily analyzing human factors for high level performance in a Dota 2 professional team. The data was collected from open sources due to limited resources. No team, organization or player was contacted for the purpose of collecting data. For example as how much a team trains together, how much a player plays in their spare time, if the players feel secure with their teammates and dare to make mistakes, their salaries and more. All these factors could be significant but are outside the scope of this report. The study only includes active professional teams that have an Elo rating. To count as a player for the team, the player needs to have played at least three games (one match) with the team in a competition.

A time series analysis approach was considered for this project, however finding historical data for the players and teams proved hard. Therefore a cross sectional regression model was chosen instead.

3.2 Variable Selection

The variables were deliberately chosen to reflect the human players and team compositions instead of playing styles, hero selection and strategies. The reason behind this decision was that the model should reflect the human and not in-game mechanics, even if both surely play a role in how successful a team is.

3.2.1 Response Variable

Average Elo rating The Elo rating changes slightly between every match played between professional teams. The average Elo rating for 30 days was chosen to get a value that is relatively stable over a longer period of time, with \( K = 32 \).

3.2.2 Explanatory Variables (Naive Model)

Estimated mmr: The naive model includes five covariates, each being the estimated mmr for a player in the team. Matchmaking rating (mmr) is a numerical value that estimates the skill level for a player [22]. Dota 2 measures two different mmr values. Solo mmr and party mmr. Solo meaning playing alone with four unknown players. Party meaning that you play together with one or up to four friends in a group. The mmr used in this model is an estimation of the solo mmr.
3.2.3 Explanatory Variables (Extended Model)

**Total matchmaking rating:** All team members combined estimated solo mmr.

**Matchmaking rating difference:** Taking the difference between the highest and lowest rated player in the team gives the range of how skilled the individual players on the team are.

**Number of games:** Total amount of games that the team has played, with any team composition. It might be an indicator how experienced the organization is in a competing environment.

**Age of the Dota 2 team:** The age of the organization’s Dota 2 commitment. The team has not necessarily been active all throughout the whole time span but the organization as such have experience of running a Dota 2 team.

**Region of the organization:** The Dota 2 community is divided into different regions. These regions are taken into consideration for example when teams get invited to major tournaments, to make sure all regions are represented [21]. These regions are North America, South America, Europe, Commonwealth of Independent States (CIS), China and Southeast Asia. See appendix for region distribution.

**Number of different nationalities:** How many different nationalities that are represented in a team. The reason for including this is because during all team sport some kind of communication takes place and Dota 2. Different nationalities would mean that this communication will probably require the players to talk in a second language.

**Total number of players:** The sum of total people that have played in the team, either as a stand-in or as a team player.

3.3 Data Collection

The data was collected from two websites that are using the Steam WebAPI, which gives them direct access to the Steam data base. The average Elo rating was downloaded as an excel file from www.datdota.com and the organization age was manually collected from the same website when going through each team in the excel file.

The excel file contained all the teams unique team ID number. A Python program was written to iterate over the team IDs and to use the ID to download json files by using opendota’s own API. All the data was not contained in a single file so the program needed to download a new file for each team and player. First the total matches was calculated by summing the amount of matches the teams had won and lost. The program proceeded to download the files containing information about how many players the team had been associated with and
how many were active. Using the limitation that a player must have to played at least three games, the program counted how many players that had represented the team. If there was exactly five active team members, which is the required amount to play a game of Dota 2, the program stored their personal player ID and proceeded with downloading the files containing the information about the players. From these files the estimated mmr was extracted. The program calculated the total mmr in the team by summing each players estimated mmr. The range of the mmr was calculated by taking the difference between the largest mmr value and the lowest mmr value. All the data was saved in the same excel file as earlier.

Nationality of the players and the organizations was manually collected through the websites www.gosugamers.net and www.liquidpedia.net due to lack of options. The data of nationality does not seem to be stored in the Steam data base or is at least not public.

3.4 Model Selection

An initial model was first fitted using all covariates from the data set. To verify if the model met the assumptions necessary a residual analysis was conducted, using the diagnostic plots: The Q-Q plot, Residual plot and Scale-location plot. The VIF was calculated for the initial model to make sure that there did not exist any multicollinearity between the covariates, using 10 as the cut off value. Cooks distance was calculated to diagnose if there existed any influential points that could have an unnecessarily high impact on the model. $R^2$, $R^2_{Adj}$ and the cross validation value $CV_{(n)}$ was calculated to be compared with the reduced model.

With ”All possible regression” all regression models was generated and tried, using AIC and $C_p$ to compare the different models with each other. A reduced model was chosen on the basis of having the lowest AIC and $C_p$ values, compared to the other models. The same residual analysis was performed to verify that the reduced model met the assumptions needed. $R^2$, $R^2_{Adj}$ and $CV_{(n)}$ was calculated so it could be compared with the initial model, to see if the model had improved or not. The process was performed both for the naive model and the extended model. Since the number of models was relatively low, 128 for the extended model, every model was fitted and tried with LOOCV to compare those results as well.
4 Results

4.1 Naive Model

To decrease the variance a linear transformation with the natural logarithm was used on the response variable.

\[
\log(\hat{y}) = \hat{\beta}_0 + \sum_{j=1}^{5} \hat{\beta}_j x_j
\]  

(31)

Table 1: Covariates, naive initial model

<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 ) Estimated mmr, position 1</td>
</tr>
<tr>
<td>( x_2 ) Estimated mmr, position 2</td>
</tr>
<tr>
<td>( x_3 ) Estimated mmr, position 3</td>
</tr>
<tr>
<td>( x_4 ) Estimated mmr, position 4</td>
</tr>
<tr>
<td>( x_5 ) Estimated mmr, position 5</td>
</tr>
</tbody>
</table>

4.1.1 Validating Naive Initial Model

The model needs to fulfill the assumptions presented in section 2.4.3. First assumption is that there exists a linear relationship between the covariates and the response variable. The null hypothesis is that there exists no linear relationship between the response variable and at least one covariate. Using equation (15), \( F_0 \) is calculated and compared to \( F_{0.05,5,49} \).

\[
7.269 > 2.404
\]  

(32)

indicating that at least one \( \beta_j \neq 0 \). The p-value is \( 3.684 \cdot 10^{-5} \), concluding that there exists a relationship between the response variable and the covariates.

To further verify this the residuals are plotted against the fitted values.
Figure 4: Naive initial model, residual vs fitted

Figure 4 shows an almost straight line that has a small deviation towards the end. This could be a result of a small sample size of only 55 observations. The plot is interpreted to show a linear relationship.

The second assumption is that the errors are normally distributed. One method to verify this assumption is to use the normal Q-Q plot [12]. The plot shows the residuals plotted against the quantiles. If the errors are normally distributed the points in the plot should form a straight line.

Figure 5: Naive initial model, Normal Q-Q

A straight line with a small deviation in the end is shown in figure 5. It is sufficiently normally distributed to continue evaluate the naive initial model.

The third assumption is that the errors have a constant variance. The scale-location plot shows the spread of the points across the predicted values range. If
the points form a horizontal random pattern the errors have constant variance [12].

Figure 6: Naive initial model, Scale-Location

The horizontal pattern in figure 6 indicates that there is constant variance for the errors, which verifies the third assumption.

To diagnose if there exists any multicollinearity among the covariates the VIF statistic was calculated.

Table 2: VIF, Naive initial model

<table>
<thead>
<tr>
<th>Estimated mnr</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>1.62</td>
</tr>
<tr>
<td>Position 2</td>
<td>1.70</td>
</tr>
<tr>
<td>Position 3</td>
<td>1.92</td>
</tr>
<tr>
<td>Position 4</td>
<td>1.77</td>
</tr>
<tr>
<td>Position 5</td>
<td>2.11</td>
</tr>
</tbody>
</table>

Using 10 as the cut off value, it is concluded that the model is free from multicollinearity.

Cooks distance, $D_i$, was calculated for to diagnose if there existed any outliers in the data that could distort the the outcome of the model. The largest value for $D_i$ was $D_{30} = 0.085 < 1$. Hence, there exists no influential points in this model.

The naive initial model meets the assumptions necessary. The properties of the model is presented in table 3 and 4.
Table 3: Naive initial model, estimations

|          | \(\beta\)  | Std. Error | Pr(>|t|) | 2.5%   | 97.5%   |
|----------|-------------|------------|----------|--------|--------|
| (Intercept) | 6.161       | 1.723 \cdot 10^{-1} < 2 \cdot 10^{-10} | 5.815 | 6.518 |
| Position 1  | 1.810 \cdot 10^{-5} | 1.920 \cdot 10^{-5} | 0.350 | -2.059 \cdot 10^{-5} | 5.669 \cdot 10^{-5} |
| Position 2  | 6.085 \cdot 10^{-6} | 2.994 \cdot 10^{-5} | 0.840 | -5.408 \cdot 10^{-6} | 6.626 \cdot 10^{-6} |
| Position 3  | 2.371 \cdot 10^{-5} | 2.509 \cdot 10^{-5} | 0.349 | -2.671 \cdot 10^{-5} | 7.412 \cdot 10^{-5} |
| Position 4  | 9.988 \cdot 10^{-6} | 2.046 \cdot 10^{-5} | 0.628 | -3.112 \cdot 10^{-5} | 5.110 \cdot 10^{-5} |
| Position 5  | 5.588 \cdot 10^{-5} | 2.156 \cdot 10^{-5} | 0.0125 | 1.255 \cdot 10^{-5} | 9.920 \cdot 10^{-5} |

Table 4: Naive initial model

\[
\begin{array}{ll}
R^2 & 0.4251 \\
R^2_{\text{Adj}} & 0.3664 \\
\text{CV}_{(55)} & 0.00469 \\
\end{array}
\]

42.51% of the variation is explained by the covariates in this model and the \(\text{CV}_{(55)}\), which is average mean square error, is 0.00469.

4.2 Naive Reduced Model

The naive initial model only contained 5 covariates. Since the number of covariates, all possible regression method can be applied to reduce the model. All possible models was built and the model with the lowest AIC and \(C_p\) was selected. The reduced model with AIC value of -140.34 and \(C_p\) of 1.47 only contained the covariates of position 3 and position 5.

\[
\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_3x_3 + \hat{\beta}_5x_5
\]

4.2.1 Validating Naive Reduced Model

The same analysis was performed as for the full naive model. First the assumption that there exists a linear relationship between the response variable and the covariates must be verified. As before the F-statistic was calculated. \(F_0 = 17.91 > 3.175 = F_{0.05,2,52}\). The p-value is \(1.211 \cdot 10^{-6}\). The null hypothesis is rejected and at least one of the \(\beta_j\) is not zero. To confirm this the residuals are plotted against the fitted values.
Figure 7: Naive reduced model, residual vs fitted

Figure 7 that there exist a linear relationship between the covariate and the response variable.

As before the errors need to have a normal distribution. The Q-Q plot is used to verify this assumption.

Figure 8: Naive reduced model, normal Q-Q

Almost the same straight line is shown in the Q-Q plot. The errors have a normal distribution.

The naive reduced model needs to have constant variance. The scale-location plot is used again to verify this assumption.
Figure 9 shows a small inclination but this could, as noted before, be a result of having a sample size of only 55 observations. Since the plot points looks randomly distributed in a horizontal band the assumption of constant variance.

VIF was calculated to diagnose if there existed any multicollinearity between the covariates in the naive reduced model.

**Table 5: VIF, Naive Reduced Model**

<table>
<thead>
<tr>
<th>Estimated mmr</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 3</td>
<td>1.47</td>
</tr>
<tr>
<td>Position 5</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Using the cut off value of 10 it is concluded that there exists no multicollinearity.

To diagnose if there existed any influential points in the reduced model Cooks distance was calculated. The largest $D_i$ value was $D_{30} = 0.122 < 1$. There exists no influential points in this model. This concludes that the naive reduced model is valid and the properties of the model is presented in table 6 and 7.

**Table 6: Naive reduced model estimations**

|             | $\beta$  | Std. Error | Pr($>|t|)$        | 2.5%   | 97.5%   |
|-------------|----------|------------|-------------------|--------|---------|
| (Intercept) | 6.242    | 0.1318     | $< 2 \cdot 10^{-16}$ | 5.978  | 6.507   |
| Position 3  | $3.311 \cdot 10^{-5}$ | $2.167 \cdot 10^{-5}$ | 0.1326 | -1.038$\cdot 10^{-5}$ | 7.659$\cdot 10^{-5}$ |
| Position 5  | $6.916 \cdot 10^{-5}$ | $1.773 \cdot 10^{-5}$ | $2.767 \cdot 10^{-4}$ | 3.358$\cdot 10^{-5}$ | 1.047$\cdot 10^{-4}$ |
Table 7: Naive reduced model properties

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.4078</td>
</tr>
<tr>
<td>$R^2_{\text{Adj}}$</td>
<td>0.3851</td>
</tr>
<tr>
<td>CV(55)</td>
<td>0.00442</td>
</tr>
</tbody>
</table>

4.3 Extended Initial Model

The extended model includes more covariates than the naive model. The estimated mmr have been summarized to a single value. The rest of the covariates are explained in section 3.2.3. The Chinese region is used as a benchmark for the dummy variables. The response variable was transformed using the natural logarithm to make the model more linear.

$$\log(\hat{y}) = \hat{\beta}_0 + \sum_{j=1}^{11} \hat{\beta}_j x_j$$

(34)

Table 8: Covariates, extended initial model

<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$ Total mmr sum</td>
</tr>
<tr>
<td>$x_2$ Mmr difference</td>
</tr>
<tr>
<td>$x_3$ Games</td>
</tr>
<tr>
<td>$x_4$ Age of the organization</td>
</tr>
<tr>
<td>$x_5$ Region North America (dummy)</td>
</tr>
<tr>
<td>$x_6$ Region South America (dummy)</td>
</tr>
<tr>
<td>$x_7$ Region Europe (dummy)</td>
</tr>
<tr>
<td>$x_8$ Region CIS (dummy)</td>
</tr>
<tr>
<td>$x_9$ Region Southeast Asia (dummy)</td>
</tr>
<tr>
<td>$x_{10}$ Sum of nationalities</td>
</tr>
<tr>
<td>$x_{11}$ Total amount of players</td>
</tr>
</tbody>
</table>

4.3.1 Validating Extended Initial Model

The model needs to fulfill the same assumptions as the naive model before it can be reduced. The same line of reasoning was applied as the naive initial model, therefore only the results and comments about the results will be presented.

$$F_0 = 5.604 > 2.019502 = F_{0.05,11,43}$$

(35)
The p-value for the model is $1.712 \cdot 10^{-5}$. The $F$ statistic and the low p-value indicates that it exists a relationship between the covariates and the response variable. Figure 10 shows a horizontal pattern supporting that the relationship is linear.

The almost straight band of points along the line in figure 11 indicates that the errors follow a normal distribution.
Horizontal pattern in figure 12 indicates that the variance is constant for the extended full model.

Table 9: Extended full mode, VIF

<table>
<thead>
<tr>
<th></th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mmr sum</td>
<td>3.43</td>
</tr>
<tr>
<td>Mmr difference</td>
<td>2.02</td>
</tr>
<tr>
<td>Games</td>
<td>4.13</td>
</tr>
<tr>
<td>Age of the organization</td>
<td>5.50</td>
</tr>
<tr>
<td>Region of the organization</td>
<td>3.94</td>
</tr>
<tr>
<td>Sum of nationalities</td>
<td>2.22</td>
</tr>
<tr>
<td>Total amount of players</td>
<td>3.79</td>
</tr>
</tbody>
</table>

With a chosen cut off value of 10, the extended full model is free from multicollinearity.

The largest Cooks value in the extended model is $D_{40} = 0.107$ which is smaller than 1. Hence, there exists no influential points in the extended model.
Table 10: Extended full model estimations

|                        | $\hat{\beta}$ | Std. Error | Pr($>|t|)$ | 2.5% | 97.5% |
|------------------------|---------------|------------|-----------|------|-------|
| (Intercept)            | 5.8037        | 0.2379     | $< 2 \cdot 10^{-16}$ | 5.32 | 6.28  |
| Total mmr sum          | $3.088 \cdot 10^{-5}$ | $6.696 \cdot 10^{-6}$ | $3.57 \cdot 10^{-5}$ | 1.74 | 4.44 \cdot 10^{-5} |
| Mmr difference         | $4.138 \cdot 10^{-5}$ | 1.801 \cdot 10^{-5} | 0.0265 | 5.06 | 7.10 \cdot 10^{-5} |
| Games                  | $5.701 \cdot 10^{-5}$ | 3.421 \cdot 10^{-5} | 0.1029 | -1.10 | 1.265 \cdot 10^{-4} |
| Age of the organization | 1.877 \cdot 10^{-5} | 0.0007     | 0.9780 | -1.38 | 1.38 \cdot 10^{-3} |
| Region CIS             | -0.0277       | 0.0305     | 0.3688 | -8.91 | 3.38 \cdot 10^{-2} |
| Region Europe          | 0.0092        | 0.0303     | 0.7635 | -5.10 | 7.03 \cdot 10^{-2} |
| Region North America   | 0.0415        | 0.0325     | 0.2084 | -2.40 | 1.07 \cdot 10^{-1} |
| Region South America   | 0.0878        | 0.0348     | 0.0154 | 1.76  | 1.58 \cdot 10^{-1} |
| Region Southeast Asia  | 0.0216        | 0.0301     | 0.4774 | -3.92 | 8.23 \cdot 10^{-2} |
| Sum of nationalities   | -5.988 \cdot 10^{-4} | 9.306 \cdot 10^{-3} | 0.9490 | -1.94 | 1.82 \cdot 10^{-2} |
| Total amount of players| -0.0007       | 0.0013     | 0.5615 | -3.32 | 1.84 \cdot 10^{-3} |

Table 11: Extended full model, properties

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.5891</td>
</tr>
<tr>
<td>$R^2_{\text{Adj}}$</td>
<td>0.4840</td>
</tr>
<tr>
<td>CV(55)</td>
<td>0.0045</td>
</tr>
</tbody>
</table>

4.4 Extended Reduced Model

Since the number of covariates are low in the full model, it is computationally viable to perform all possible regressions. This resulted in 128 different models. To determine which model is best fit relatively to the other models the AIC and Mallows $C_p$ was used. The model with the lowest AIC and $C_p$ values was chosen to be the final model. With a AIC value of -147.87 and $C_p$ of -1.55 the final model is:

$$\log(\hat{y}) = \hat{\beta}_0 + \sum_{j=1}^{8} \hat{\beta}_j x_j$$  \hspace{1cm} (36)
Table 12: Covariates, extended reduced model

| Data             |  
|------------------|------------------|
| $x_1$            | Total mmr sum    |
| $x_2$            | Mmr difference   |
| $x_3$            | Games            |
| $x_4$            | Region North America (dummy) |
| $x_5$            | Region South America (dummy) |
| $x_6$            | Region Europe (dummy) |
| $x_7$            | Region CIS (dummy) |
| $x_8$            | Region Southeast Asia (dummy) |

The model uses the Chinese region as a benchmark, the same as for the initial model.

4.5 Validating Extended Reduced Model

The same procedure is followed earlier.

$$F_0 = 5.604 > 2.147288 = F_{0.05,8,46} \quad (37)$$

![Residuals vs Fitted](image)

Figure 13: Extended reduced model, residuals vs fitted

The p-value for the reduced model is $9.427 \cdot 10^{-7}$. In combination with the $F$ statistic, the p-value and the residual plot it is concluded that a linear relationship exists between the covariates and the response variable.
Figure 14 indicates that the errors of the model have a normal distribution.

Figure 15 shows a random horizontal pattern. Hence, the errors of the model have constant variance.

All the VIF values are under 10. There exists no multicollinearity in the reduced model.
Table 13: Extended reduced mode, VIF

<table>
<thead>
<tr>
<th></th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mmr sum</td>
<td>2.31</td>
</tr>
<tr>
<td>Mmr difference</td>
<td>1.55</td>
</tr>
<tr>
<td>Games</td>
<td>1.43</td>
</tr>
<tr>
<td>Region of the organization</td>
<td>1.79</td>
</tr>
</tbody>
</table>

The lowest Cooks distance value is $D_{50} = 0.106$ which is smaller than 1. Therefore it does not exist any influential points in the reduced model. The reduced model is presented in table 13 and 14.

Table 14: Extended reduced model estimations

|                      | $\beta$ | Std. Error | Pr($>|t|)$ | 2.5%  | 97.5%  |
|----------------------|---------|------------|-----------|-------|--------|
| (Intercept)          | 5.812   | 1.938e-01  | < 2e-16   | 5.42  | 6.20   |
| Total mmr sum        | 3.054\cdot10^{-5} | 5.345\cdot10^{-6} | 7.77\cdot10^{-7} | 1.98\cdot10^{-5} | 4.13\cdot10^{-5} |
| Mmr difference       | 3.679\cdot10^{-5} | 1.533\cdot10^{-5} | 0.02052 | 5.93\cdot10^{-6} | 6.77\cdot10^{-5} |
| Games                | 0.0493  | 0.0213     | 0.0252    | 4.85\cdot10^{-6} | 8.37\cdot10^{-5} |
| Region CIS           | -3.171\cdot10^{-2} | 2.700\cdot10^{-2} | 0.24629 | -8.61\cdot10^{-2} | 2.27\cdot10^{-2} |
| Region Europe        | 8.986\cdot10^{-3} | 2.526\cdot10^{-2} | 0.72368 | -4.19\cdot10^{-2} | 5.98\cdot10^{-2} |
| Region North America | 4.163\cdot10^{-2} | 2.743\cdot10^{-2} | 0.13596 | -1.36\cdot10^{-2} | 9.68\cdot10^{-2} |
| Region South America | 8.194\cdot10^{-2} | 2.925\cdot10^{-2} | 0.00742 | 2.31\cdot10^{-2} | 1.413\cdot10^{-1} |
| Region Southeast Asia| 1.542\cdot10^{-2} | 2.511\cdot10^{-2} | 0.54228 | -3.51\cdot10^{-2} | 6.10\cdot10^{-2} |

Table 15: Extended reduced model, properties

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.5848</td>
</tr>
<tr>
<td>$R^2_{Adj}$</td>
<td>0.5126</td>
</tr>
<tr>
<td>CV(_{(55)})</td>
<td>0.0039</td>
</tr>
</tbody>
</table>
4.6 Model Comparison and Analysis

All models are summarized in a single table with their features.

Table 16: Model comparisons

<table>
<thead>
<tr>
<th></th>
<th>Naive Initial Model</th>
<th>Naive Reduced Model</th>
<th>Extended Initial Model</th>
<th>Extended Reduced Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>-135.97</td>
<td>-140.34</td>
<td>-142.44</td>
<td>-147.87</td>
</tr>
<tr>
<td>$C_p$</td>
<td>6.00</td>
<td>1.47</td>
<td>4.00</td>
<td>-1.55</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4251</td>
<td>0.4078</td>
<td>0.5891</td>
<td>0.5848</td>
</tr>
<tr>
<td>$R^2_{Adj}$</td>
<td>0.3664</td>
<td>0.3851</td>
<td>0.4840</td>
<td>0.5126</td>
</tr>
<tr>
<td>CV$^{(55)}$</td>
<td>0.0047</td>
<td>0.0044</td>
<td>0.0045</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

The $R^2$ increased substantially when going from the naive model to the extended model. This is a result of adding more covariates in the model. In general, $R^2$ does not decrease if a covariate is added [23]. The $R^2_{Adj}$ increases in both the naive reduced model and the extended reduced model, indicating that removing the covariates in the initial models leads to a better fit. The largest $R^2_{Adj}$ belongs to the extended reduced model. Hence, including region of the organization, games played and the skill difference leads to a better model than just looking at the players individual skill. This is also evident when looking at the average mean square error from the cross validation. The lowest value of all 128 models was 0.0039 for the chosen reduced model. The average mean square error decreased with 11.36% with the extended reduced model compared to the naive reduced model, and even more comparing it with the other two models.
5 Discussion

The naive model showed that the individual skill of the players could be used to predict the performance of the team. In the extended model the variable for skill used was the total individual skill instead of a covariate for each individual player. This was due to the fact that the data sample was limited in size and having five covariates representing individual skill was deemed suboptimal.

The result of the extended model shows that the number of nationalities present in the team does not have a significant effect on the teams performance. The covariate for the total number of nationalities had a p-value close to 0.95 and was therefore removed from the model. This does not mean that cultural differences and language barriers can’t impact a teams performance since people from different countries can have languages in common. Due to a small sample size of 55 teams we can not examine the effect of language barriers since most professional teams have members speaking the same languages.

The study does not include the players age, which is quantifiable, since it is not stored in the Steam database. We tried to collect the age of all the players manually through different websites but the Dota 2 scene is too immature to store accurate information of that kind, compared to football for example where age, height and weight can be found on sport sites containing player data. A lot of the information regarding Dota 2 is still stored in wiki-sites like liquidpedia.

In Project Aristotle the key findings were factors that relates to the social dynamics and corporate culture of the groups and teams. Our research and model focused to more easily quantifiable factors. That is because we chose to study factors that are more available to an outside observer. In order to study the internal social dynamics of the Dota 2 teams we would need to do as vast data collection project and survey teams from all around the world. Google researched their own teams while we would have to study competing teams from different organizations which we believe would only add to the difficulty. Due to the scope of such a project and the possible problems that could arise such as language barriers and teams not wanting to answer surveys regarding internal social dynamics we deemed that it was not possible for us hence we instead used publicly available data. We do however believe that the social dynamics factors put forward by Project Aristotle could be helpful in modeling the skill and performance of Dota 2 teams. The factors psychological safety and dependability could possibly be very helpful since Dota 2 is a complex game with many different strategies and teammates need to be able to trust each other and be able to speak their mind without being seen as ignorant, incompetent, negative, or disruptive if they want to be able to compete at the highest levels. Structure and clarity in terms of setting up achievable goals for the team is also factors that can help the success of a team. The meaning and impact of work however might not be as significant in professional gaming since it is much less abstract then for example a big research project or preforming a highly specialized task.
so the players should be able to see the bigger picture of their work much easier. Also worth noting is that while the researchers of Project Aristotle found that individual performance of team members were not significantly connected with team effectiveness we found that the variable individual skill had significant explanatory power for the Dota 2 teams performance.

The research on both the Olympics and E-Sport found evidence to support the thesis that regional differences has an effect on the performance of athletes. While GDP and population was deemed to be the most influential factors when it comes to traditional sport the same did not hold true for E-Sport. The fact that a nation’s GDP was significant for traditional sport but not for E-Sport can possibly be explained by comparing the needs of traditional sport and E-Sport respectively. While traditional sport at the Olympic level has a need for world class training facilities, E-Sport participation costs are much lower. The Olympics is also more regulated then E-Sport. It requires a national Olympic committee that can work together with the International Olympic Committee in order to be eligible to compete and to determine the number of athletes a nation are allowed to send. No such governing body exists for E-Sport which likely helps keep the participation costs low. For the Olympics a host nation advantage was noted, we did not examine whether such an affect is present for Dota 2 as well. If a host nation advantage exists for Dota 2 we believe that the effect of it would have been greater if we had chosen a monetary measurement to represent the team’s success since most of the large Dota 2 tournaments are played in North America. Since we instead chose Elo rating which is not affected by the money at stake at a given tournament, host nation advantage might not have been as important for the model.
6 Conclusion

It is possible to create a linear fit both from the naive model and from the extended model. The naive model resulted in a worse fit than the extended models. This indicates that it is reasonable to take the nationality of the organization and the experience the organization into account along with the players combined estimated skill and skill difference.

The extended reduced model explains 58.48% of the variation in the data. Dota 2 is a team-based game and the outcome depends on how well the players can perform in the circumstances provided and the preparations leading up to the game. The factors in the result are connected to the preparations since they are not circumstantial. The p-value for matches is 0.1029, it is not possible to reject the null hypothesis for the covariate but including it leads to a better estimation. In the extended reduced model all coefficients, except the dummy variables, are positive. Hence, it is evident that the higher the estimated skill that the players have combined, the more matches the organization have played and the larger skill difference all have a positive effect on the teams rating. With the exception that the skill difference can be large but if the cause is that one or more players have an estimated skill bellow the average professional estimated skill, the teams combined skill will suffer. Therefor a large skill range is not necessarily a good thing for the team.

Using the Chinese region as the benchmark the study finds that all other regions except CIS predicts a better performance. The p-values for the different regions seen in table 10 is high. Region Europe has a p-value of 0.76 so we can’t reject the null hypothesis that there is no performance difference between the two regions. Region South America has the lowest p-value for the regions at 0.015 which makes us believe that regional differences might still be a factor. The team distribution shows that over a third of the teams in the data are from China. This gives the study more detailed information about the Chinese teams while the other countries lack the same detail due to the small sample size. Future research in this area could be to attempt to explain the regional differences that was observed. These regional differences can possibly be explained by examining demographic and economical factors.

For future research a more refined model could be obtained by looking at all registered teams in the Steam database, including amateur teams and teams that only participated in one competition, usually the International qualifiers. The soft factors found in Project Aristotle could also be interesting to examine to see how factors related to group dynamics can predict performance.
References


A Appendix

![Distribution of countries](image)

Figure 16: Distribution of countries

Teams

1. Virtus Pro
2. Team Secret
3. LGD Gaming
4. Evil Geniuses
5. SG E-sport Team
6. Team Liquid
7. VGJ Thunder
8. Fnatic
9. Newbee
10. OG
11. Invictus Gaming
12. Complexity Gaming
13. Pain Gaming
14. Optic Gaming
15. Vici Gaming
16. Mineski
17. TNC Predator
18. Team Spirit
19. Keen Gaming
20. Vega Squadron
21. Rock Young
22. Rock Gaming
23. Team Empire
24. Team Kinguin
25. Execration
26. Immortals
27. LGD.Forever Young
28. Digital Chaos
29. Vici Gaming Potential
30. IG Vitality
31. Skiter Evil
32. The Final Tribe
33. Geem Fam
34. Newbee Young
35. Effect
36. Team Braveheart
37. Eclipse
38. Ehome
39. Team Waoo
40. Mousesports
41. Mad King E-sport
42. T Show
43. Keen Gaming Luminous
44. VGJ Storm
45. Natus Vincere
46. Infamous
47. Sacred
48. Happy Feet
49. BOOM ID
50. Alliance
51. Clutch Gamers
52. CDEC Gaming
53. Gambit E-sport
54. Midas Club
55. Team Max