**Separating the best from the rest**

**An analysis into behaviors that best predict the skill of computer game players**

In any sport there is a recognized need for practice in order to improve. Serious competitors spend a great deal of time optimizing their practice, focusing on the most important areas in order to realize the most gain within the shortest timeframe. With the advent of computers and competitive computer gaming there is a wealth of new data where we can record, observe, and analyze the actions of players within various skill levels with unprecedented precision. By employing analytic techniques such as principal component analysis on this data we can identify which behaviors best separate the skill levels of players, and use that as a basis to determine a training regimen for statistically maximized improvement rates.

**Problem Statement:**

Using recorded data from competitive matches within a computer game we will convert a selection of observed player behaviors into principal components. Using an index of skill provided by the game itself as the response variable we will then apply principal component analysis to determine which groupings of those variables best describe the variation of the skill group for the players in that match. This should provide a good estimate of the behaviors that are most important to focus on during training in order to improve skill ranking within the game.

**Data Set Used:**

The data set that we will use consists of over 300 recorded matches of the popular competitive computer game “Starcraft 2”. These matches were recorded by a research team at Simon Fraser University who donated the data set to the UC Irvine Machine Learning Repository. *Starcraft* separates players into graduated “leagues” representing a general level of that player’s skill. We will use a continuous ordinal value representing these leagues as our response variable, with 1 representing “Bronze League” at the lowest skill level and 7 representing “Grand Master League” at the highest level.

Within each match of *Starcraft* both players control worker units who construct buildings and soldier units who aim to destroy the opponent’s buildings. Each player can assign keys on their keyboard as “hotkeys” to more quickly select these units, and each has a “minimap” that shows the entire playing area that has been explored by their units. Since the speed of the game is variable the researchers recorded time in terms of timestamps, where a game played on the fastest setting has roughly 88.5 timestamps per 1 real-time second. The researchers also recorded PACs or Perception Action Cycles which are defined as any one fixation by a player upon an area or unit to accomplish a task that consisted of one or more actions.

We will use the following 18 explanatory variables as part of our analysis:

1. Age: The age of each player
2. HoursPerWeek: Reported hours spent playing *Starcraft* per week.
3. TotalHours: Reported total hours spent playing *Starcraft*.
4. APM: The number of actions each player input into the game each minute.
5. SelectByHotkeys: The number of object selections made using hotkeys per timestamp.
6. AssignToHotkeys: The number of objects assigned to hotkeys per timestamp.
7. UniqueHotkeys: The number of unique hotkeys used per timestamp.
8. MinimapAttacks: The number of attack actions made on the minimap per timestamp.
9. MinimapRightClicks: The number of right-clicks on the minimap per timestamp.
10. NumberOfPACs: The number of PACs per timestamp.
11. GapBetweenPACs: Mean duration in milliseconds between PACs.
12. ActionLatency: Mean latency from the onset of a PAC to the first action in milliseconds.
13. ActionsInPAC: Mean number of actions within each PAC.
14. TotalMapExplored: The number of game coordinate grids viewed per timestamp.
15. WorkersMade: The number of worker units produced per timestamp.
16. UniqueUnitsMade: The number of unique units produced per timestamp.
17. ComplexUnitsMade: The number of complex units produced per timestamp.
18. ComplexAbilitiesUsed: The number of complex unit abilities used per timestamp.

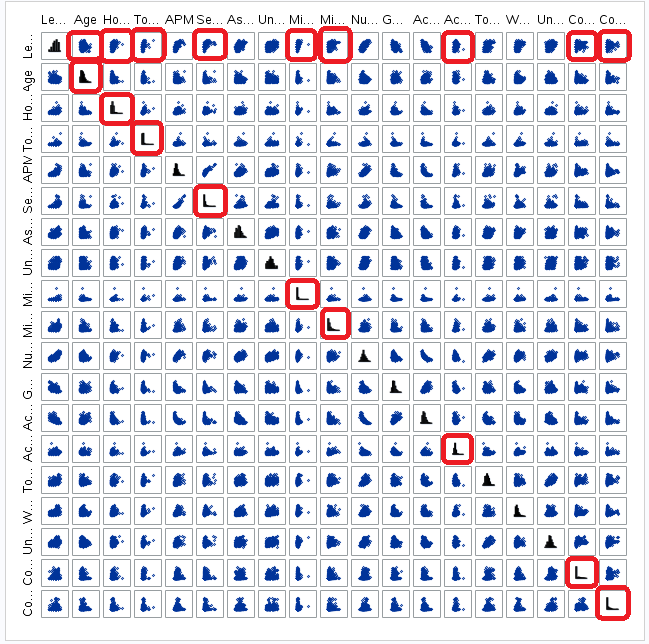
While all of these explanatory variables are assumed to be correlated with a player’s skill level we are not concerned with possible confounding factors as the process of reducing the variables to principal components will mathematically eliminate any collinearity between the resulting components.

**Scope:**

This was an observational study, and as such any inference made as a result of this study can only be one of association and not causation. There is no specific mention within the data set or its description detailing how the subjects were selected. Since we cannot say with certainty that the subjects were randomly selected from within a greater population we must limit any inferences of association to the specific population of players and games observed within the dataset. Since the game studied was not randomly selected any inferences to skill levels for games outside of *Starcraft* are speculative.

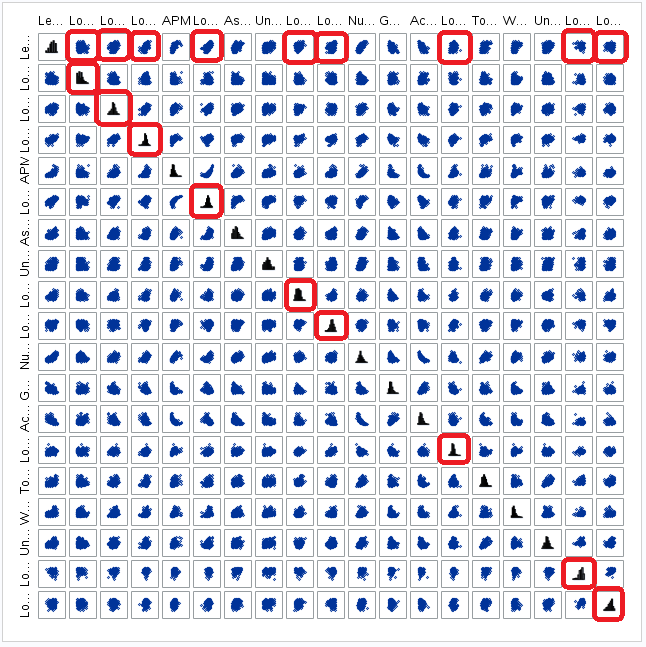
**Exploratory Data Analysis:**

Because all 18 variables were specifically selected to try and predict skill levels they are likely to be correlated. This, combined with the sheer number of variables involved, makes using Principal Components Analysis appealing as it will naturally result in principal components that are not correlated. PCA does, however, assume that all explanatory variables have a normal distribution and a linear association with the response. As seen in the scatterplot matrix below there is visual evidence to suggest that some variables do not meet these assumptions.



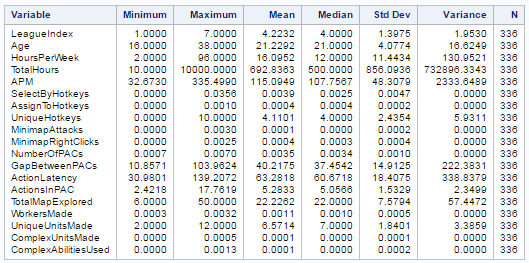
Display 1: Scatterplot Matrix

After performing a log transformation on the variables of concern they now appear to be much closer to the normal distribution and linear relationship needed for PCA, as seen below.



Display 2: Scatterplot matrix after log transformation of selected variables

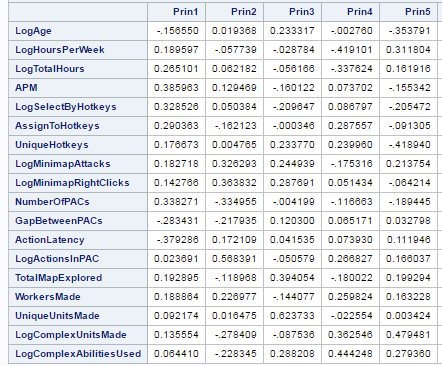
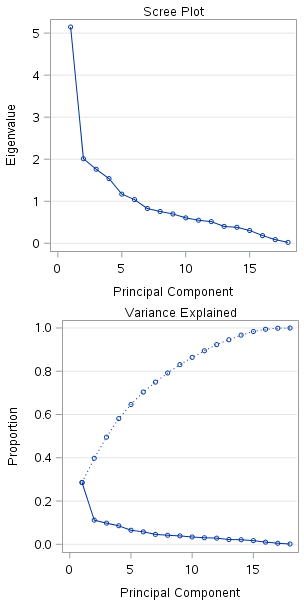
In examining the minimum and maximum values of the explanatory variables we see that all of them appear to be continuous, a requirement of PCA. There is some concern about the scale of the data in reference to each other, but so long as a correlation matrix is used PCA is extremely robust to deviations in scale as all of these values will become normalized to a single scale as part of the analytic process.



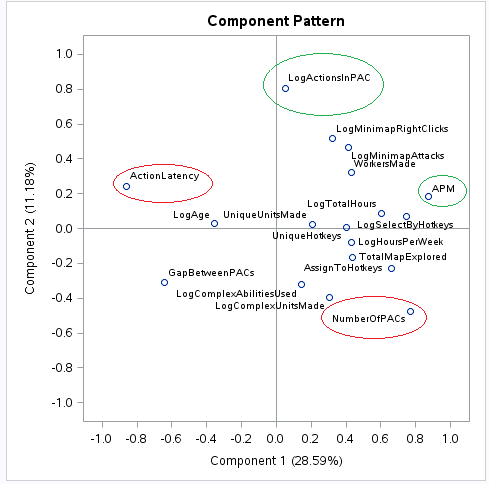
Display 3: Summary statistics including minimum and maximum values

**PCA Analysis:**

As seen in the Scree plot below there is a clear inflection point after principal component 2 where any additional principal components provide diminishing returns in regards to explaining more of the variation in the model.

Display 4: Selected Eigenvalues and Scree plot

In examining the first two principal components both with the eigenvalues recorded in Display 4 and the component patterns charted in Display 5 we see that while there are multiple loadings in each component the first component primarily deals with individual actions, where the APM (Actions Per Minute) variable has the most pronounced positive effect and ActionLatency (The delay between actions) shows the largest negative effect. The second component primarily deals with groups of actions focused on the same specific task, with the number of actions in each PAC (Perception Action Cycle) having the greatest positive effect and the number of PACs themselves having the most pronounced negative effect. In other words, having more actions in each task fixation but fewer task fixations overall is desirable.

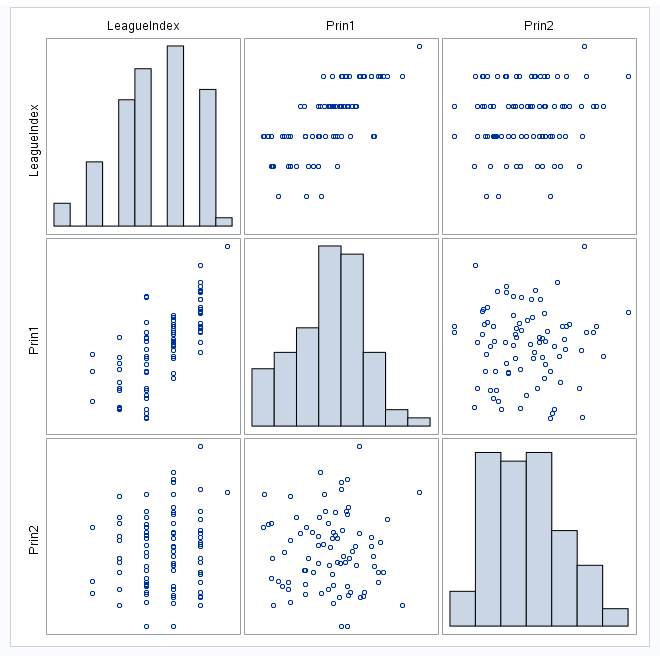
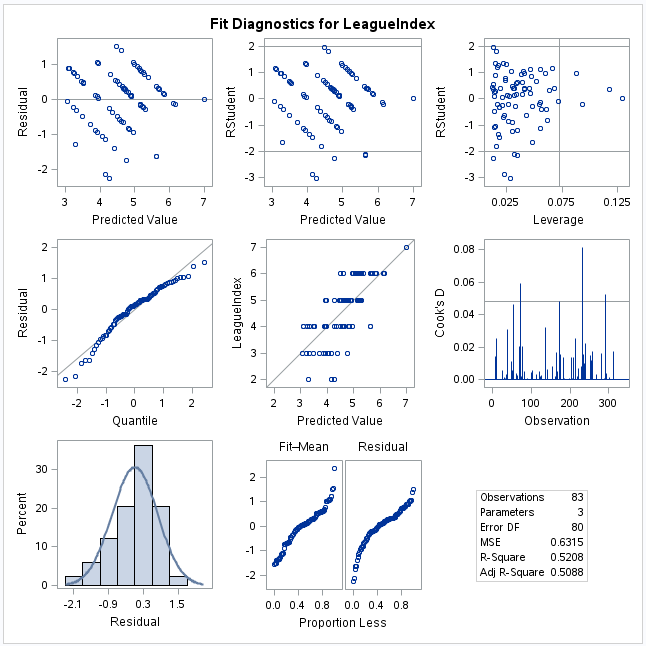
Display 5: Component pattern graph

**Regression Analysis:**

Based upon the results of the PCA analysis we will take P1 and P2 to represent principal component 1 and principal component 2 then move forward with the following model for our linear regression analysis:

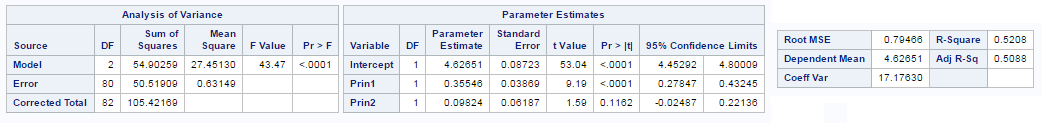
LeagueIndex = β­0 + β1­P1 + β2P2

Using that model we observe no visual evidence against a linear relationship between each principal component and the response variable. Furthermore we see no evidence against normality in the histogram or QQ plot, and we see no obvious outliers in the residual plot. The highest Cook’s D statistic is less than 0.1 so there is no evidence of any problematic outliers that would have leverage on our chosen regression model.

Display 6: Scatterplot matrix and fit diagnostics for our chosen regression model

While the model itself was found to be statistically significant at the alpha 0.05 level (P Value less than 0.001) only the first principal component was found to be statistically significant (P Value less than 0.001). The second principal component on its own came close but was ultimately found not to be statistically significant at the alpha 0.05 level (P Value of 0.1162).



Display 7: Parameter estimates for the selected regression model

**Conclusion:**

Within the dataset there were two principal components identified as being most predictive of player skill as measured by League Index. Principal Component One primarily measured how many actions a player made with the least delay between them. Principal Component Two primarily identified skilled players as having fewer fixations on tasks with more actions in each fixation, while unskilled players had more fixations on tasks with fewer actions per fixation.

There is overwhelming evidence that the main effects of Principal Component One are statistically significant at the alpha 0.05 level (P Value less than 0.0001). There is suggestive but not conclusive evidence that Principal Component Two is also significant (P Value 0.1162). The combined model including both principal components explains 50.88% of the variation in League Index, and there is strong evidence that the model itself is statistically significant at the alpha 0.05 level (P Value less than 0.0001).

These results suggest that in order to increase one’s skill at *Starcraft* it is most important to practice making as many actions as possible within as few task fixations, or Perception Action Cycles, as possible. Because this was an observational study where the game observed was not randomly selected and it is unknown if the players observed were randomly selected, no claims of causation may be made. Furthermore, any inferences of association must be limited to *Starcraft* and the players observed during the study.