In professional tennis the rankings are typically based on set parameters. Throughout the year there are tournaments that are worth different numbers of points. There are four tiers of events, Grand Slams, Masters 1000, ATP 500 and ATP 250. You receive the most points in your Association of Tennis Professionals (ATP) rankings for winning a Grand Slam, and the least from winning an ATP 250 tournament. In the ATP, Grand Slam tournaments are played in a best of 5 format, and non Grand Slam tournaments in a best of 3. In the WTA, all tournaments are played in a best of 3 format. The source of this data has calculated an ELO ranking, this ranking considers not only what tournament you are playing in, but who you are playing. For instance, if 4th ranked Coco Gauff beats 1st ranked Iga Swiatek, it will be worth more to her ELO ranking than if she beat 14th ranked Emma Navarro in the same tournament.

In Tennis, there are also three different types of surfaces that are played on. The options are Grass, Hard, and Clay. The surfaces are important to keep track of as the speed of tennis changes, e.g., clay generally slows the ball down whereas grass speeds it up. Certain players perform better on certain surfaces. This dataset contains information for each player on each surface.

In this worksheet, we will look at creating a multiple linear regression model to determine win percentage in the WTA.

(In order to be included in the data set, players must have played a minimum of 10 matches overall or 5 matches on a particular surface. This data was filtered so only players who have recorded data on all three surfaces are present)

1. Below is a table of the available variables and a table that shows what types of strengths and weaknesses a player has on each type of surface. Discuss which variables might be helpful in creating a model to predict win percentage.

|  |  |
| --- | --- |
| Variable | Description |
| Player | The name of the player |
| Surface | The surface of the tennis court |
| Matches | The number of matches the player has played |
| EloRank | The ELO ranking of this player |
| WinPercentage | The win percentage of this player |
| DoubleFaultPercentage | The percentage of the players’ services that result in a double fault |
| ReturnPointsWonPercentage | The percentage of return points that the player wins |
| AcesPerDoubleFault | The ratio of how many aces the player serves to double faults |

|  |  |  |
| --- | --- | --- |
| Surface | Characteristics | Best for |
| Clay | Slow, High bounce | Baseline players, Drop shots |
| Grass | Fast, Low bounce | Serve and volley, Big Servers |
| Hard | Medium speed, Highest bounce | Baseline players, Longer rallies |

1. Interpret the coefficients for return points won percentage and grass and clay surfaces.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** |
| Constant | -0.1898 | 0.0994 | -1.91 | 0.057 |
| ReturnPointsWonPercentage | 2.079 | 0.253 | 8.23 | 0.000 |
| EloRank | -0.001332 | 0.000120 | -11.07 | 0.000 |
| Surface |  |  |  |  |
| Grass | 0.1016 | 0.0209 | 4.85 | 0.000 |
| Hard | 0.0269 | 0.0166 | 1.62 | 0.106 |

Return Points: A 1 unit increase in the return points won percentage is associated with a 2.079 increase in win percentage holding all else constant.

Grass: Compared to playing on Clay courts, we would expect the win percentage for playing on grass courts to be 0.1016 higher, on average, holding all else constant.

1. Interpret the R-Squared value and decide if you think that this is a good model.

|  |  |  |
| --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** |
| 0.125993 | 48.20% | 47.57% |

About 47.57% of the variability of win percentage is explained by the multiple linear regression model.

This isn’t a great R-Squared value, so we should look at different models or transformations to our model that might be able to explain the data better.

1. Based off of the following plots, do you believe that this model meets the assumptions of a multiple linear regression model?

A graph with blue dots

Description automatically generatedA graph with a line

Description automatically generated

As we can see in the Q-Q plot on the left, this looks normally distributed as a vast majority of the points follow the line. On our plot of fitted values vs. residuals, the variance seems consistent with a slight narrowness towards the lower values.