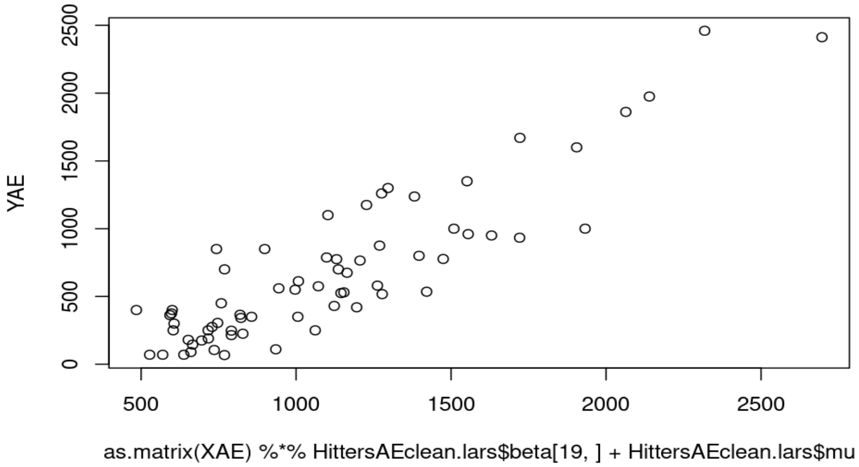
Hitters Assignment 2

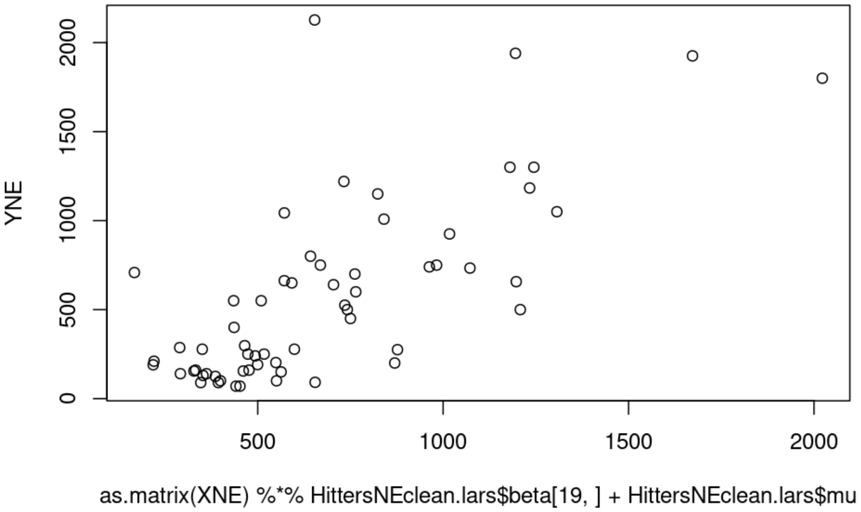
Step 2: Min Cp Lars model for AE league/division

AE CP Model



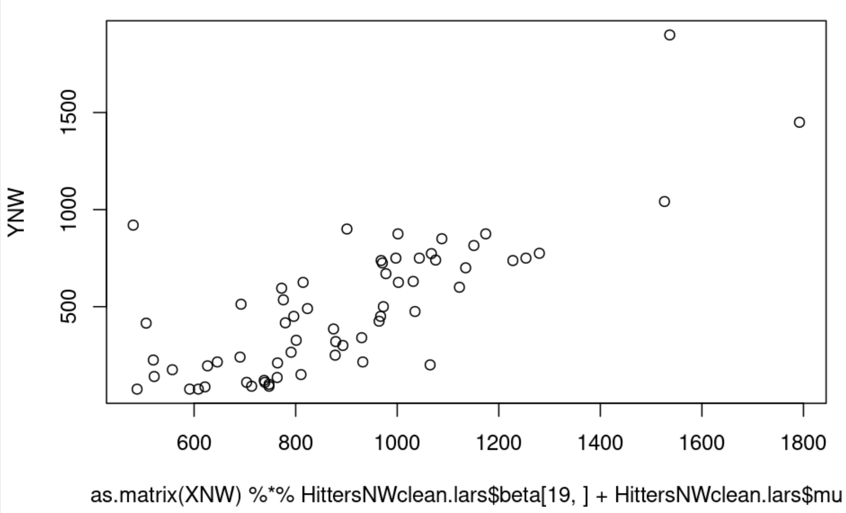
This is the AE CP Model, and we can clearly see an upward, positive, and almost linear relationship between the actual and predicted salaries of the player. There is a cluster of data values between the x-values of 500 to 1000, and then they start becoming more spread apart. Apart from the 500 to 1000 range, the data values start becoming more spread apart, in an upwards trend. This ends up creating a positive, almost linear relationship between the x and y values of the actual and predicted salaries of the players. For the most part, the data values are concentrated between 500 and 2000, but there are a couple of outliers placed slightly higher than 2000 and one outlier past 2500.

NE CP Model



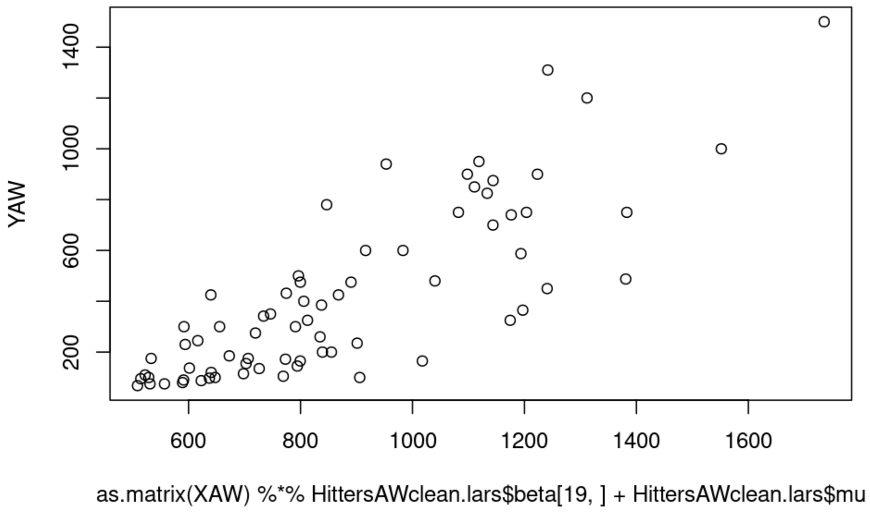
This is the NE CP model, and upon closer inspection, there isn't a clear relationship between the predicted and actual salaries of the players in this division. Instead, the data values are kind of spread out all throughout the graph. Again, there is a cluster located between the values of 0 to approximately 520. Multiple data values overlap with each other in that area. Moving out from that area, there are scattered values around the y values of 500 to 1500. Like all graphs, there are outliers located in this graph as well. We can see four visible outliers located at the (approximate) values of (600, 2200), (1200, 2000), (1600, 2000), and (2000, 1900). Apart from these outliers, there is still no visible trend between the other data points.

AW CP Model



This scatterplot represents the AW CP model, and there is a slight positive correlation that we can notice between the predicted and actual salaries for the players in this division. Unlike the other graphs, this one doesn't necessarily have a cluster of data points near one location, but the range is not as large as the other graphs. For this scatterplot, most of the data values are located between 400 to 1300. There are three visible outliers located at the following (approximate) points: (1500, 1000), (1600, 1800), and (1800, 1500). Overall, we can say that there is a slight, positive, linear relationship between the two variables of predicted versus actual salaries for the AW division.

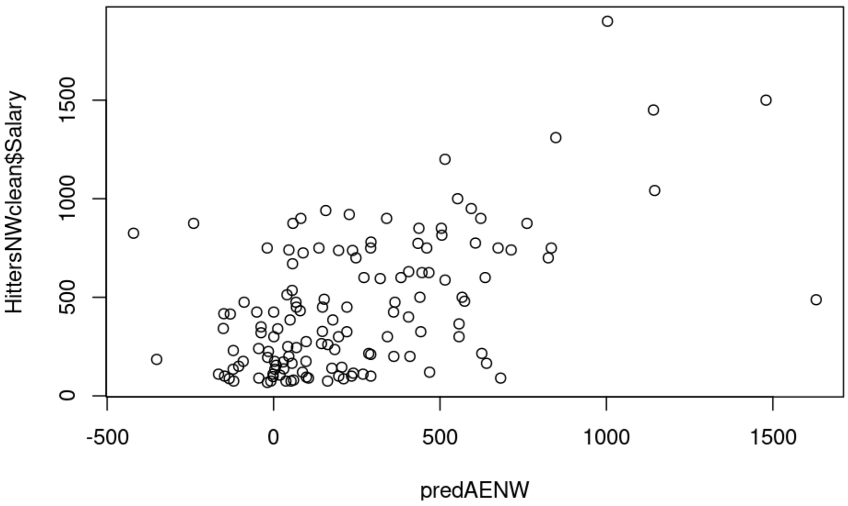
NW CP Model



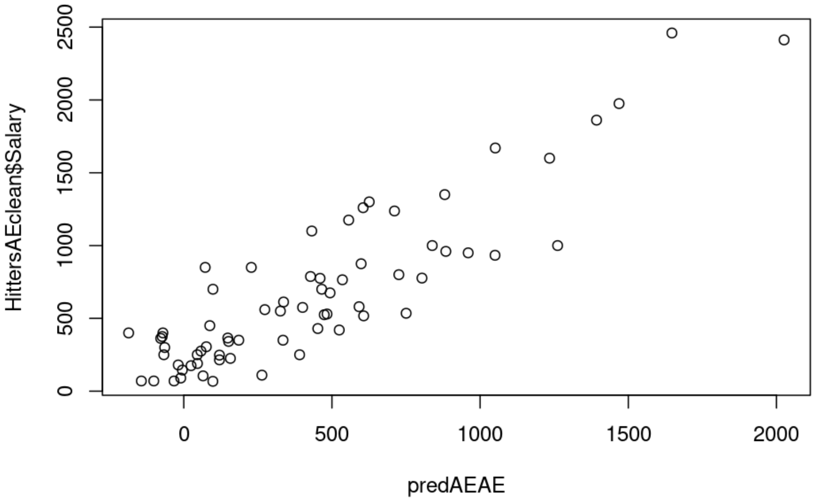
This scatter plot represents the NW CP model between the actual and predicted salaries for the players in the NW division. A large majority of the data values are located at the lower range of the scatterplot, and there is even some overlap in that range. As the x and y values get larger, the data values start becoming more spread out. There is less overlap between the data values and more distance between them. We can see one visible outlier at the (approximate) location (1700, 1400). Certain data values are further away from the majority of the points, so while they might not be completely far, they can be considered far enough to be considered outliers. Overall, we can say there is a weak, but positive correlation between the predicted values and the actual salary values for the players in the NW division.

Step 3: Four prediction plots for each division

AE Coefficient against all the Divisions (4 plots)

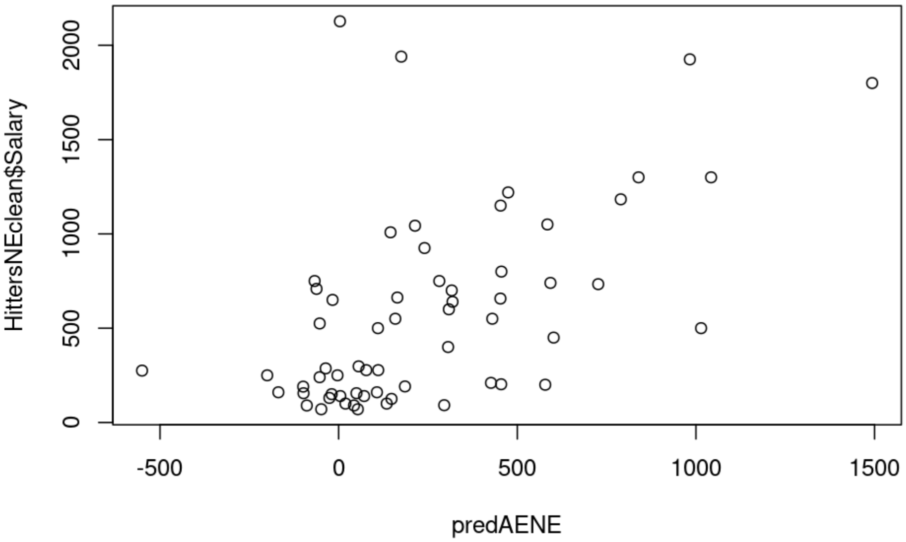


Comparing the data for the A League and E division using the N League and W division coefficient, against the N League and W Division Salary, there is a slight positive trend as there is values throughout about -200 to 500 all have points with the same salaries, but on average, the salaries increase as the predAENW values increase. As the predicted data for AENW increases, the salary for the N League and W Division increases. This trend holds especially true for the X values before 1000. However, towards the higher X values, the Salary is more unpredictable. For example, an outlier is a point around 1600, where though the predictor data is high, the Salary is still just as low as it is around where the predictor data is 0.

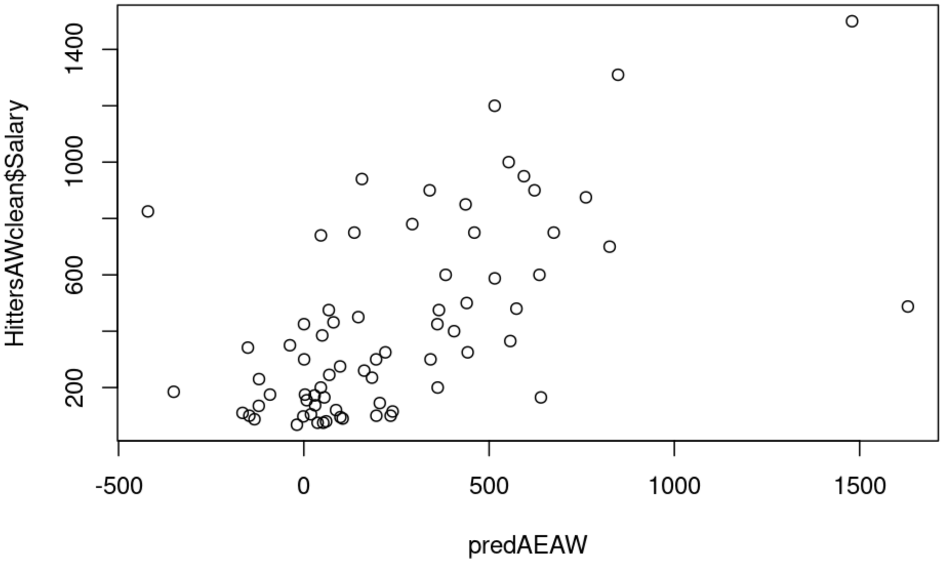


This plot compares AE performance with the AE coefficient where A is the League, and E is the Division. It is very clear that there is a positive and linear correlation between predAEAE and the Salary for the A-League and E Division. Throughout the range of values for predAEAE, the Salary seems to increase as the prediction values increase. There is a cluster around the predAEAE value of

0, but from around 250 and beyond, the Salary increases clearly, whereas a higher predAEAE value assumes a high Salary, on average. This plot is very close to being perfectly linear.

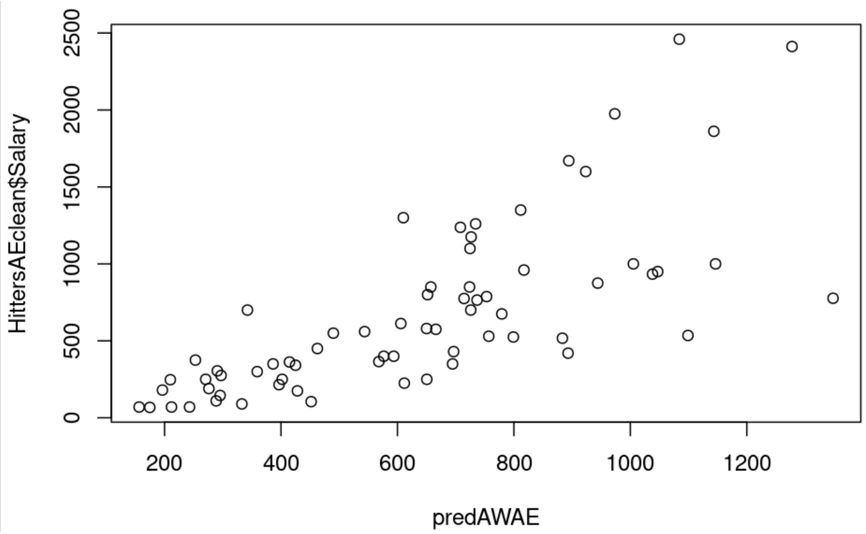


In the range from about -200 to 900 for predAENE, there is an increase in the Salary for the N League and E division as predAENE values increase. This indicates that higher predAENE values generally equate to higher salaries. However, the trend is not as clear around the value 0 as it is for values above 200 due to the cluster around the value 0. There are a couple of outliers, which I assume, are from rare circumstances or biases.

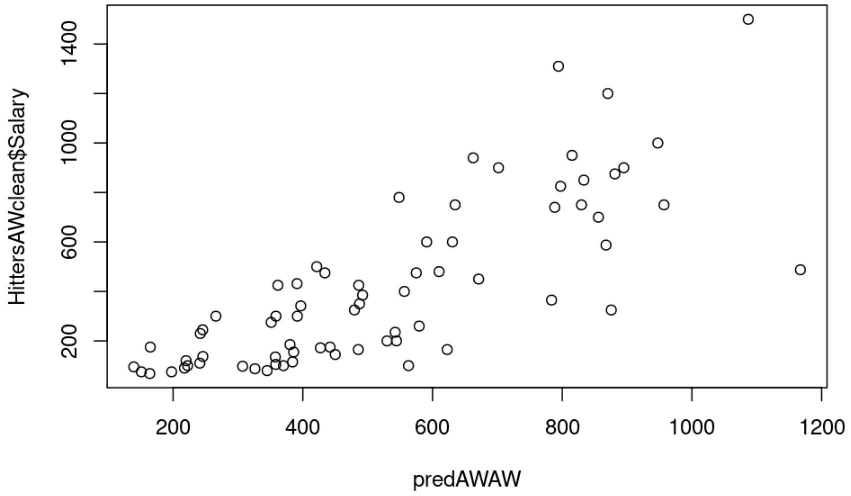


Again, it is evident that the Salary for the A-League and W Division increases as predAEAW values also increase. The salaries of those of predAEAW values of 0 or around that value seem to all earn similar salaries except for some. There seem to be clear outliers in the data, however. For example, for a predAEAW value of around -450, the Salary is the same as those who have a higher performance of around 0-500.

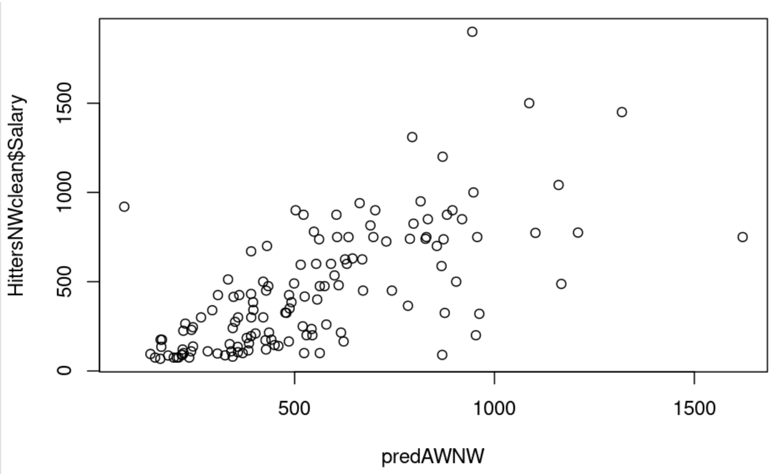
AW Coefficient against all the Divisions (4 plots)



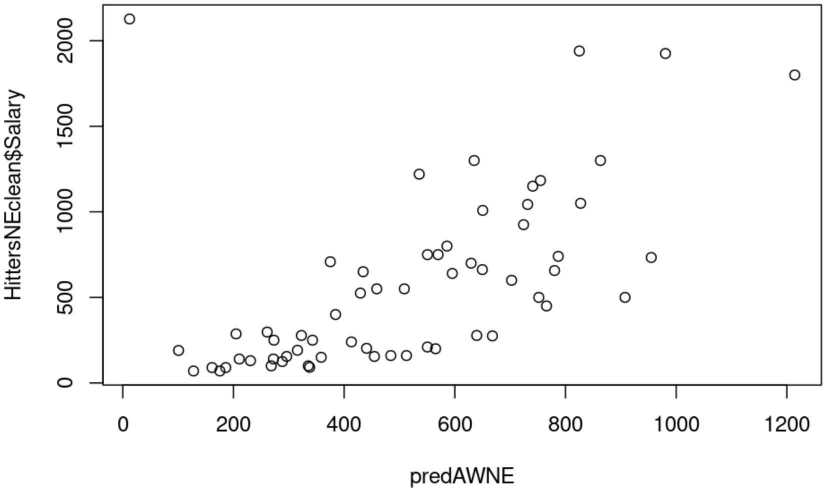
In this plot, where we compared the A division W league to the A division E league, we see a positive correlation that progressively gets weaker as the prediction and the salaries increase. We see more outliers as we reach higher Y values, notably the point at the furthest X value but at a relatively low salary. We see a wider spread as well.



Here, we see a similar trend when comparing the A division W league to the A division W league. We see a strong clustering at lower values and a wider spread as we increase, which detriments the linear trend at higher values. We are also noticing many isolated points as we increase our X and Y values.

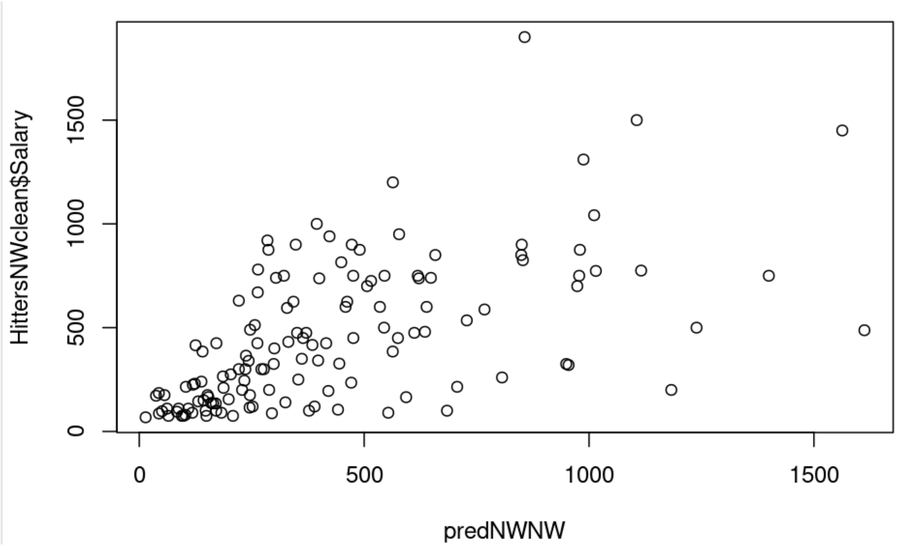


In this plot, we are comparing the A division w league to the N division W league. We see a clustering of data points at lower values, showing that those around 250 to 500 have very low salaries, but we don't see a strong linear trend. As values get larger, we see more spread and an increased appearance of outliers.

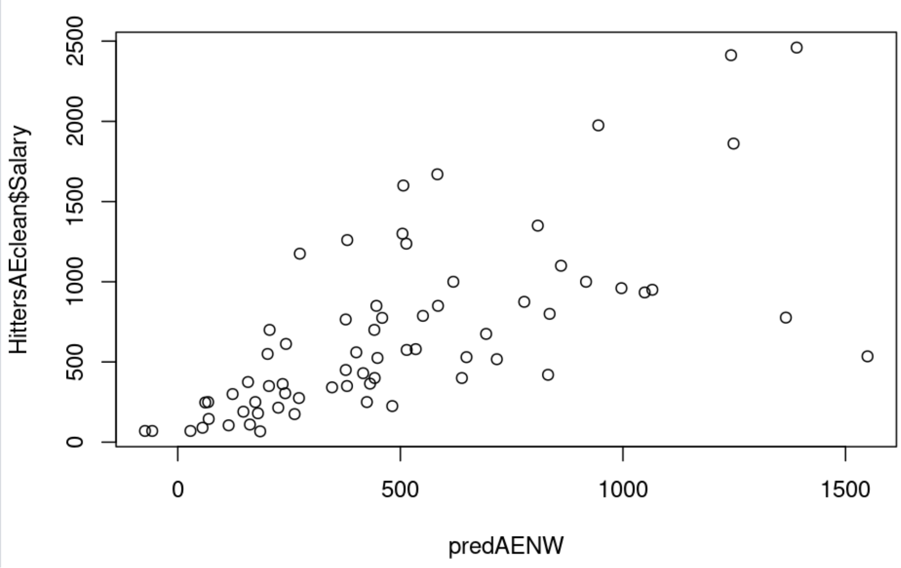


We are comparing the A division W league to the N division E league. We see a positive correlation between the salaries of the two divisions, but it appears as a clustering at salaries sub-1000 with X values in the hundreds. As we increase, we see three prominent outliers at high salaries. There is also a point to note with a prediction around 0 with a very high salary.

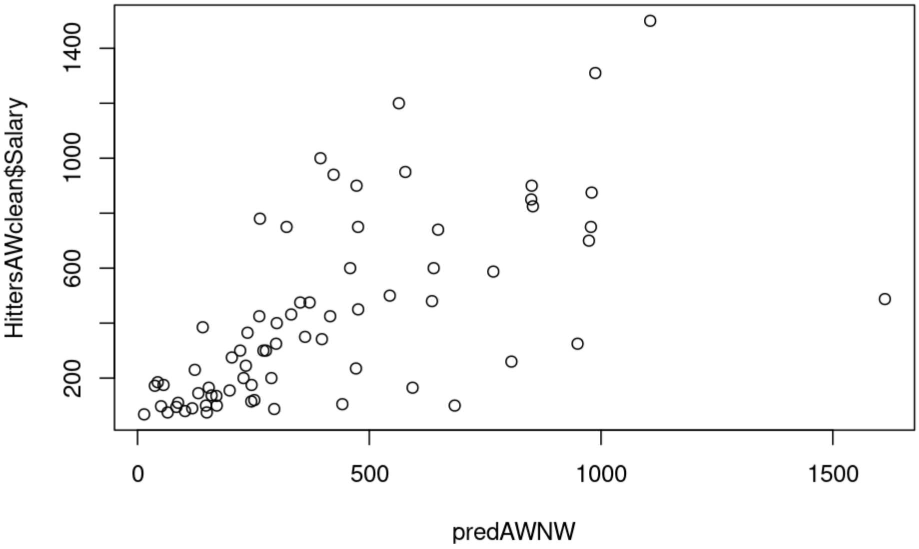
NW Coefficient against all the Divisions (4 plots)



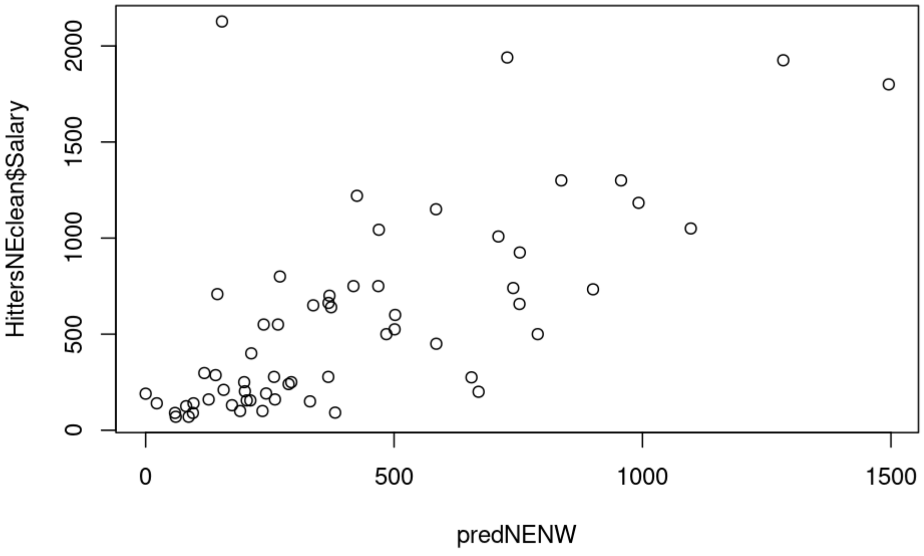
This plot compares the data for the N division and the W league against the N division and the W league coefficient. We see a lot of clustering towards the smaller salary values and lower predicted data values. However, overall, we are inclined to say that there is some sort of positive correlation between the predicted NWNW data and the NW salary. This is because if you were to draw a line of best fit, then it would be upward-sloping, possibly flattening out towards the higher predicted values. It is important to note that for x values between 1000 and 1500, salary tends to be more unpredictable, as we see a lot of random outliers that don't necessarily demonstrate a definitive trend.



This plot compares the data for the A division and the E league against the N division and the W league coefficient. Compared to the previous graph, we can see a more upward trend in this graph, at least until the prediction value of 1200. Essentially, what this positively correlated graph represents is that as the predicted values for AENW increase, so does the salary for the A division and E League. It is important to point out that the slope here is not steep, meaning that the ratio of rise over run is fairly small, and the increase in salary is not as significant when corresponding to the increase in predicted value.

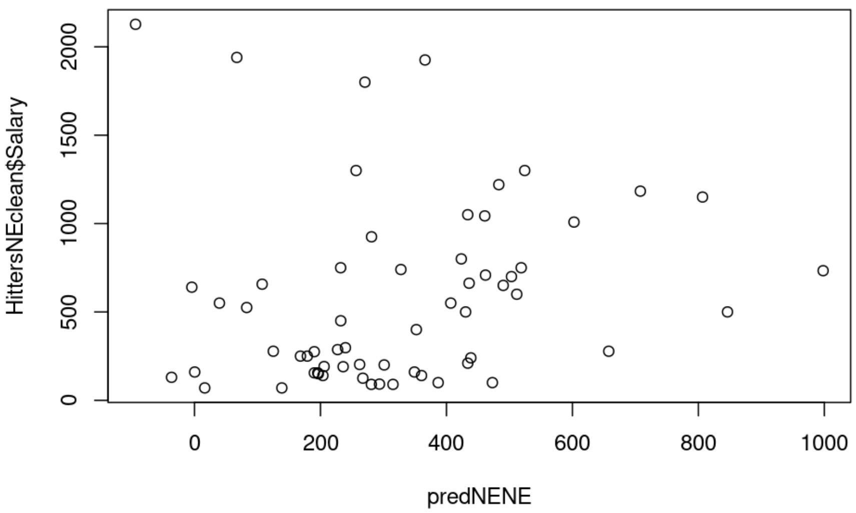


This plot compares the data for the A division and the W league against the N division and the W league coefficient. In this plot, we can see some clustering towards the prediction values from 0-300, and then after that, the data points start to spread a little more. Overall, a weak positive correlation is still evident in this graph, as the general trend is that as the predicted AWNW data values increase, so does the AW salary. A notable feature of this graph is the outlier that it produces at an x value of around 1650. We see that most of the correlated data points tend to be around around 1100, but this one data point is an outlier. It is intriguing to see why and makes us want to explore this subset of data even more closely.

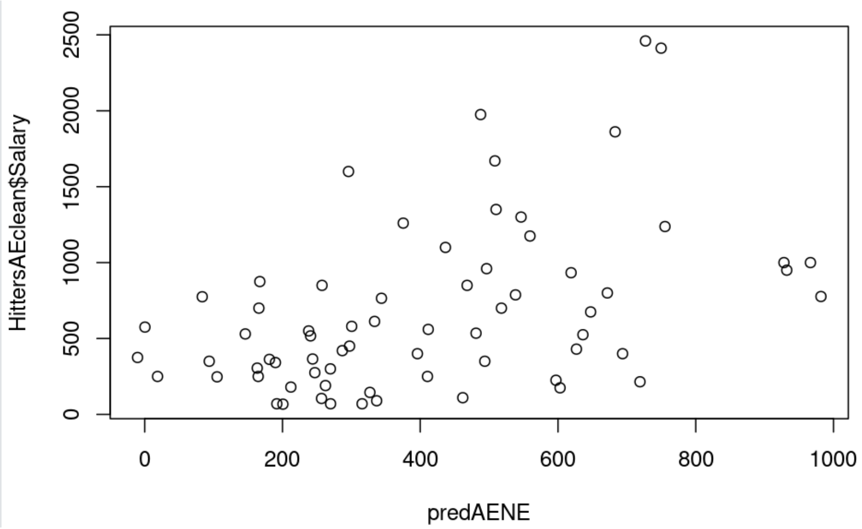


The last plot for the N division W league coefficient is this one, where it is compared against the N division and E league. Looking at this graph, we see a moderate positive correlation whereas the predicted NENW value increases, so does the salary. Some notable outliers are seen at the x value of about 200, where we see that the salary is the highest it ever reached, about approximately 2000. This is really surprising, and we wonder what caused this dramatic outlier and increase in salary at a predicted value of 200 when all of its neighboring values have a corresponding salary of around 500-1000.

NE Coefficient against all the Divisions (4 plots)

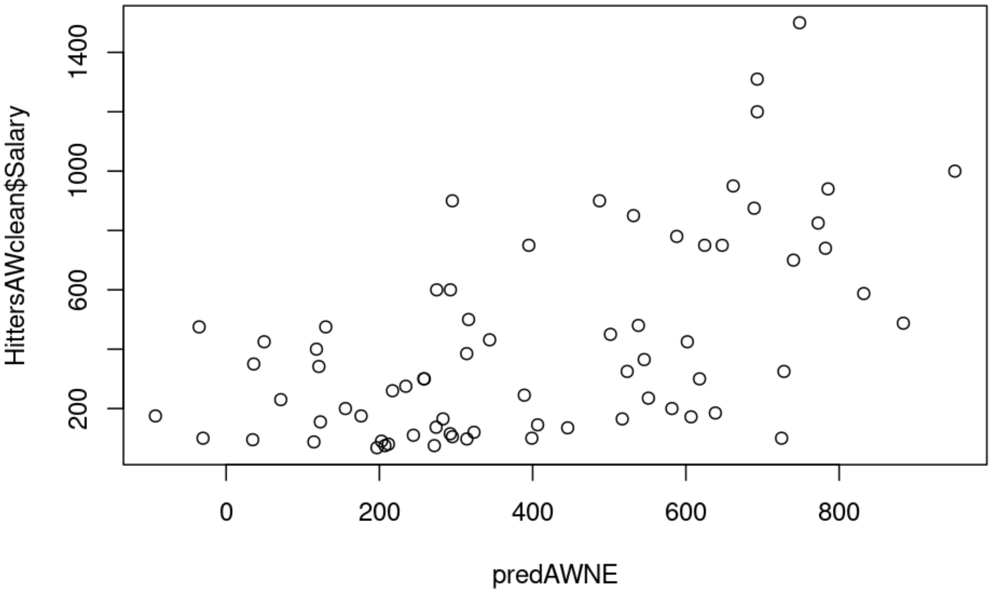


This plot compares the data for the N division and the E league against the N division and the E league coefficient. If there were to be a line of best fit, it would show neither a positive nor negative correlation. The data points seem randomly dispersed rather than clustered around a potential trend line. Therefore, if a line of best fit were added, it would likely be relatively flat, indicating neither a positive nor negative correlation between the x and y variables. Two outliers exist where x > 0 but has a low y-value compared to other points, and where x is around 1000, which has a higher y-value than surrounding data points. There is a cluster of data points between x-values of 200 and 400.

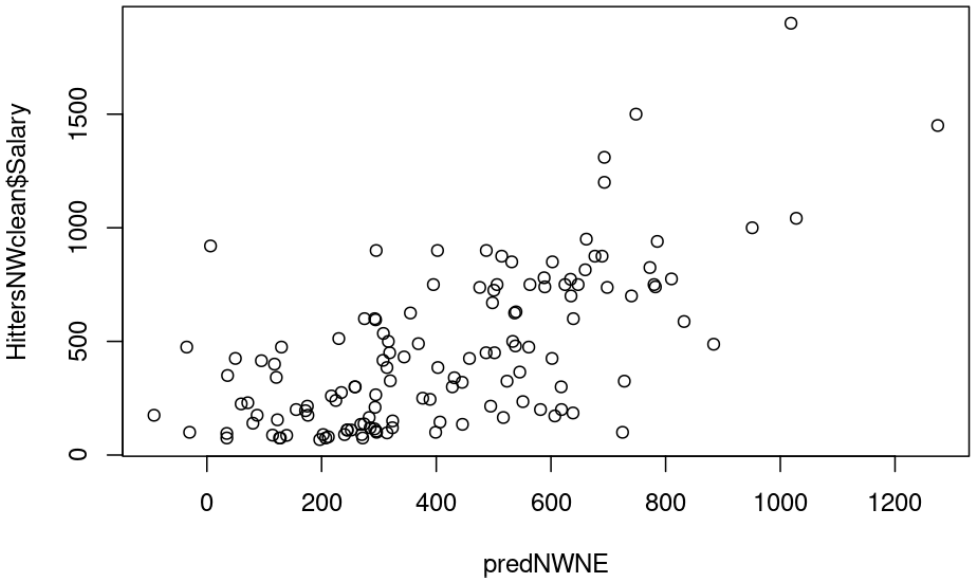


This plot compares the data for the A division and the E league against the N division and the E league coefficient. If there were to be a line of best fit, it would show a weak positive correlation. As the A division and E league coefficients increase, the N division and E league coefficients also tend to increase slightly. However, this positive relationship is quite weak, as evidenced by the wide dispersion of data points around the implied trend line. One outlier is the point where x is a little below 1000, which has a much higher y-value than surrounding points. The other potential outlier is the point where x is below 800. There are no real clusters of data points within this scatterplot.

The points are relatively evenly dispersed throughout the plot area. This lack of clustering suggests there is not a strong relationship or predictability between the two variables at any specific value.

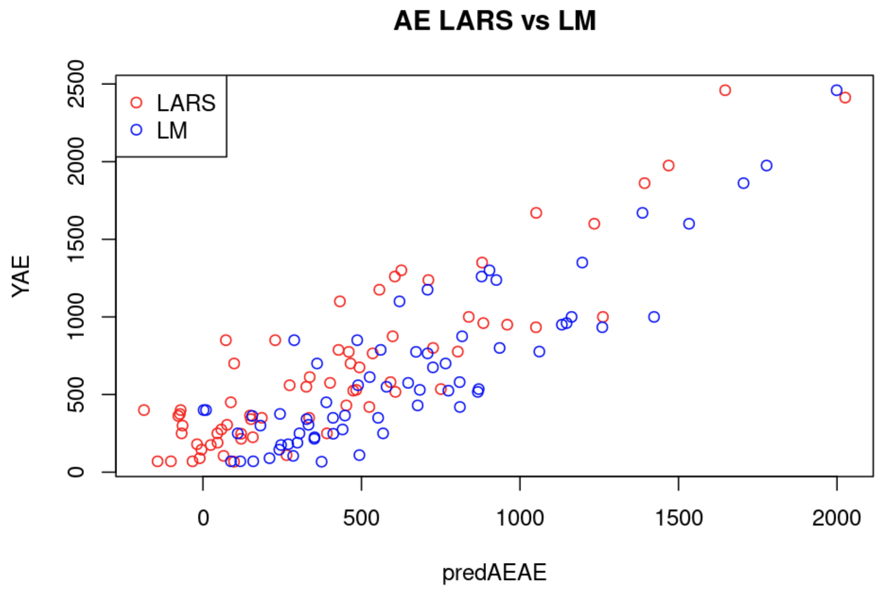


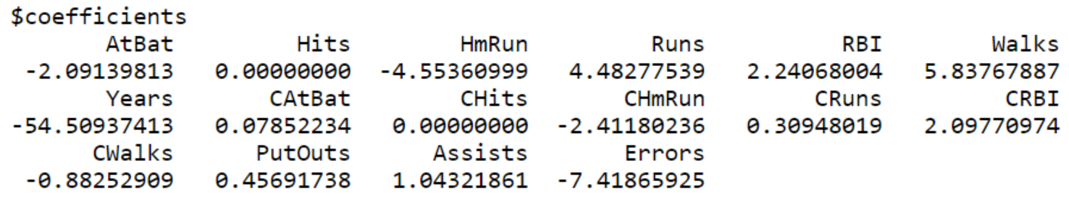
This plot compares the data for the A division and the W league against the N division and the E league coefficient. If there were to be a line of best fit, it would show a moderately positive correlation. Upon visual inspection, there appears to be a moderately positive correlation between the two variables. As the A division and W league coefficients increase, the N division and E league coefficients also tend to increase. The data points are dispersed, indicating there is variability in the strength of the correlation. The correlation does not appear to be extremely strong. There is a slight clustering of data points around x=200. This cluster is not very dense compared to some of the other scatterplots, meaning there are not substantially more observations around x=200 relative to other x-values.



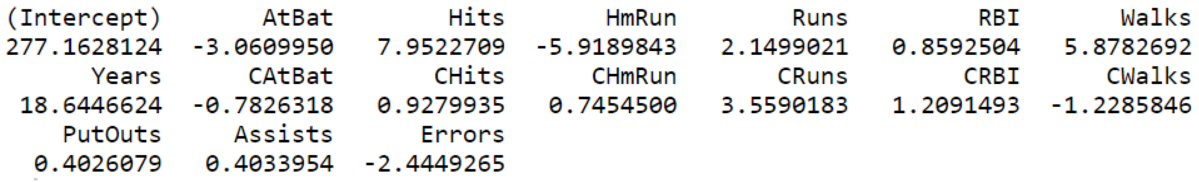
This plot compares the data for the N division and the W league against the N division and the E league coefficient. If there were to be a line of best fit, it would show a strong and positive correlation. As the N division and W league coefficients increase, the N division and E league coefficients tend to increase as well. This is evidenced by the general upward slope from left to right of the data points. There are a couple of outliers where x is a little above 1000 and where x is above 1200. There is also a visible cluster of points between the points 200 and 400 on the x-axis and below the 500 mark on the y-axis. The cluster suggests more observations exist with coefficient values in the 200-400 and below 500 range.

Lars Vs. Lm model for AE





```
$coefficients
        AtBat          Hits        HmRun         Runs          RBI        Walks
  -2.09139813    0.00000000  -4.55360999   4.48277539   2.24068004   5.83767887
        Years        CAtBat        CHits       CHmRun        CRuns         CRBI
 -54.50937413    0.07852234   0.00000000  -2.41180236   0.30948019   2.09770974
       CWalks       PutOuts      Assists       Errors
  -0.88252909    0.45691738   1.04321861  -7.41865925
```

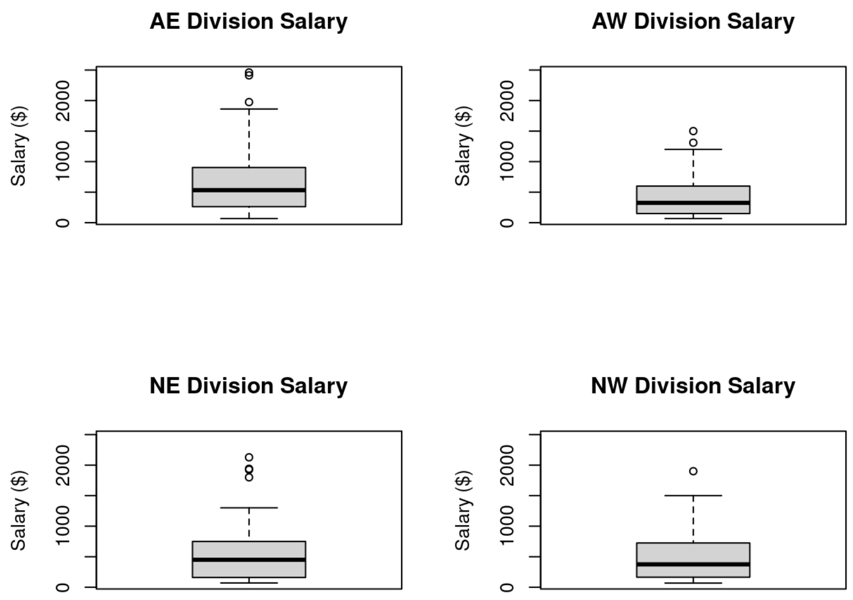


```
(Intercept)       AtBat        Hits      HmRun        Runs         RBI       Walks
277.1628124  -3.0609950   7.9522709 -5.9189843   2.1499021   0.8592504   5.8782692
      Years      CAtBat       CHits     CHmRun       CRuns        CRBI      CWalks
 18.6446624  -0.7826318   0.9279935  0.7454500   3.5590183   1.2091493  -1.2285846
    PutOuts     Assists      Errors
  0.4026079   0.4033954  -2.4449265
```

Compare the prediction (eg. just AElars vs AElm) and the coefficients. What have the coefficients done between Lars and LM?

When comparing the predictions, the trend for the LM model is slightly lower than the LARS model. They both have strong positive correlations, and there are more data points and clusters below 500 on the x-axis. Upon looking, the slopes seem to be around the same, but the salary for the LM model is lower, on average. The coefficients are very similar for most of the categories, except for the Hits value being 0 for the LARS plot and 7.95 for the LM plot, the Walks value being 2.10 for the LARS plot and -1.23 for the LM plot, so there is a positive correlation versus negative correlation, along with a couple of other categories with varying coefficient values, such as for Years, CRuns, CHmRun, and Errors.

Extra Credit Boxplots



<u>The Relevant Discussion:</u>

The American East (AE) Division's high outlier salaries compared to the other divisions could be attributed to the fact that it contains several historically successful and popular baseball teams like the Boston Red Sox, New York Yankees, and Toronto Blue Jays. These big market teams likely have higher revenues that allow them to spend more on top player salaries, driving up the outlier salaries and average salary for the whole AE Division. The American West (AW) Division having the lowest high outlier salaries of all divisions makes sense, given it does not contain the same level of historically successful and popular teams. The division lacks the big market teams that would drive up top salaries, so its maximum salaries are lower.

The popularity and success of teams like the Red Sox, Yankees, and Blue Jays allow them to command higher TV ratings, ticket sales, merchandise sales, and more. This higher revenue likely enables them to spend more on their players and salaries. It makes sense that the AE Division would then have higher outlier salaries and a higher average salary compared to other divisions without those high-revenue teams. The lack of historically dominant and popular teams in the AW Division leads to lower revenues and an inability to spend at the same level on top talent. This results in lower maximum salaries and a lower average salary for the division compared to the AE Division, which has those big market teams. The economic success and popularity of teams like the Red Sox, Yankees, and Blue Jays are key drivers of the higher outlier and average salaries in the AE Division compared to less prominent divisions like the AW. The lack of similar high-revenue teams in divisions brings down their salary metrics.