Using multiple linear regression to examine salaries of baseball players in 1986 1987

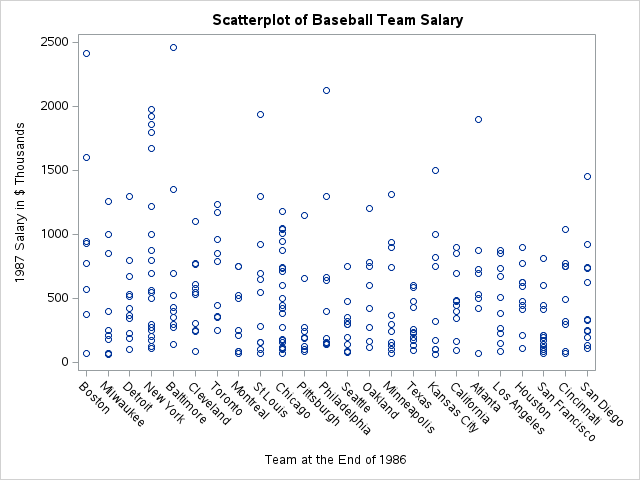
Data Analysis

**Introduction**

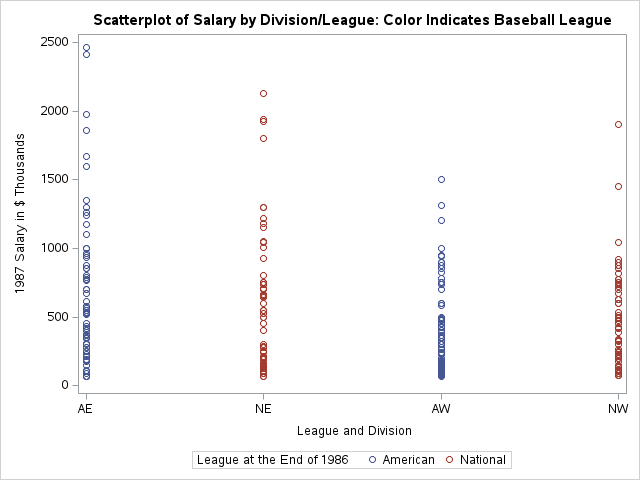
The goal of this data analysis is to understand which variables contributed to a baseball player’s salary from 1986 to 1987 and to create a multiple linear regression model that can sufficiently predict a player’s salary at this time based on these variables. These variables are: Team (Team at the End of 1986), nAtBat (Times at Bat in 1986), nHits (Hits in 1986), nHome (Home Runs in 1986), nRuns (Runs in 1986), nRBI (RBIs in 1986), nBB (Walks in 1986), YrMajor (Years in the Major Leagues), CrAtBat (Career Times at Bat), CrHits (Career Hits), CrHome (Career Home Runs), CrRuns (Career Runs), (CrRbi (Career RBIs), CrBB (Career Walks), League (League at the End of 1986), Division (Division at the End of 1986), Position (Position(s) in 1986), nOuts (Put Outs in 1986), nAssts (Assists in 1986), nError (Errors in 1986), Salary (1987 Salary in $ Thousands), Div (League and Division), logSalary (Log Salary). Note that LogSalary is the natural log.

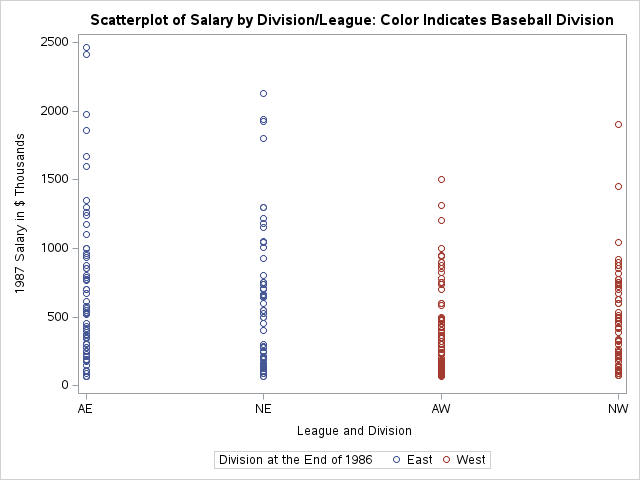
**Results of Analysis**

A scatterplot of salary for each team shows that while most players receive up to $1000 (in thousands) in salary, certain teams are able to pay select players much higher salaries, such as New York and St Louis. This suggests that the range of salaries might vary by team. This suggests that certain teams are able to pay more money to their players, likely because the teams are provided greater funds.

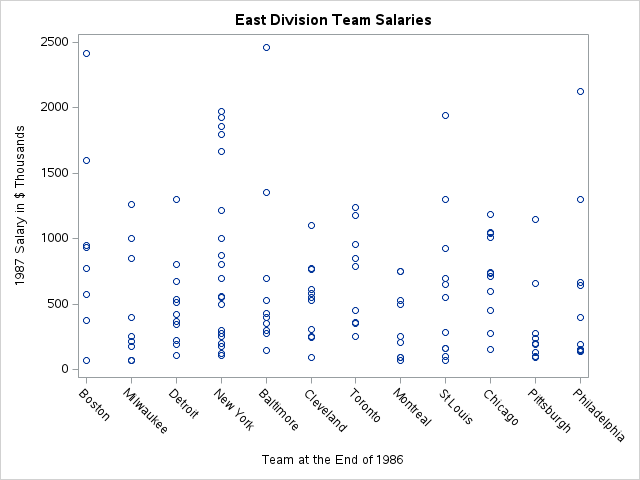


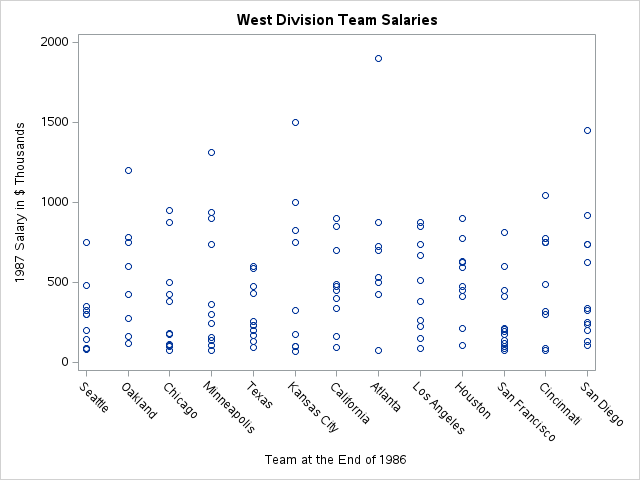
One probable cause for a difference in team funds might be due to differences in team location (baseball division). The funds that teams are able to raise is likely related to how much merchandise and many ball game tickets they are able to sell. Other related factors could be the baseball league. Similarly to how one company can offer a higher salary than another, one league might have more funds. To examine this possibility, two scatterplots of salary by div(division and league) were created, the first grouped by league, the second grouped by division. The results suggest that there is a significant difference in salary among division, but not among league.



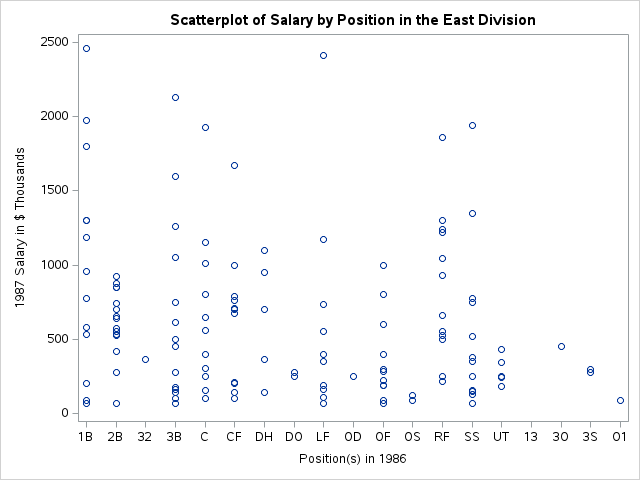


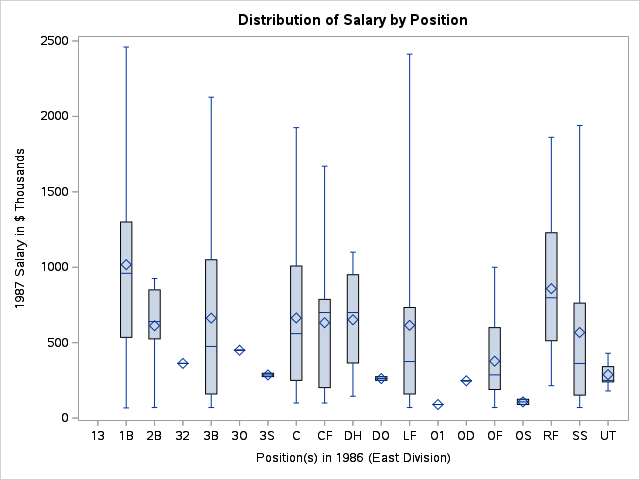
It is possible the variation in salary seen among teams is in fact due to a variation in salary for different divisions. To examine this, scatterplots were created for the teams in each division. We can see that the variation in salary among teams has decreased greatly when teams are compared only with other teams within the same division.

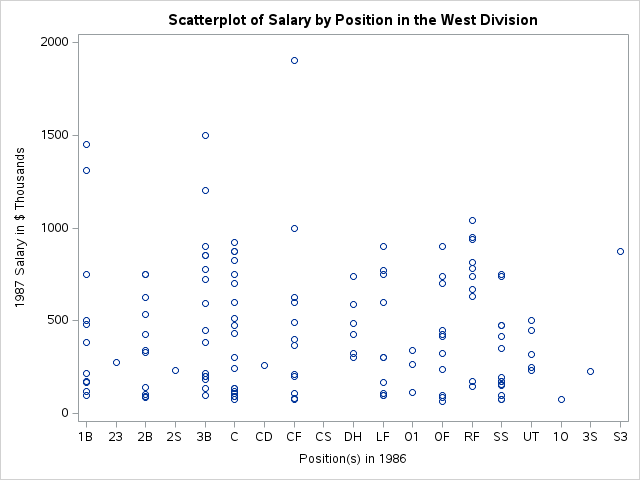


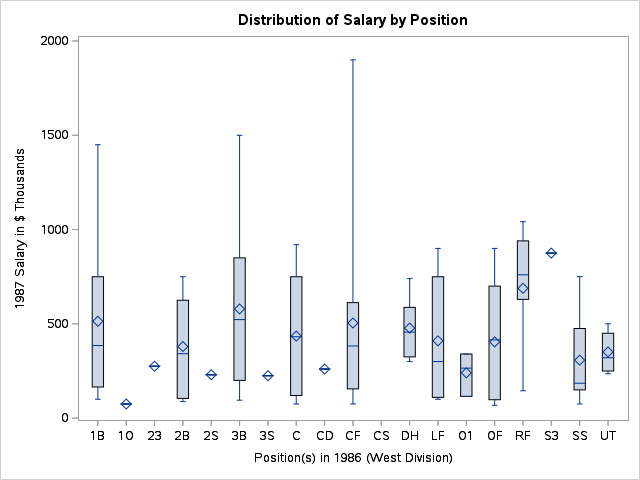


It is possible that within-team variation in salary might be due to differences in salary based on position. A scatterplot and a boxplot of salary vs. position(s) were created for the East and West divisions. There are multiple positions with few number of observations, with ten (East and West) containing only one observation. Because of this, a difference in salary for position might be better explained by other statistics (such as batting average) that should be correlated with position.



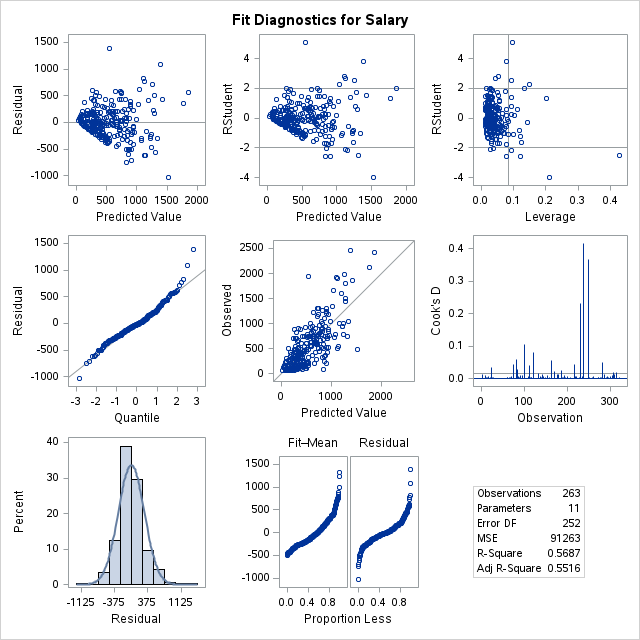


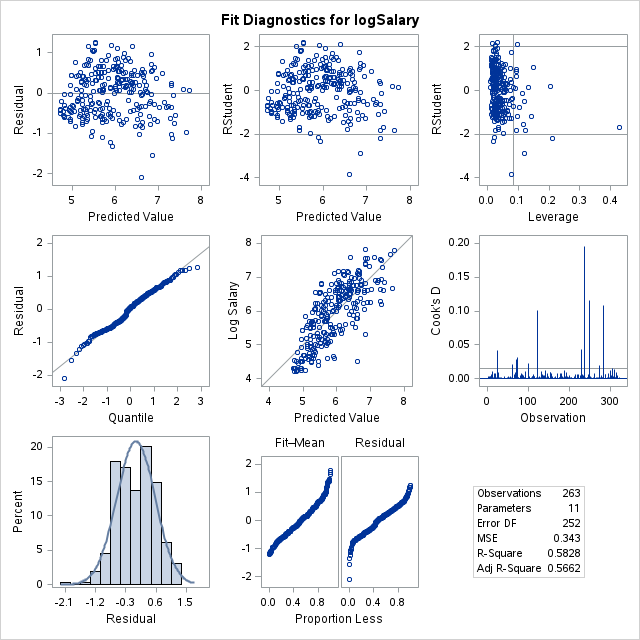




When making the regression model, we must determine if we will transform the variable salary by taking the log of salary. To examine this, we run two multiple regression models with division2, YrMajor, nAtBat, CrAtBat, nHome, CrHome, nRuns, CrRuns, CrHits, and nHits as independent variables and either salary or log salary as the independent variable. The purpose of this is simply to get an idea of the distribution of the residuals for Salary and LogSalary.

The graphs of Residual vs. Predicted Value show a clear indication of heteroscedasticity for both Salary and LogSalary, however, the trend is stronger for Salary than LogSalary. The Cook’s Distance appears more extreme for Salary than LogSalary. The QQ-plot for Salary appears slightly s-shaped.



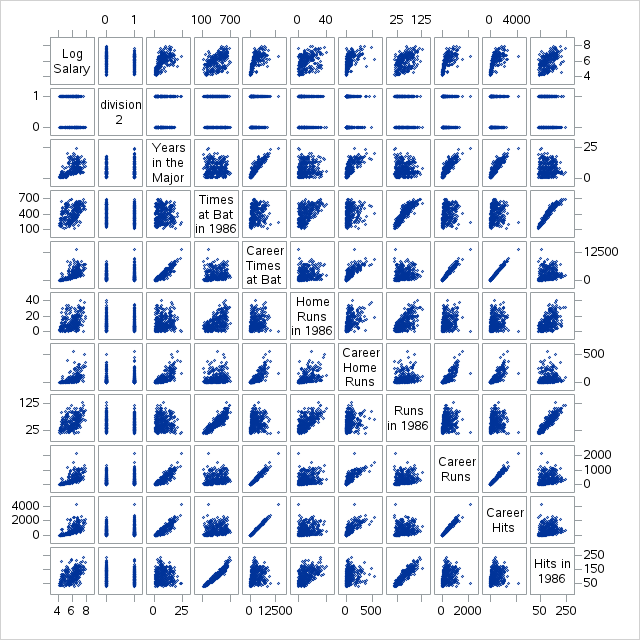


We next run a simple BoxCox transformation using div2 as the independent variable to examine the normality of the data. The selected lambda is 0, suggesting that a transformation of the natural logarithm should be used to stabilize the variance and make the data more normally distributed. Thus, we will use LogSalary as the dependent variable.



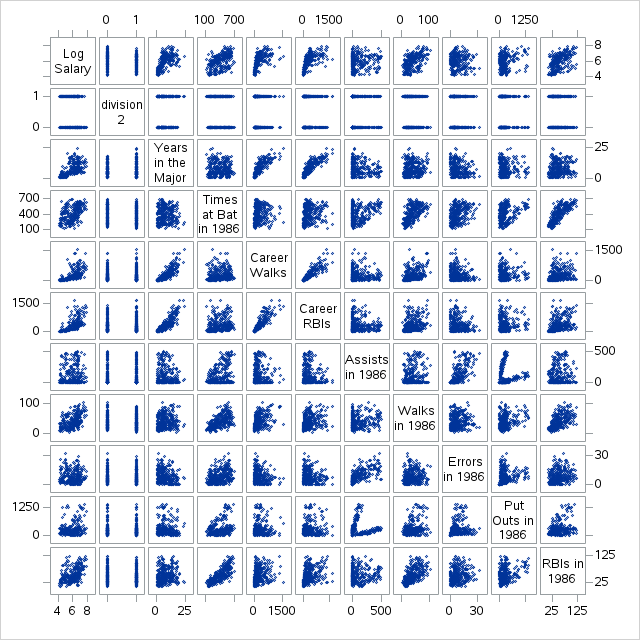
The remaining variables that so far have not been examined are likely to have a large correlation between them (such as CrRun and CrAtBat). Examining a matrix scatterplot and Pearson Correlation Coefficients, we see that nAtBat has a strong, significant positive association with nHit and nRun and a moderate, significant positive association with nHome. We also see that CrAtBat has a strong, significant positive association with CrHit, CrHome, and CrRun. Both of these results are likely because players with better batting statistics are at bat more often. CrAtBat also has a strong, positive association with YrMajor, undoubtedly because a player who has been in the major league longer will have had more times at bat in their career.

nAtBat and YrMajor will be in the model, while CrAtBat, CrHit, CrHome, CrRun, nHit, nRun, and nHome will not.



| **Pearson Correlation Coefficients**  **Prob > |r| under H0: Rho=0**  **Number of Observations** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **logSalary** | **division2** | **YrMajor** | **nAtBat** | **CrAtBat** | **nHome** | **CrHome** | **nRuns** | **CrRuns** | **CrHits** | **nHits** |
| **logSalary**  **Log Salary** | 1.00000    263 | -0.15010  0.0148  263 | 0.56436  <.0001  263 | 0.46183  <.0001  263 | 0.63801  <.0001  263 | 0.37124  <.0001  263 | 0.54641  <.0001  263 | 0.46268  <.0001  263 | 0.64709  <.0001  263 | 0.64598  <.0001  263 | 0.49233  <.0001  263 |
| **division2** | -0.15010  0.0148  263 | 1.00000    322 | 0.02602  0.6418  322 | -0.05280  0.3450  322 | 0.02700  0.6293  322 | -0.02927  0.6008  322 | -0.00752  0.8931  322 | -0.08738  0.1176  322 | -0.00287  0.9591  322 | 0.01758  0.7533  322 | -0.08248  0.1397  322 |
| **YrMajor**  **Years in the Major Leagues** | 0.56436  <.0001  263 | 0.02602  0.6418  322 | 1.00000    322 | -0.00848  0.8795  322 | 0.92066  <.0001  322 | 0.09768  0.0801  322 | 0.72149  <.0001  322 | -0.04267  0.4454  322 | 0.87947  <.0001  322 | 0.90352  <.0001  322 | -0.00803  0.8858  322 |
| **nAtBat**  **Times at Bat in 1986** | 0.46183  <.0001  263 | -0.05280  0.3450  322 | -0.00848  0.8795  322 | 1.00000    322 | 0.18556  0.0008  322 | 0.57307  <.0001  322 | 0.20577  0.0002  322 | 0.90260  <.0001  322 | 0.22315  <.0001  322 | 0.20568  0.0002  322 | 0.96447  <.0001  322 |
| **CrAtBat**  **Career Times at Bat** | 0.63801  <.0001  263 | 0.02700  0.6293  322 | 0.92066  <.0001  322 | 0.18556  0.0008  322 | 1.00000    322 | 0.20287  0.0002  322 | 0.79222  <.0001  322 | 0.14153  0.0110  322 | 0.98069  <.0001  322 | 0.99489  <.0001  322 | 0.17943  0.0012  322 |
| **nHome**  **Home Runs in 1986** | 0.37124  <.0001  263 | -0.02927  0.6008  322 | 0.09768  0.0801  322 | 0.57307  <.0001  322 | 0.20287  0.0002  322 | 1.00000    322 | 0.49270  <.0001  322 | 0.63965  <.0001  322 | 0.25413  <.0001  322 | 0.20179  0.0003  322 | 0.54165  <.0001  322 |
| **CrHome**  **Career Home Runs** | 0.54641  <.0001  263 | -0.00752  0.8931  322 | 0.72149  <.0001  322 | 0.20577  0.0002  322 | 0.79222  <.0001  322 | 0.49270  <.0001  322 | 1.00000    322 | 0.20649  0.0002  322 | 0.82093  <.0001  322 | 0.77573  <.0001  322 | 0.17398  0.0017  322 |
| **nRuns**  **Runs in 1986** | 0.46268  <.0001  263 | -0.08738  0.1176  322 | -0.04267  0.4454  322 | 0.90260  <.0001  322 | 0.14153  0.0110  322 | 0.63965  <.0001  322 | 0.20649  0.0002  322 | 1.00000    322 | 0.21660  <.0001  322 | 0.16193  0.0036  322 | 0.91167  <.0001  322 |
| **CrRuns**  **Career Runs** | 0.64709  <.0001  263 | -0.00287  0.9591  322 | 0.87947  <.0001  322 | 0.22315  <.0001  322 | 0.98069  <.0001  322 | 0.25413  <.0001  322 | 0.82093  <.0001  322 | 0.21660  <.0001  322 | 1.00000    322 | 0.98208  <.0001  322 | 0.22119  <.0001  322 |
| **CrHits**  **Career Hits** | 0.64598  <.0001  263 | 0.01758  0.7533  322 | 0.90352  <.0001  322 | 0.20568  0.0002  322 | 0.99489  <.0001  322 | 0.20179  0.0003  322 | 0.77573  <.0001  322 | 0.16193  0.0036  322 | 0.98208  <.0001  322 | 1.00000    322 | 0.21111  0.0001  322 |
| **nHits**  **Hits in 1986** | 0.49233  <.0001  263 | -0.08248  0.1397  322 | -0.00803  0.8858  322 | 0.96447  <.0001  322 | 0.17943  0.0012  322 | 0.54165  <.0001  322 | 0.17398  0.0017  322 | 0.91167  <.0001  322 | 0.22119  <.0001  322 | 0.21111  0.0001  322 | 1.00000    322 |

We examine the remaining unexamined variables (division2, CrBB, CrRBi, nAssts, nBB, nError, nOuts, nRBI) with nAtBat and YrMajor using the same method. We see that CrBB and CrRBi both have a strong, significant positive association with YrMajor, likely for the same reason as CrAtBat. YrMajor will be used in the model and not CrBB or CrRBi. nErrors is very strongly not associated with logsalary and will not be in the model. nAtBat has a strong, significant association with nRBI, a moderate, significant association with nBB and a weak, significant association with nAssts, and nOut. The strong and moderate associations likely have the same causes as previously listed (a player at bat with better statistics will be at bat more often), so nRBI and nBB will not be in the model. nAssts and nOut are both actions performed by fielders, and thus the association is likely because a fielder who plays more games will have more chances to perform these actions. Because the correlation is weak and because the final model should contain a variable that accounts for the statistics of outfielders, either nAssts or nOuts will be tested for the final model. However both nAssts and nOut should not be in the final model. The pattern of their matrix scatterplot suggests that there is some interaction in effect.



| **Pearson Correlation Coefficients**  **Prob > |r| under H0: Rho=0**  **Number of Observations** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **logSalary** | **division2** | **YrMajor** | **nAtBat** | **CrBB** | **CrRbi** | **nAssts** | **nBB** | **nError** | **nOuts** | **nRBI** |
| **logSalary**  **Log Salary** | 1.00000    263 | -0.15010  0.0148  263 | 0.56436  <.0001  263 | 0.46183  <.0001  263 | 0.57158  <.0001  263 | 0.62937  <.0001  263 | 0.04997  0.4197  263 | 0.46920  <.0001  263 | -0.02076  0.7375  263 | 0.22448  0.0002  263 | 0.48533  <.0001  263 |
| **division2** | -0.15010  0.0148  263 | 1.00000    322 | 0.02602  0.6418  322 | -0.05280  0.3450  322 | -0.00833  0.8817  322 | 0.00984  0.8604  322 | -0.01388  0.8041  322 | -0.06768  0.2259  322 | -0.02312  0.6794  322 | -0.00563  0.9198  322 | -0.09138  0.1017  322 |
| **YrMajor**  **Years in the Major Leagues** | 0.56436  <.0001  263 | 0.02602  0.6418  322 | 1.00000    322 | -0.00848  0.8795  322 | 0.83299  <.0001  322 | 0.86822  <.0001  322 | -0.09730  0.0813  322 | 0.10870  0.0513  322 | -0.18612  0.0008  322 | -0.00995  0.8588  322 | 0.11291  0.0429  322 |
| **nAtBat**  **Times at Bat in 1986** | 0.46183  <.0001  263 | -0.05280  0.3450  322 | -0.00848  0.8795  322 | 1.00000    322 | 0.12337  0.0269  322 | 0.20317  0.0002  322 | 0.35758  <.0001  322 | 0.63578  <.0001  322 | 0.35789  <.0001  322 | 0.34395  <.0001  322 | 0.80281  <.0001  322 |
| **CrBB**  **Career Walks** | 0.57158  <.0001  263 | -0.00833  0.8817  322 | 0.83299  <.0001  322 | 0.12337  0.0269  322 | 1.00000    322 | 0.88500  <.0001  322 | -0.04550  0.4158  322 | 0.41501  <.0001  322 | -0.14027  0.0117  322 | 0.04573  0.4134  322 | 0.23648  <.0001  322 |
| **CrRbi**  **Career RBIs** | 0.62937  <.0001  263 | 0.00984  0.8604  322 | 0.86822  <.0001  322 | 0.20317  0.0002  322 | 0.88500  <.0001  322 | 1.00000    322 | -0.09126  0.1021  322 | 0.28955  <.0001  322 | -0.12324  0.0270  322 | 0.10088  0.0706  322 | 0.37475  <.0001  322 |
| **nAssts**  **Assists in 1986** | 0.04997  0.4197  263 | -0.01388  0.8041  322 | -0.09730  0.0813  322 | 0.35758  <.0001  322 | -0.04550  0.4158  322 | -0.09126  0.1021  322 | 1.00000    322 | 0.14036  0.0117  322 | 0.70635  <.0001  322 | -0.02520  0.6523  322 | 0.09887  0.0765  322 |
| **nBB**  **Walks in 1986** | 0.46920  <.0001  263 | -0.06768  0.2259  322 | 0.10870  0.0513  322 | 0.63578  <.0001  322 | 0.41501  <.0001  322 | 0.28955  <.0001  322 | 0.14036  0.0117  322 | 1.00000    322 | 0.11260  0.0435  322 | 0.30121  <.0001  322 | 0.59070  <.0001  322 |
| **nError**  **Errors in 1986** | -0.02076  0.7375  263 | -0.02312  0.6794  322 | -0.18612  0.0008  322 | 0.35789  <.0001  322 | -0.14027  0.0117  322 | -0.12324  0.0270  322 | 0.70635  <.0001  322 | 0.11260  0.0435  322 | 1.00000    322 | 0.10974  0.0491  322 | 0.18245  0.0010  322 |
| **nOuts**  **Put Outs in 1986** | 0.22448  0.0002  263 | -0.00563  0.9198  322 | -0.00995  0.8588  322 | 0.34395  <.0001  322 | 0.04573  0.4134  322 | 0.10088  0.0706  322 | -0.02520  0.6523  322 | 0.30121  <.0001  322 | 0.10974  0.0491  322 | 1.00000    322 | 0.34861  <.0001  322 |
| **nRBI**  **RBIs in 1986** | 0.48533  <.0001  263 | -0.09138  0.1017  322 | 0.11291  0.0429  322 | 0.80281  <.0001  322 | 0.23648  <.0001  322 | 0.37475  <.0001  322 | 0.09887  0.0765  322 | 0.59070  <.0001  322 | 0.18245  0.0010  322 | 0.34861  <.0001  322 | 1.00000    322 |

We next examine if either nAssts or nOuts should be added to the model. We do this by running two models with the variables previously determined should be included. The p-value of nAssts is not significant (p-value = 0.0747), while the p-value of nOuts is (p-value = 0.0332). The adjusted R-square value for the model with nOuts is also larger (0.5493 compared to 0.5469). Thus, we will include nOuts in the final model.

| **Root MSE** | 0.59854 | **R-Square** | 0.5538 |
| --- | --- | --- | --- |
| **Dependent Mean** | 5.92722 | **Adj R-Sq** | 0.5469 |
| **Coeff Var** | 10.09810 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | 4.06548 | 0.13465 | 30.19 | <.0001 |
| **division2** |  | 1 | -0.18424 | 0.07403 | -2.49 | 0.0135 |
| **YrMajor** | Years in the Major Leagues | 1 | 0.10407 | 0.00775 | 13.43 | <.0001 |
| **nAtBat** | Times at Bat in 1986 | 1 | 0.00305 | 0.00027783 | 10.97 | <.0001 |
| **nAssts** | Assists in 1986 | 1 | -0.00048923 | 0.00027340 | -1.79 | 0.0747 |

| **Root MSE** | 0.59696 | **R-Square** | 0.5562 |
| --- | --- | --- | --- |
| **Dependent Mean** | 5.92722 | **Adj R-Sq** | 0.5493 |
| **Coeff Var** | 10.07147 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | 4.05908 | 0.13438 | 30.21 | <.0001 |
| **division2** |  | 1 | -0.18440 | 0.07384 | -2.50 | 0.0131 |
| **YrMajor** | Years in the Major Leagues | 1 | 0.10547 | 0.00770 | 13.69 | <.0001 |
| **nAtBat** | Times at Bat in 1986 | 1 | 0.00268 | 0.00027370 | 9.81 | <.0001 |
| **nOuts** | Put Outs in 1986 | 1 | 0.00029803 | 0.00013918 | 2.14 | 0.0332 |

The final model contains division2 (p = 0.0131), YrMajor (p = <0.0001), nAtBat (p = <0.0001), and nOuts (p = 0.0332) as independent variables. The adjusted R-square is 0.5493, meaning that the model predicts 54.93% of the variability in log salary.

The final equation is:

Where Division is a dummy variable where 0 = East and 1 = West.

The intercept is 4.05908, meaning that for a player who has just joined the East Division of the major leagues and has not yet been at bat or had any put outs, their expected salary is 1000\*e^4.05908 = $57,921.00.

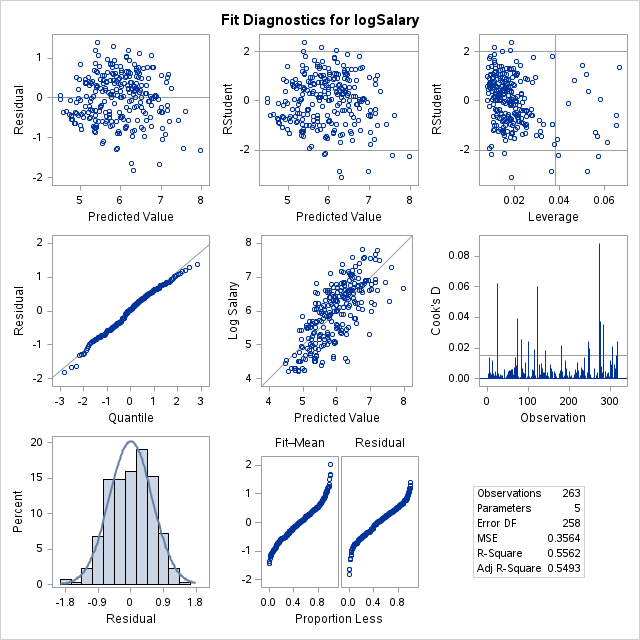
The regression coefficient for Division is -0.18440, meaning that a player in the West Division is expected to make $831.60 less than a player of the East Division.

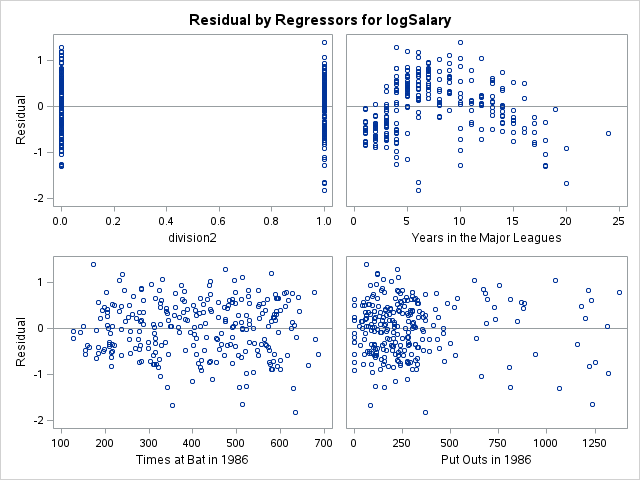
The regression coefficient for YrMajor suggests that when comparing two players of the same division, same nAtBat, and same nOuts, who have a difference of one year of playing in the major leage, the player who has been in the major league one year longer will make $1,111.23 more than the player who has been in the league one year less.

The regression coefficient for nAtBat suggests that when comparing two players of the same division, same YrMajor, and same nOuts, who have a difference of being at bat of one, the player who has been at bat one more time will make $1,002.68 more than the player who has been at bat one less time.

The regression coefficient for nOuts suggests that when comparing two players of the same division, same YrMajor, and same nAtBat, who have a difference in Put Outs in 1986 by one, the player with one more put out will have a salary of $1,000.30 more than the player with one less put out.

Examining the residuals, we see that the graph of Cook’s distance suggests the existence of three possible large outliers and twelve possible minor outliers. The Leverage graph shows sixteen points with high leverage. In the graph of Residual vs. Quantile, almost all points fall on the line, suggesting that the assumption of normality of residuals is met. The left side of the Fit-Mean/Residual plot is higher than the right, suggesting that a good amount of variation in log salary is explained by the model. As mentioned previously, there are signs of heteroskedasticity, suggesting that the assumption of equal variance of residuals is violated.





**Conclusion**

The adjusted R-square of the final model is 0.5493, meaning that the model predicts 54.93% of the variability in log salary. This is acceptably high, but there are still many ways that the model is lacking. The strongly correlated variables were suspected of conferring similar information to the model, and thus most were removed. In order to improve the predictive power of the model, variables outside those given in the dataset likely need to be considered. For example, would a baseball player considered ‘handsome’ sell a greater number of baseball cards, and thus be a more profitable inclusion to the team? Furthermore, In the future, it might be interesting to examine the outliers.

Another point of interest is that YrMajor appears to be only weakly linear with some large outliers, as better seen in the scatterplot bellow. It might be of interest to examine this variable further in the future. Perhaps there is some third variable to take into account in the relationship between YrMajor and LogSalary.

