Dependent Variable: Salary

Independent Variables: Length, G(Games), RBI(Total Bases), BB(Walks), DV1B(Dummy Variable) , DVSS(Dummy Variable), DVDH(Dummy Variable)

While selection of variable all the variables were analyzed and variables which did not give redundant information were picked based on limited domain knowledge and information gathering using search engines.

1. **Find outliers in X, outliers in Y and influential points. You do not need to remove them just list a few of the ones found and indicate how you determine the points were outliers and/or influential**

Outliers are identified based on the following table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Rule** |  |  |  | **Color Code** |
| X outlier | Hii > 2(k+1)/n = 0.069 | | |  | **14** |
| Y outlier | Rstudent > t(n-k-2) =1.9707 or < -1.9707 | | | | **17** |
| Influence | Cook's D > F(0.50,k+1,n-k-1) = 0.92 | | | | **0** |

Please note there are no influential points that could be detected using the Cook’s Distance. A possible failure to detect an influential point is due to the presence of more than one point together.

Variance Inflation factor for multicollinearity is under check as not all the variables have a VIF value greater than 10.

A complete list of outliers is attached in an excel file in the Appendix.

1. **Check for assumption violations**

**a. if necessary use transformation to fix any assumptions that you note. Explain why you did or not find any assumption violations.**

Ans- Assumptions violated are-

Salary-

* Normality – The Normality plot is not a straight line indicating the violation of the normality assumption, however as the sample size is 232 this violation would not have an effect.
* Independence -Data is sampled using RAND excel formula from the set of data. However, these is no information whether the data is recorded with respect to time or not. Hence, we cannot check the independence assumption.
* Equal Variance – Salary vs independent variables
  + Length – The variance assumption is getting violated where in there is slight consistency apart from few point between the range 4 to 8 length value but variance around the values 2 and 10 have a much lower variance and inconsistent variance as per the residual plot. Hence, the variation in the value of salary is not the same for the value of length.
  + G – The variance till game 50 is much less compared to variance greater than 50 which results in inconsistency in the residual plot
  + RBI – Variance below the RBI value of 25 is much less and compared to above 25 showing inconsistency.
  + BB - Apart from few outliers, the variation in the value of salary is same for the value of BB.
  + DV1b – The variance is not consistent between the 0 and 1 values for the DV1b dummy variable.
  + DVSS - The variance is not consistent between the 0 and 1 values for the DVSS dummy variable.
  + DVDH - The variance is not consistent between the 0 and 1 values for the DVDH dummy variable.
* Linearity – Salary vs Independent variable
  + Length –As per the Salary vs Scatter plot, data is widely spread and scattered and there is no evident straight line relationship built between them. Hence the average salary doesn’t have a first order relationship with Length.
  + G – As per the Salary vs Scatter plot, data is widely spread and scattered and there is no evident straight line relationship built between them. Hence the average salary doesn’t have a first order relationship with G.
  + RBI - As per the Salary vs Scatter plot, data is widely spread and scattered and there is no evident straight line relationship built between them. Hence the average salary doesn’t have a first order relationship with RBI.
  + BB - As per the Salary vs Scatter plot, data is widely spread and scattered and there is no evident straight line relationship built between them. Hence the average salary doesn’t have a first order relationship with BB.
  + DV1b – We cannot comment on linearity assumption for Dummy Variables
  + DVSS – We cannot comment on linearity assumption for Dummy Variables
  + DVDH – We cannot comment on linearity assumption for Dummy Variables

b**. check assumptions again after your transformation**

c. **comment on whether the assumption transformations helped or not. If they do not help, then revert back**

We did both logarithmic and SQRT transformation and objects plots for the same.

Although none of the transformation could satisfy all the assumption, but there is slight improvement in terms of linearity and variance in the SQRT transformation which makes the data look relatively stable.

The excel data and the SAS reports are attached in the appendix

1. **Using all-possible regression, reduce the list of possible models down to three, explaining your reasoning for choosing those three.**

We have come up with the following 3 models using SAS results on the basis of high R2 values and low variance K.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Index** | **Number in Model** | **R-Square** | **MSE** | **Variables in Model** |
|
| **8** | **2** | **0.6791** | **0.63279** | **SQRT(Length) SQRT(RBI)** |
| **29** | **3** | **0.6871** | **0.61961** | **SQRT(Length) SQRT(RBI) DVDH** |
| **64** | **4** | **0.6912** | **0.61423** | **SQRT(Length) SQRT(G) SQRT(RBI) DVDH** |

1. **Examine the three models in detail and choose one model. Explain why you choose that model and write down the least squares line for that model**

**Model 1: SQRT(Length) and SQRT(RBI)**

VIF: None of the variables are multi-collinear.

F Test – A F Test value of 242.29 indicates that average salary depends upon Length and/or RBI

T Test – T test result shows that both the variables individually impact the average salary keeping the other constant.

Residual Plots

* Length – Relatively showing equal variance and linearity except for few outliers
* RBI – Relatively showing equal variance and linearity but has many outliers which are scattered

R2 value is 0.6791 ie approx. 68 percent of the variance in Average Salary is due to variance the variation in Length and RBI

**Model 2: SQRT(Length) ,SQRT(RBI) and DVDH**

VIF: None of the vairables are multi-colinear

F Test – A F Test value of 166.92 indicates that average salary depends upon Length and/or RBI and/or DVDH

T Test – T test result shows that all the the variables individually impact the average salary keeping the others constant.

Residual Plots

* Length – Relatively showing equal variance and linearity except for few outliers
* RBI – Relatively showing equal variance and linearity but has many outliers which are scattered
* DVDH – Very few data point for the value of 1 showing un equal variance. Not enough data to support the use of this variable

R2 value is 0.6871 ie approx. 68.7 percent of the variance in Average Salary is due to variance the variation in Length, RBI and DVDH

**Model 3: SQRT(Length) SQRT(RBI) SQRT(G) DVDH**

VIF: None of the vairables are multi-colinear

F Test – A F Test value of 127.03 indicates that average salary depends upon either Length and/or RBI

T Test – T test result shows that Length, RBI and DVDH individually impact the average salary keeping the other constant. But incase of G the T test value is -1.73 which is less that -1.97 T-Test value at 95% significance and 231 DOF hence cannot reject the null of slope =0

Residual Plots

* Length – Relatively showing equal variance and linearity except for few outliers
* RBI – Relatively showing equal variance and linearity but has many outliers which are scattered
* DVDH – Very few data point for the value of 1 showing un equal variance. Not enough data to support the use of this variable
* G – The residual plot seems decent with value above 8 wherein the variance and linearity assumption seems to be met.

R2 value is 0.6912 ie approx. 69 percent of the variance in Average Salary is due to variance the variation in Length, RBI, G and DVDH

Considering all the three models we select model 1 as model 2 and model 3 increases the complexity but does not have a significant rise in the R2 value.

Also in case of model 2 DVDH residual plot does not adhere to the assumptions and in case of Model 3 the independent variable G is not significant as per the T Test results.

Least Square line for Model 1

SQRT(Salary) = -1.01984 +1.49311\*SQRT(Length) + 0.13262\*SQRT(RBI)

**Validation dataset questions:**

5. Using the least squares equation found in 4 from your training set, predict the values of the dependent variable in your validation data set. Average the absolute differences between the actual value and the predicted value. Discuss the average prediction you would expect in the future using this model.

Confidence interval for absolute error in the population is 1.666817683 to 3.515862368 with 95% confidence. We can say with 95% confidence that the value of salary is between 1. 666817683 and 3. 515862368 for an observation with values of the

6. Using the validation data set, see if the model found in 4 is useful.

Using the validation data set F test value of 69.58 shows that the model still holds significant for the validation data.