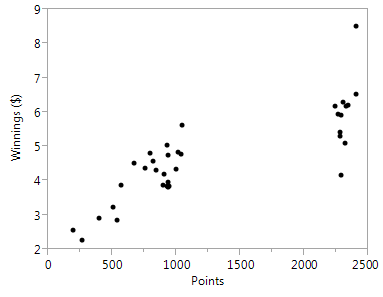
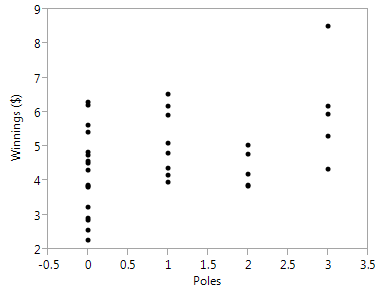
1.

**Graph 1a: Bivariate Fit of Winnings ($) by Points**



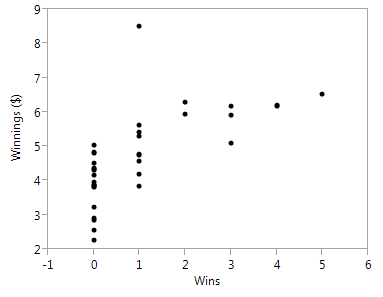
R square = 0.67

**Graph 1b: Bivariate Fit of Winnings ($) by Poles**



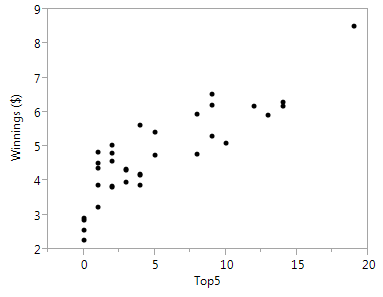
R square = 0.16

**Graph 1c: Bivariate Fit of Winnings ($) by Wins**



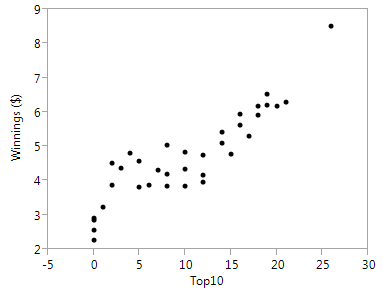
R square = 0.44

**Graph 1d: Bivariate Fit of Winnings ($) by Top 5**



R square = 0.74

**Graph 1e: Bivariate of Winnings ($) by Top 10**



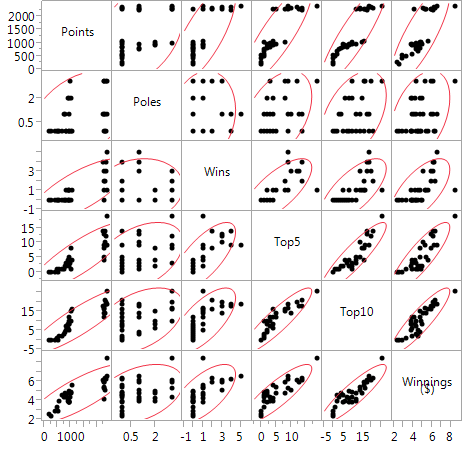
R square = 0.81

2.

Correlations

|  | **Points** | **Poles** | **Wins** | **Top5** | **Top10** | **Winnings ($)** |
| --- | --- | --- | --- | --- | --- | --- |
| Points | 1.0000 | 0.3887 | 0.7529 | 0.8496 | 0.8826 | 0.8165 |
| Poles | 0.3887 | 1.0000 | 0.1331 | 0.4373 | 0.4578 | 0.4059 |
| Wins | 0.7529 | 0.1331 | 1.0000 | 0.7252 | 0.6972 | 0.6615 |
| Top5 | 0.8496 | 0.4373 | 0.7252 | 1.0000 | 0.9017 | 0.8613 |
| Top10 | 0.8826 | 0.4578 | 0.6972 | 0.9017 | 1.0000 | 0.8977 |
| Winnings ($) | 0.8165 | 0.4059 | 0.6615 | 0.8613 | 0.8977 | 1.0000 |

**Scatterplot Matrix**



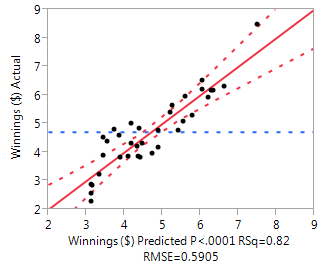
3.

Based on the graphs from question #1 and the correlation matrix from #2, I believe that we can use linear regression to predict winnings. The correlation between Winnings and the variables in general is pretty high, especially Top 10 (highlighted in the table) and Top 5.

Variable Top 10 appears to have the strongest relationship with Winnings (R square = 0.81). The graph and the analysis show that the relationship between Winnings and Top 10 is the strongest linear positive relationship.

4. Full Model

Actual by Predicted Plot



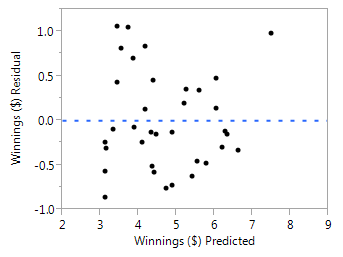
Summary of Fit table

|  |  |
| --- | --- |
| RSquare | 0.820813 |
| RSquare Adj | 0.789919 |
| Root Mean Square Error | 0.590496 |
| Mean of Response | 4.706286 |
| Observations (or Sum Wgts) | 35 |

**Parameter Estimates table**

| **Term** |  | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | 3.1134062 | 0.227236 | 13.70 | <.0001\* |
| Points |  | 7.0216e-5 | 0.000313 | 0.22 | 0.8238 |
| Poles |  | -0.015093 | 0.109149 | -0.14 | 0.8910 |
| Wins |  | 0.0033826 | 0.120972 | 0.03 | 0.9779 |
| Top5 |  | 0.0702731 | 0.05201 | 1.35 | 0.1871 |
| Top10 |  | 0.1126996 | 0.03869 | 2.91 | 0.0068\* |

**Residual by Predicted plot**



5.

I believe this model could be useful for predicting winnings due to its high R square (0.82). Even though it might be useful to predict Winnings, some variables are not good to include in the model. (Explain more below)

According to the Actual vs. Predicted chart, there’s a strong correlation between the model’s predictions and its actual results.

According to the Residual Plot, all the points are centered on the line 0.0 throughout the range of fitted values. The model is correct on average for all fitted values. These points cluster towards the 0 line and around the lower single digits of the Y-axis (0.25, -0.25, 0.5, -0.5), and they are pretty symmetrically distributed. There is not any clear patterns.

My concerns are:

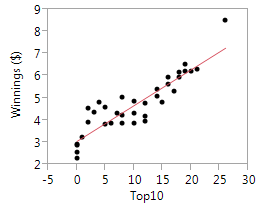
* There is some outliners that may affect the prediction
* Some plots exhibit “heteroscedasticity” that the residuals tend to be larger when the predicted values are between 3 and 4
* Maybe sometimes external environments (like condition of the car, the road, scoring machine, weather, driver’s age, and gender) affect the winnings possibility and cause the inconsistency/ outliers.
* There are 4 out of 5 variables (except Top 10) having p value > than alpha 0.05 so they failed the hypothesis test and they are not good variables so they should not be included in the model.

Based on all of these reasons, we can concluded that this model can be useful for predicting Winnings but due to high p value of some variables, it is not the optimal model.

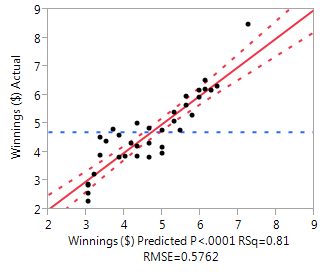
6.

New model:

**Regression Plot**

****

**Actual by Predicted Plot**

****

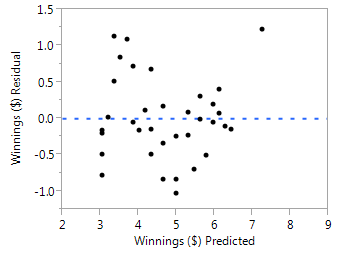
**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.805863 |
| RSquare Adj | 0.79998 |
| Root Mean Square Error | 0.576184 |
| Mean of Response | 4.706286 |
| Observations (or Sum Wgts) | 35 |

**Parameter Estimates**

| **Term** |  | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | 3.0508482 | 0.17173 | 17.77 | <.0001\* |
| Top10 |  | 0.1618445 | 0.013828 | 11.70 | <.0001\* |

**Residual by Predicted Plot**



Based on my observations, I have concluded each variable’s rank of significance:

1st: Top 10

2nd: Top 5

3rd: Points

4th: Wins

5th: Poles

I have removed all variables but kept Top 10, the R square does not change much and the residuals appeared to be about the same with before I removed other variables. It can be concluded that without other variables, the driver can still be able to win by having HIGH Top 10 finishes. Only Top 10 is kept because all the other variables have p-value > than alpha 0.05 thus they fail the hypothesis test and they are not good variables to be included in the model. Importantly, the new model (with Top 10 only) has higher R-square Adjusted (0.79998) than the full model (0.789919), which means the new term improves the model more than would be expected by chance. It can be concluded as the best model.

My concerns: Even though the new model meets all requirements (high R-square, R-square Adj, Correlation, p-value, etc.), sometimes it might be helpful to include other variables depends on the business’s needs. In fact, it would be not fair for a person who has high Top 10 finishes but less Top 5 to win than a person who also has same Top 10 finishes but higher Top 5. Therefore it seem to be the optimal model but it still has flaw like this. Also, Top 5 has the 2nd smallest p-value. Therefore, **Top 5 should be considered.**

7.

I would advise NASCAR driver to increase his Top 10 finishes as much as he can, and maybe Top 5 if possible. In this case, he could increase additional:

Winnings (in millions $) = 3.05 + (0.16)\* (number of Top 10 finish)

So based on this model, if he gets 1 additional Top 10 finish, he could earn 3.21 million of $ more.

8.

**Model using Top 10 + Top 5:**

| **Term** |  | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | 3.1361601 | 0.176082 | 17.81 | <.0001\* |
| Top5 |  | 0.0735057 | 0.045949 | 1.60 | 0.1195 |
| Top10 |  | 0.1167511 | 0.03126 | 3.73 | 0.0007\* |

Y = 3.136 + 0.0735057\* Top 5 + 0.1167511\* Top 10

= 3.136 + 0.0735057\* 13 + 0.1167511\* 22

= 6.66 (millions of $)

Conclusion: he should expect to win $6.66 millions