Paper 439-2013

Regression of NASCAR: Looking into Five Years of Jimmie Johnson

Yun Gao, California State University Long Beach; Alex Herrington, North Carolina State University; Luichun Yang, California State University Long Beach

ABSTRACT

In this paper, we investigate the winnings associated with different factors for NASCAR drivers. We want to predict the winnings that a driver can earn in a season given other, related factors, such as the number of races the driver competes in, the average finish position, or the make of car. We obtained 190 observations with 15 factors and randomly split the data into learning data and test data. Using the learning data set, we conducted multiple regression analyses to build a predictive model. Then we examined the final model with the test data set to see how well the model would work in the future. The model shows a high degree of accuracy in predicting the future.

INTRODUCTION

The goal of this paper is to model the monetary winnings of NASCAR drivers. In particular we decided to look at the five years Jimmie Johnson won the championship title. He won a championship in 2006, 2007, 2008, 2009, and 2010. Not all of the drivers competed in each year so we narrowed down our data set to only contain the drivers who had raced during all five years. We were left with 38 drivers which give us a sample size of 190. For the project we randomly divided the date set in half, one half was used to make a model for winnings and the other was used to validate our model. The variables in our data set are:

Winnings - amount of money a driver made during the season

Year - year for the season (2006, 2007, 2008, 2009, and 2010)

Driver - driver name

Points - points obtained by the driver in the season

Starts - number of races a driver started in the season

Poles - number of times a driver was first in the lineup of cars during the season

Wins - number of races a driver won in the season

Top 5 - number of times a driver finished in the top five during the season

Top 10 - number of times a driver finished in the top ten during the season

Make - brand of the racecar (Ford, Chevrolet, Dodge and Toyota)

Avg.st. - average starting position for the season

Avg. Fin. - average finish position for the season

Avg. Pos. - average race position for the season

Laps - number of laps completed during the season

LED - number of laps led during the season

Rating - a ranking calculated by NASCAR for each driver

Average Spots Gained - average number of spots gained during a race during the season

We decided to model winnings to see which variables have the largest effect on winnings. We hypothesized that there would be some obvious variables that influence winnings such as the number of races. We were curious to look at some of the other variables such as make of the car.

METHODS

In this section we are going to discuss the different types or regression models and their assumptions. We ran simple linear regression models for an exploration analysis to get an idea of what variables are significant predictors of winnings. Then we made our predictive model.

SIMPLE LINEAR REGRESSION

For a simple linear regression model there are only two parameters, β_0 and β_1. A simple linear regression model would be:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ \ (i = 1 \text{ to } n)$$

Where:

Y_i is the value of the response variable (winnings) in the ith trial

β_0 is the intercept parameter

β_1 is the coefficient parameter for the selected variable

X_i is the known value of the predictor variable in the ith trial

ε_i is the random error

When testing β_1 a t-test can be used and there are four assumptions that need to be checked. An F-test can also be used for β but would be the same as a t-test in simple linear regression. The assumptions that need to be assessed include normality, linearity, equal variance, and independence. For these simple models we are just going to examine residual plots and not conduct any formal tests. Normality can be assessed by looking at an Anderson Darling Statistic or normal probability plot. A p-value greater than 0.01 suggests that the data is normal. Linearity and equal variance can be checked by looking at the residuals versus fitted values plot. If the data has a random scatter of points the assumptions are valid. If there is a curve in the data linearity is violated. If there is a certain upward, downward or fan shape, equal variance is violated. Independence can be checked by looking at a Durbin Watson Tests Statistic or the residuals versus order plot. If the value is near two, or greater than one, the assumption of independence is valid.

A test for β_1 would have the following hypothesis

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0$$

Fail to reject the null hypothesis if the p-value is greater than or equal to 0.01 and reject the null hypothesis if the p-value is below 0.01. If the conclusion rejects the null hypothesis there is evidence that β_1 is an important predictor of the response variable Y. This simple test will be used in each of our simple linear regression models.

The R^2 value can also be looked at to see how well the model predicts Y. The closer the value is to one or 100% the better the model fits. Ideally we want an R^2 value above 70%.

MULTIPLE REGRESSION

Then we will consider the linear regressions with more variables and compare them with the simple linear regression. The multiple regressions can be written in the vector form as

$$Y = X\beta + \varepsilon$$

Where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & \cdots & x_{1,p-1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{n,p-1} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Y is the vector of responses

X is the matrix of constants

β is a vector of parameters

ε is a vector of independent normal random variables with expectation zero and variance-covariance matrix $\sigma^2 I$

Here, we firstly construct the regression model using selected variables from simple regression results. Highly correlated predictor variables may cause multicollinearity. To test whether the phenomena of multicollinearity exists, we compute the VIF values and plot the scatter plot matrix. We also conduct the residuals analysis and find out the response Winnings need to be transformed. The Box-Cox procedure suggests the transformation form. The result of Breusch-Pagan test and the high Pearson coefficient of correlation support the assumptions.

ANALYSIS/RESULTS

In this section we are going to present the results from the regression models. First we will talk about the simple linear regression models. Then we will go over our model used for predictions.

SIMPLE LINEAR REGRESSION

This section discusses each of our simple linear regression models. We decided to use earnings in a million dollars as the response variable. We decided to examine variables that we thought would be most interesting. All of the assumption discussions and Minitab output for this section can be found in the appendix. We wanted to perform these tests to see if the variables would be significant in the multiple regression model if they were significant in the simple linear regression models. From the simple linear regression models we found average finish, driving a Toyota, number of starts, and the transformation of the number of laps led to be significant predictors of the amount a driver wins. Now we will go into each individual simple linear regression model.

Winnings vs. Average Finish Position

First we start off with our model of winnings versus average finish position. We hypothesized that the lower an average finish, meaning average finish near one, would result in higher winnings. This is self explanatory because the better a driver finishes the more money he should make through sponsorships and winnings. Before we go into any analysis we wanted to view the data graphically to see if there is any indication of a linear relationship. Figure 1 displays a scatter plot of the relationship. Just by looking at the graph there appears to be a somewhat strong negative linear relationship. A driver makes more money when they finish closest to first. This outcome was expected and now we will investigate the significance of the variable average finish.

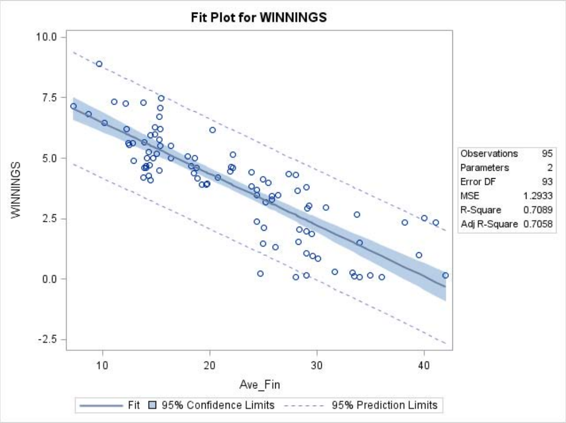


Figure 1 Scatter plot of winnings vs. Avg. Finish

The regression equation and model output can be seen in Table 1. The regression equation is

Winnings = 8.609 - 0.213*Average Finish

The slop β_1 here is negative, which means the smaller the average finish position is, the more winnings the driver gets, which we saw in the scatter plot earlier. For every position decrease in the average finishing position a driver will lose on average $212,702 ($0.213 million dollars).

Regression Equation				
WINNINGS = 8.609 – 0.213 Avg. Fin.				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Constant	8.60856	0.33372	25.80	<0.001
Average Finish	-0.21270	0.01413	-15.05	<0.001
R^2 = 70.9%				

Table 1 Regression of Winnings vs. Average Finish[①]

[①] For computation efficiency, we use “million dollars” as a unit in SAS program, same for other predictor variables.

Now that we have a regression equation we will see if the predictor variable is actually significant. Before a test on the β_1 can be run the assumptions need to be looked at. In the appendix we discussed the assumptions for the model and found that none were violated. The residual plot in Figure A1 of the appendix provides information the residuals and provides evidence for valid assumptions. Looking at the p-value for Average Finish, less than 0.001, we can conclude that average finish position is a significant predictor of winnings because the p-value is less than 0.01.

Lastly we can look at the R^2 value to see how well the model works. The R^2 is .709, which effectively proves that there is a great liner association between average finish position and winnings. From this simple linear model we can conclude that average finish position has a strong relation with winnings.

Winnings vs. Average Spots Gained

Next we decided to look at winnings versus average spots gained. We thought that the more a car moved round on the track the more the driver would make because fans would be excited to follow this driver. We started by making a scatter plot of the data and found that there may possibly be an outlier. The graph can be seen in Figure A2 of the appendix. After discussion we decided to remove the outlier thinking that it was a data entry error by NASCAR. The scatter plot with the removed point can be seen in Figure 2. The plot does not show a very strong linear relationship. The overlaid regression line is horizontal meaning that there is a poor linear relationship. We will look at a formal test to see if average spots gained is actually not significant like it appears in the graph.

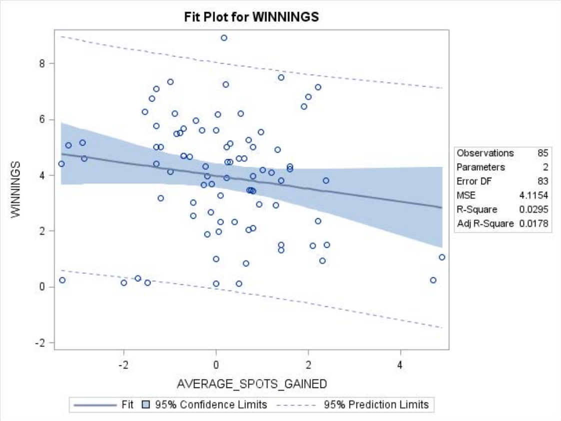


Figure 2 Scatter plot of Winnings vs. Average Number of Spots Gained

The regression equation and output can be seen in Table 2. The regression equation is

Winnings = 3.983- 0.234*Average Spots Gained

The regression equation and model output can be seen in Figure 2. The β_1 can be interpreted as, on average a driver is expected to lose $234,096 ($0.234million dollars) for every increase in average positions gained. Before we test the predictor variable we looked assumptions. The all appeared to be valid and can be read about in the appendix. The graph of the residuals used to validate the assumptions is Figure A3. From Table 2 the p-value of 0.116 suggests that average spots gained is not a significant predictor of winnings. This is different than what we expected to see. We were expecting more movement to lead to more money. We will still look at the R^2 value to see how the model performs.

<table>
<tr><td colspan="5">Regression Equation</td></tr>
<tr><td colspan="5">WINNINGS = 3.983 – 0.234 AVERAGE SPOTS GAINED</td></tr>
<tr><th>Predictor</th><th>Coefficient</th><th>Standard Error</th><th>T-Value</th><th>P-Value</th></tr>
<tr><td>Constant</td><td>3.983123</td><td>0.221068</td><td>18.02</td><td>&lt;0.001</td></tr>
<tr><td>Average Spots Gained</td><td>-0.23410</td><td>0.147302</td><td>-1.59</td><td>0.116</td></tr>
<tr><td colspan="5">R^2 = 3.0%</td></tr>
</table>

Table 2 Regression of Winnings vs. Average Spots Gained

We only obtained an R^2 value of 3.0% which is not much at all. This indicates the model does a poor job predicting winnings. Overall there is not a relationship between how much a driver makes and how much they move around positions on the race track.

Winnings vs. Starts

Next we decided to investigate winnings versus the number of starts. We thought the more races a driver started in the more money he would make. When we were making this model we ran into problems with the assumptions. A full discussion of the assumptions is in the appendix for Figure A4. We were unable to find a simple transformation to fix the assumptions so we decided to use the untransformed data. The regression equation for the model is

Winnings = -0.550 + 0.153*Starts

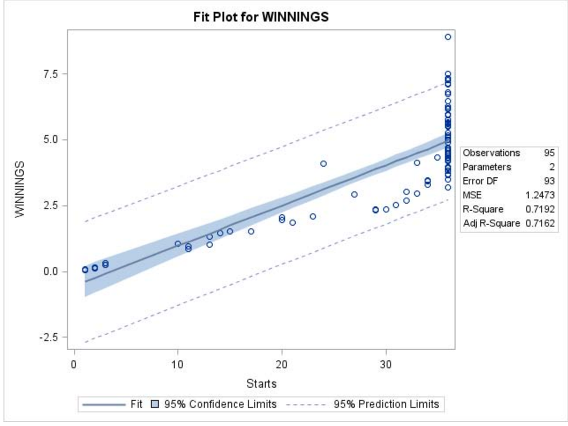


Figure 3 Scatter plot of Winnings vs. number of starts

A scatter plot of the data with the regression line is in Figure 3. The graph also contains confidence and prediction intervals.

The β_1 translates to, on average for every race a driver starts in their winnings goes up by $152,797 (0.153 million dollars). This makes sense because the more races a driver is in the more he has a chance to win and gets more publicity. Table 3 contains the statistics from the regression equation. We obtained a p-value less than 0.001 for our β_1. This indicates that the number of races a driver started in is significant to the amount of money they win.

<table>
<tr><td colspan="5">Regression Equation</td></tr>
<tr><td colspan="5">WINNINGS = -0.550 – 0.153*Starts</td></tr>
<tr><th>Predictor</th><th>Coefficient</th><th>Standard Error</th><th>T-Value</th><th>P-Value</th></tr>
<tr><td>Constant</td><td>-0.55027</td><td>0.31046</td><td>-1.77</td><td>0.0796</td></tr>
<tr><td>Starts</td><td>0.15280</td><td>0.00990</td><td>15.43</td><td><0.0001</td></tr>
<tr><td colspan="5">R^2 = 71.9%</td></tr>
</table>

Table 3 Regression of Winnings vs. Starts

For the model we obtained an R^2 value of 71.9%. This value indicates that this model does a good job of predicting the amount of winnings. In conclusion for this model, we found that it does a decent job predicting winnings from the number of races a driver starts in.

Winnings vs. Led

We also analyzed the number of laps led versus the amount of winnings. We hypothesized that the more laps a driver led the more he would make because he would score more points and more fans would follow him. When we first ran the model there was a problem with the normality assumption. This can be shown in Figure A5 of the appendix. We had to take the log of one plus the number of laps led. One was added because some drivers never led a lap. We did the log transformation because it is common practice to try a log transformation first because it is robust to fixing assumptions and usually fixes most assumptions. Fortunately for us the transformation worked. The assumptions for the new model were all valid and can be seen in Figure A6 of the appendix. The transformed regression equation is:

Winnings = 0.947 + 0.781*ln(1+Led)

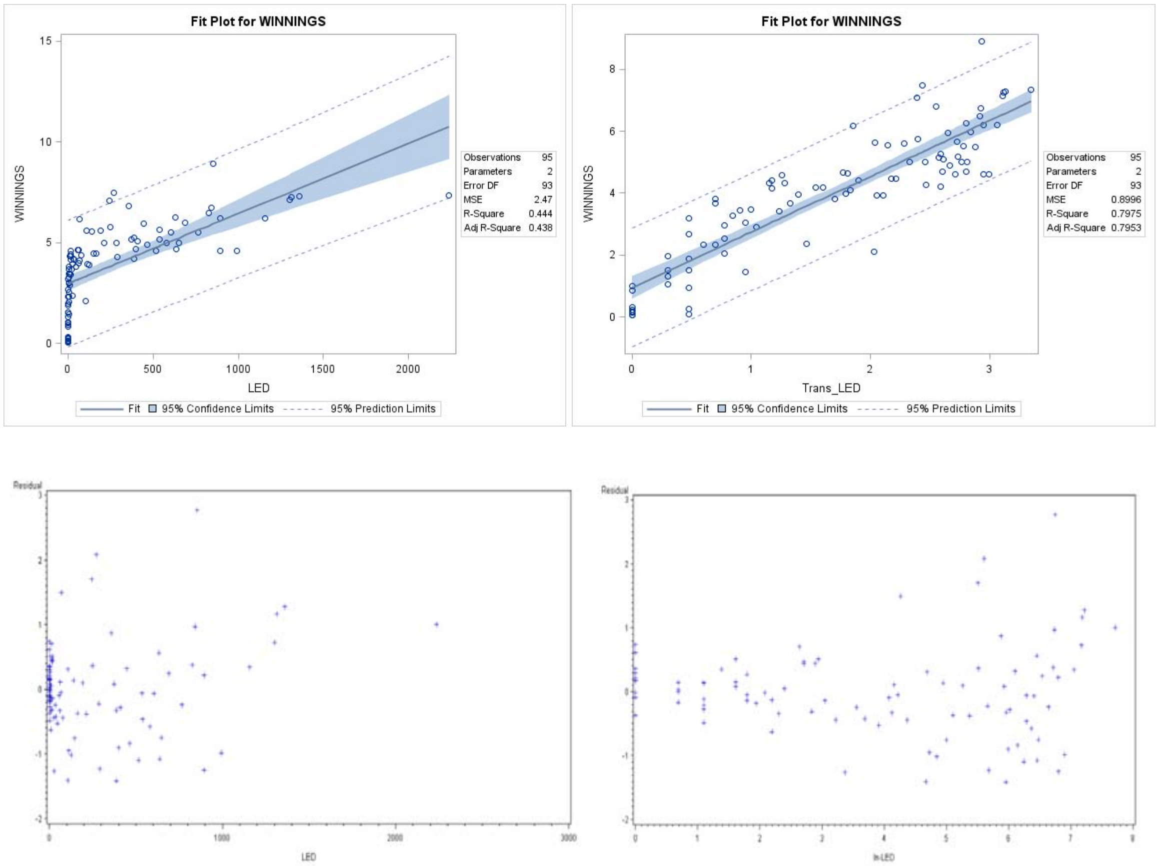


Figure 4 Scatter plots and residual plots of LED before and after transformation

This means that as a driver leads more laps their average winnings goes up. Table 4 contains the statistics for the regression model. We obtained a p-value less than 0.001 for our test on the predictor variables. This indicates that there is strong evidence to conclude that the number of laps led is a significant predictor of a driver's winnings.

Regression Equation				
WINNINGS = 0.947+0.781* ln_Led				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Constant	0.94703	0.18257	5.19	<.0001
ln(1+Led)	0.78079	0.04080	19.14	<.0001
R^2 = 79.7%				

Table 4 Regression of Winnings vs. ln_Led[2]

For this model we got an R^2 value of 79.7% which was our highest one in the simple linear regression models. In conclusion it is safe to say that the number of laps a driver leads is a significant predictor of how much money he will make.

Winnings vs. Year

Lastly we looked at winnings versus year. We wanted to see if there was a significant difference in the winnings from year to year. We decided to do this because we took data from five different years. The assumptions in Figure A7 of the appendix all appear to be valid so we should have a reliable model. The regression equation we came up with is:

Winnings = -83.443 + 0.043*Year

[2] ln_Led represents ln (1+Led), same for any ln_Led shown in this report.

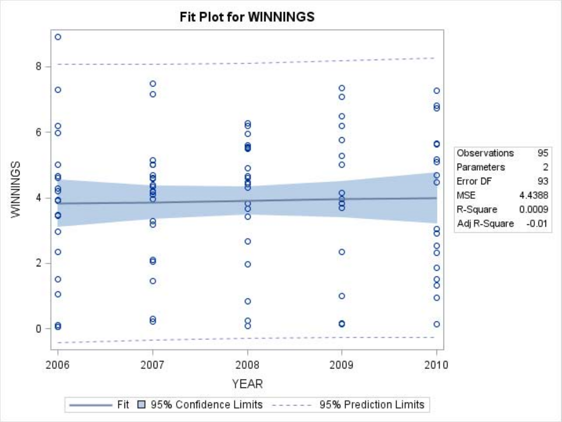


Figure 5 Scatter plot of Winnings vs. Year

This means that for every year increase a drivers winnings goes up on average by $43,435 (0.043 million dollars). To see if year is significant we ran a t-test and obtained a p-value of 0.778 which is seen in Table 5. Since this value is larger than 0.05 we would conclude that there is not enough evidence to say that year is a significant predictor of winnings.

Regression Equation				
WINNINGS = -83.443 +0. 0434 Year				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Constant	-83.44346	309.39460	-0.27	0.788
Year	0.043435	0.15409	0.28	0.778
R^2 = 0.1%				

Table 5 Regression of Winnings vs. Year

The R^2 value backs up our previous test about the predictability of year. We got an R^2 value 0.1% which is practically zero indicating that the model does a horrible job of predicting the amount of money a driver makes. In conclusion the year should not have an effect on how much money a driver makes.

Winnings vs. Make of Car

Here we investigate winnings versus the make of the car. In NASCAR there are only four brands of race cars: Ford, Dodge, Chevrolet and Toyota. To keep with the theme of simple linear regression we are looking at winnings from driving a Toyota compared to the other three brands. Toyotas are fairly new to NASCAR so we thought this would have an effect on the winnings of a driver. We also thought there would be some kind of impact because Toyota is not a domestic company. We were unsure if driving a Toyota would cause a driver to make more or lose more money.

The regression equation and model output is shown in Table 6 is:

Winnings = 4.170 - 1.689*Toyota

In the model β_1 is an indicator variable. A value of one is given if the car is a Toyota and a zero is given if the car is not Ford, Chevy, or Dodge. With that said, β_1 would be interpreted as: on average driving a Toyota reduces a driver's winnings by $1,688,913($1.689 million dollars).

Regression Equation				
WINNINGS = 4.170 – 1.689 Toyota				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Constant	4170034	225143	18.52	<0.001

Toyota	-1688913	566597	-2.98	0.004
R^2 = 8.7%				

Table 6a Regression of Winnings vs. Make of Car

Before a t-test on β_1 can be run, the assumptions mentioned in the methods section needs to be checked. The assumptions are in Figure A8 of the appendix and everything appears to be valid. The test of β_1 shows that driving a Toyota or not is related to how much a driver makes. The p-value in table 6a for driving a Toyota is 0.004 and is less than 0.05 so there is evidence that Toyota is a significant predictor of winnings.

For this model we only obtained a R^2 value of 8.7%, this is not very high and suggests that the model doesn't predict the data very well, but since our t-test gave a significant result, there is some relationship between driving a Toyota and the amount of winnings. From our model it appears that driving a Toyota will reduce a driver's winnings.

We also decided to look at a model with all of the individual makes of the cars as predictors. We set Toyota as the base line and create 3 indicator variables. The regression coefficients can be interpreted as comparisons with Toyota. The model can be seen in Table 6b.

Regression Equation				
WINNINGS = 2.481+0.809Ford+1.227Dodge+2.379Chevrolet				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Constant	2.48112	0.49846	4.98	<0.001
Ford	0.80866	0.66680	1.21	0.2284
Dodge	1.22656	0.64643	1.90	0.0609
Chevrolet	2.37857	0.58654	4.06	0.0001
R^2 = 17.91%				

Table 7b Regression of Winnings vs. Make of Car

Given other facts in the same condition, the winnings are 0.809 million dollars higher for the driver who drives Ford than the one who drives Toyota; the winnings are 1.227 million dollars higher for the driver who drives Dodge than the one who drives Toyota; the winnings are 2.379 million dollars higher for the driver who drives Chevy than the one who drives Toyota. Of the makes, only Chevy and the constant appeared to be significant predictor of winnings, but the relation between car brand and winnings is very weak with R^2 value only 17.91%.

MULTIPLE REGRESSION

In this section we talk about our multiple linear regression models based on the simple regression analysis we have just done.

Variable Selection

Through the simple regression analysis, we have known how much contribution the each variable has made to the winnings and the relationships between winnings and each predictor variable. Let us briefly review the respective R^2 we obtained before:

Simple linear regression						
Variable	**Avg_Fin**	**Avg_Spots_Gained**	**Starts**	**In_Led**	**Year**	**Make of Car**
R2	70.9%	3.0%	71.9%	79.7%	0.1%	8.7% &17.91%

Table 8 R^2 of each simple regression model

Therefore we initialized the predictor variable set as:

X_1~Starts (number of races a driver started in the season)

X_2~Avg. Fin(average finish position for the season)

X_3~In_Led (number of laps led during the season)

Appropriateness of the First-Order Model

Firstly we want to check whether the first-order model is fit for the data. So we checked the interactions between each variable using the VIF values and scatter plots matrix.

We regress Winnings on the predictor variable set and obtain the regression estimation below:

Regression Equation					
WINNINGS = 1.997+0.076*Starts-0.066*Avg _Fin+0.304*ln_Led					
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**	**VIF**
Intercept	1.99675	0.69288	2.88	0.0049	
Starts X1	0.07558	0.00966	7.83	<.0001	2.27443
Avg_Fin X2	-0.06551	0.01985	-3.30	0.0014	4.89069
ln_Led X3	0.30445	0.07871	3.87	0.0002	6.41919
R2 = 88.51%					

Table 9 Regression of Winnings vs. three selected variables

The VIF value of X_3 indicates that X_3 might have interactions with X_1 and X_2. The scatter plots matrix (Figure 6) confirms the VIF values. The plots between X_2 (Avg_Fin) and X_3 (ln_Led) show the linear association clearly exits.

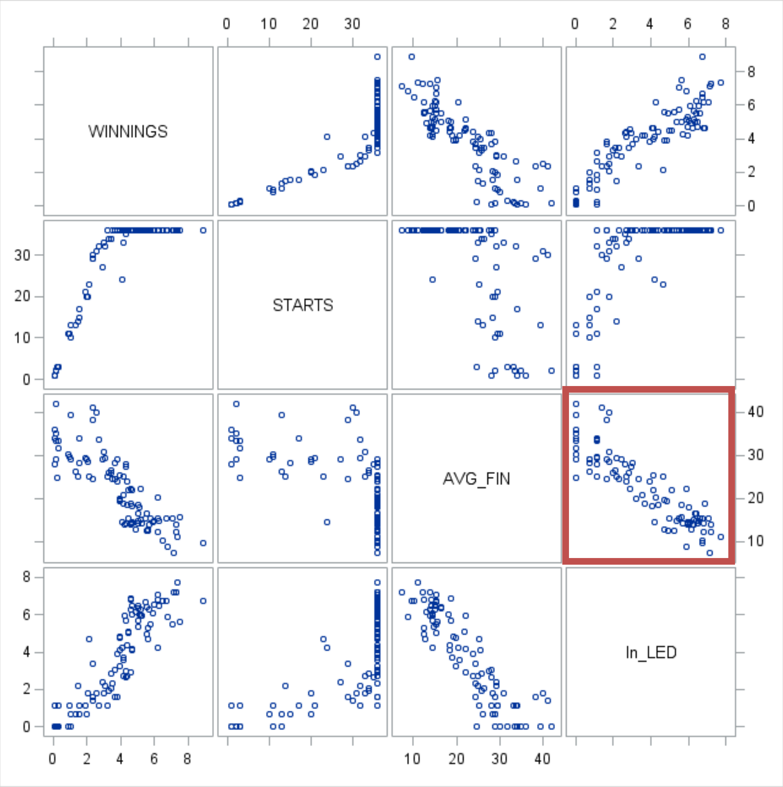


Figure 6 Scatter plots matrix

The p-values of the t-tests are shown below. The interaction term between average finish position and led should be added in the model. It makes sense in the real car race. If a driver is leading more laps this will affect their average finish position, they will most likely finish closer to first the more laps they lead.

Variable	Parameter Estimate	Standard Error	T-value	Pr>|t|
X_{12}	0.00050	0.00175	0.28	0.7773
X_{13}	0.00340	0.00800	0.42	0.6702
X_{23}	-0.02515	0.00725	-3.47	0.0008

Table 10 t-statistics and p-values of interaction terms

By adding the interaction term, we initialize the prediction model as:

Winnings=β_0+ β_1*Starts+ β_2*Avg_Fin+ β_3*ln_Led+ β_4*(Avg_Fin*ln_Led) **Model 1**

Response Transformation

By fitting the Model 1 above, we make the residual plots against predicted winnings and four predictor variables. There is a systematic pattern shown in the residual plot against predicted winnings (Figure 7) which indicates that the error variance is not constant. We also conduct a Breusch-Pagan Test and the result of the test confirms the indication of the graph. So we need a transformation on Winnings. The Box-Cox procedure suggests that $\sqrt{\text{Winnings}}$ is the best transformation, shown in Figure 8.

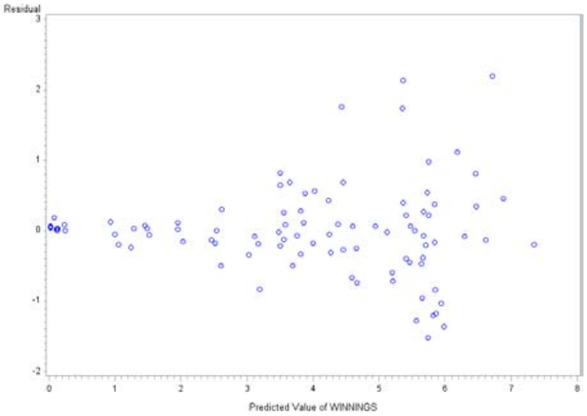


Figure 7 Residual plot against predicted winnings

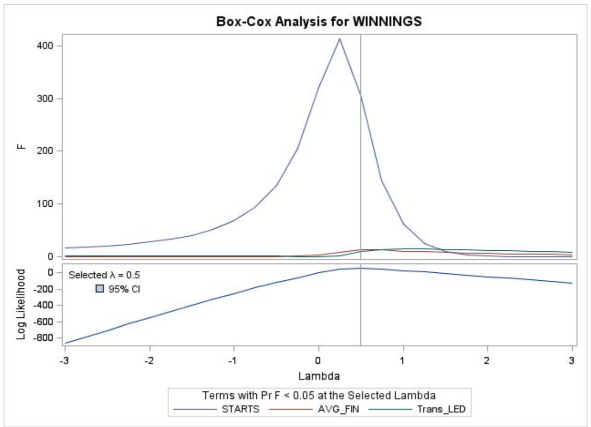


Figure 8 Box-Cox transformations

Prediction Model

After the transformation on the response Winnings, we set the regression model as:

$\sqrt{\text{Winnings}}$=β_0+ β_1*Starts+ β_2*Avg_Fin+ β_3*ln_Led+ β_4*(Avg_Fin*ln_Led) **Model 2**

Now we fit the data with Model 2, we obtain the regression estimation below. The Average Finish Position turns to be not significant in this model. But we have to keep it in the model because it’s the main effect of the interaction term.

Regression Equation				
$\sqrt{\textbf{Winnings}}$ = 0.598+0.042Starts-0.006*Avg_Fin+0.116ln_Led-0.004*(Avg_Fin*ln_Led)				
Predictor	**Coefficient**	**Standard Error**	**T-Value**	**P-Value**
Intercept	0.59820	0.18146	3.30	0.0014
Starts	0.04175	0.00248	16.85	<.0001
Avg_Fin	-0.00606	0.00539	-1.13	0.2633

ln_Led	0.11633	0.02651	4.39	<.0001
Avg_Fin*ln_Led	-0.00393	0.00130	-3.02	0.0033
R^2 = 94.87% MSE=0.02378				

Table 11 Regression of $\sqrt{\text{Winnings}}$ vs. three main effects and interaction term

Assumption Validation

Here we investigate the appropriateness of the regression Model 2 for the learning data. The Pearson coefficient of correlation between the ordered residuals and their expected values is 0.99166, larger than the critical value 0.981 when α level is 0.01. This supports the assumption of normality of error terms.

Next we check the constancy of the error variance by conducting the Breusch-Pagan test, null hypothesis assuming constancy of error variance. The test statistic $X^2{}_{BP} = \frac{SSR^*}{2} \div (\frac{SSE}{95})^2 = \frac{0.01019}{2} \div (\frac{2.14063}{95})^2 = 10.035$. Controlling the α risk at 0.01, we require $\chi^2(0.99; 4) = 13.28$. Since the $X^2{}_{BP}$ is less than critical value, so that we fail to reject null and conclude the error variance is constant.

SAS® Selections

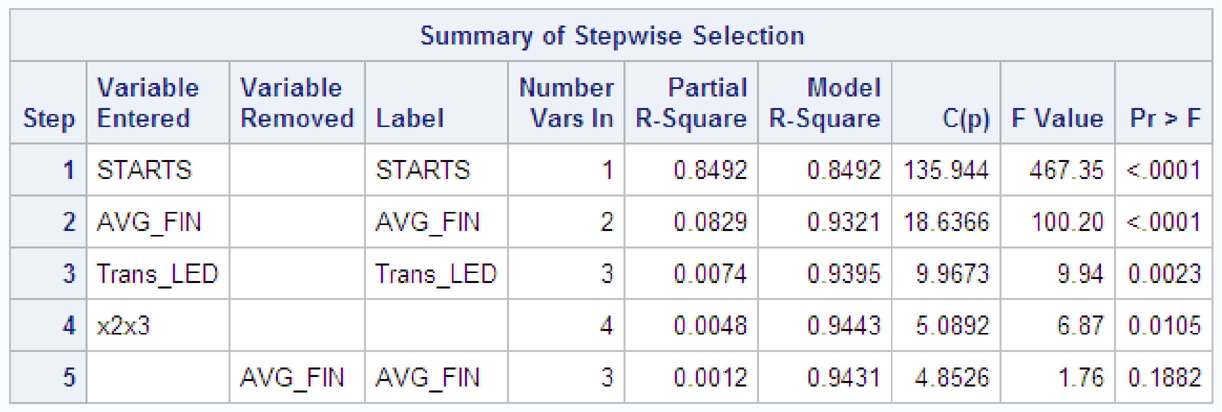
We also ask SAS to select the best combination from the set of variables by using adjusted R^2, stepwise, forward selections and backward elimination. Table 11 shows the results of the stepwise, forward selections and backward elimination by SAS. The stepwise and backward selections chose Starts, ln_Led and interaction term and delete the Avg_Fin. We have to keep Avg_Fin in the model since it is the main effect of the interaction term.

Therefore, we finalized our prediction model as the Model 2. By fitting the data, we obtain the fitted regression equation:

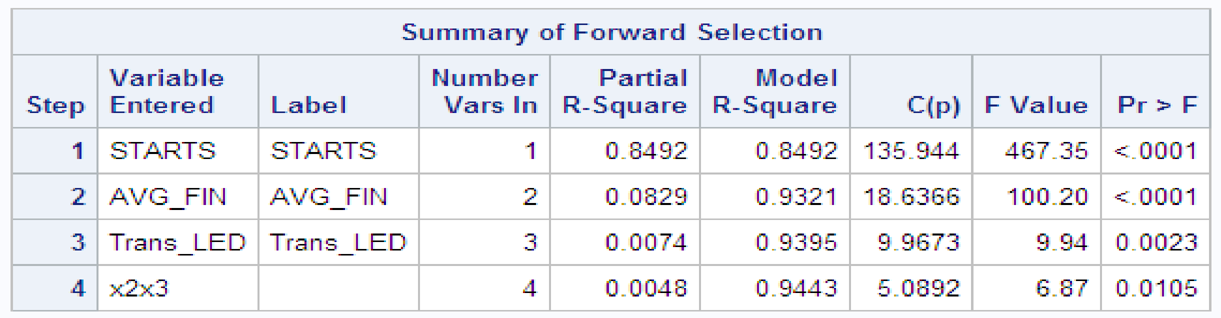
$\sqrt{\text{Winnings}}$ = 0.59820+0.04175*Starts-0.00606*Avg_Fin+0.11633*ln_Led-0.00393 (Avg_Fin*ln_Led)

Then we transform the Winnings back:

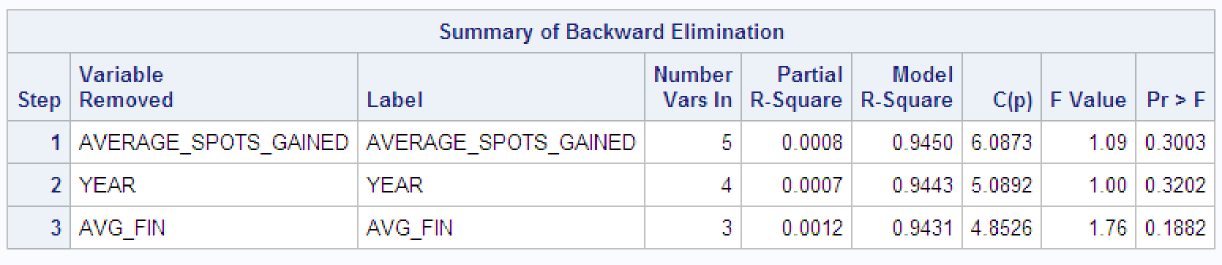
Winnings= $[\mathbf{0.59820 + 0.04175 * Starts - 0.00606 * Avg\_fin + 0.11633 * ln\ \_LED - 0.00393\ (Avg\_fin * ln\ \_LED\ )}]^2$ **Model 3**



Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STARTS		STARTS	1	0.8492	0.8492	135.944	467.35	<.0001
2	AVG_FIN		AVG_FIN	2	0.0829	0.9321	18.6366	100.20	<.0001
3	Trans_LED		Trans_LED	3	0.0074	0.9395	9.9673	9.94	0.0023
4	x2x3			4	0.0048	0.9443	5.0892	6.87	0.0105
5		AVG_FIN	AVG_FIN	3	0.0012	0.9431	4.8526	1.76	0.1882



Summary of Forward Selection								
Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STARTS	STARTS	1	0.8492	0.8492	135.944	467.35	<.0001
2	AVG_FIN	AVG_FIN	2	0.0829	0.9321	18.6366	100.20	<.0001
3	Trans_LED	Trans_LED	3	0.0074	0.9395	9.9673	9.94	0.0023
4	x2x3		4	0.0048	0.9443	5.0892	6.87	0.0105



Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AVERAGE_SPOTS_GAINED	AVERAGE_SPOTS_GAINED	5	0.0008	0.9450	6.0873	1.09	0.3003
2	YEAR	YEAR	4	0.0007	0.9443	5.0892	1.00	0.3202
3	AVG_FIN	AVG_FIN	3	0.0012	0.9431	4.8526	1.76	0.1882

Table 12 Variables selected by Stepwise, Forward and Backward methods

Validation of Test Data

At the very beginning, we have split the data into two data sets: learning data to build the model and test data to examine our model. All jobs we have done were using the test data set. Now we want to examine our prediction model by test data set.

We use Model 3

Winnings= $[0.59820 + 0.04175 * \text{Starts} - 0.00606 * \text{Avg\_Fin} + 0.11633 * \ln\ \text{\_Led} - 0.00393\ (\text{Avg\_Fin} * \ln\ \text{\_Led}\ )]^2$

to predict each case in test data set and calculate the mean of the squared prediction errors (MSPR_L). Then we compare the MSPR_L with MSPR_T to see how well the prediction model works.

$\text{MSPR}_L = \frac{\sum (\text{Yi}-\hat{\text{Y}})2}{\text{n}} = \frac{40.00626}{95}$=0.4211 (calculated from learning data)

$\text{MSPR}_T = \frac{\sum (\text{Yi}-\hat{\text{Y}})2}{\text{n}^*} = \frac{36.1616}{95}$=0.3806 (calculated from test data)

Ratio=$\frac{\text{MSPR}_L - \text{MSPR}_T}{\text{MSPR}_L} = \frac{0.4211-0.3806}{0.4211}$=9.62%

Here we can see the MSPR_T is fairly close to MSPR_L, which implies that MSPR_L based on the learning data set is a reasonably valid indicator of the predictive ability of the fitted regression model. The ratio is 9.62%. Therefore our model is approximately 90.38% accurate.

DISCUSSION

Our first model (Model 1) is regressing Winnings on Starts, Average finish position, transformed Led and the product of last two variables. The assumptions are violated and we tried several ways to fix it.

The average finish is not significant in this first model and the p-value of the t-test as large as 0.9568. However, the sample regression of winnings vs. average finish shows a strong linear association with high R^2 value and small p-value. We thought it might result by the multicollinearity between average finish and its interaction term. So we tried centered predictor variables. The coefficients of correlation between each predictor variable were significantly reduced down. But the centered model didn't improve the error variance constancy.

We have also tried the weighted least squares. According to the residual plots, we found the residual plot against predicted winnings has a megaphone shape. So we regress absolute residuals on winnings to define the weight. The error variance of the weighted model seems very constant, but the error terms doesn't follow normal distribution and seems to be a gamma distribution.

Therefore, we decided to use the Box-Cox transformation.

CONCLUSION

At the beginning of this study, we obtain the relations between the response Winnings and each selected variables Average Finish Position, Average Spots Gained, Starts, ln_Led, Year and Make of Car by using the method of simple linear regression. The P-values of the six regression models tell only three factors are significant, which are Starts, Average Finish Position, and ln_Led.

Based on the R^2 values of simple regression models, we choose three variables as the main effects for our multiple regression model. By checking the multicollinearity, the interaction term of X_{23}, which is the product of Avg_Fin and ln_Led, is kept. The transformation on the response Winnings to $\sqrt{\text{Winnings}}$ suggested by Box-Cox procedure provides a more appropriate response with better normality and variance constancy of errors.

At last we finalize our prediction model with the response $\sqrt{\text{Winnings}}$ and the variable set of Starts, Average Finish Position, ln_Led, and Avg_Fin*ln_Led. This model predicts well for the test data set with the accuracy of 90.38%.

Therefore, we can conclude that our prediction model is effective and accurate.

APPENDIX

This section contains a discussion about assumptions, outliers, and transformations, output, and references.

ASSUMPTIONS, OUTLIERS, AND TRANSFORMATIONS

Figure A1 contains the residual plots for the regression of winnings by average finish position. By looking at the normal probability plot it appears that the assumption of normality is valid because the points follow the diagonal. The large p-value, greater than 0.01, for the Anderson Darling test also suggests that the data follows a normal distribution. The random scatter in the versus order plot suggest that the independence assumption is valid. A Durbin Watson statistic of 1.46 was obtained meaning that the data is independent because the value is larger than one. The random scatter in the versus fits plot suggests that the assumptions of equal variance and linearity are valid.

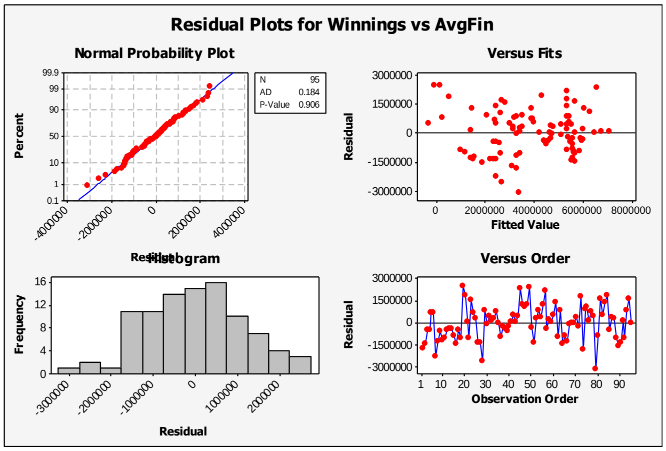


Figure A1 Residual Plots of Winnings vs. Avg. Finish

Figure A2 is a scatter plot of winnings versus the average number of spots gained. It is clear that there is an outlier. The value is at negative twenty five and the average is a .145. This made us wonder if there is a data entry error. After investigating the data set and looking at descriptive statistics for the variable we decided to remove the data point because it could possibly be a data entry error.

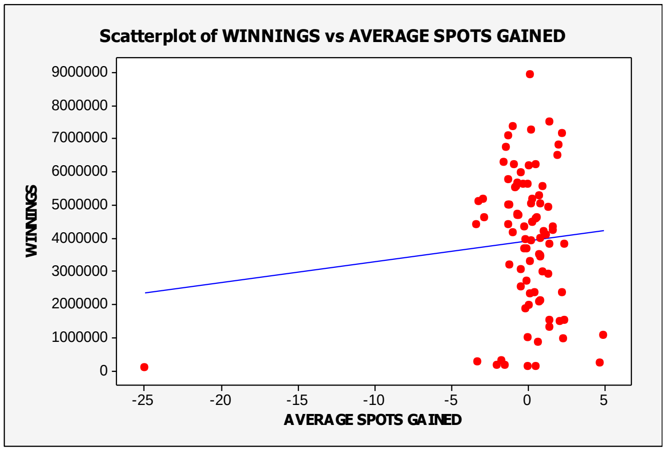


Figure A2 Scatter Plot of Winnings vs. Average Spots Gained

Figure A3 illustrates the residual plots for the regression of winnings by average spots gained. The given Anderson Darling p-value greater than .01 suggests that the data is normal. The obtained Durbin Watson statistic of 0.88

suggests that there may be a problem with independence because the value is less than one. We are going to continue with our analysis even though this assumption may be violated. We believe that time should not have an effect on the average number of spots gained. The random scatter in the versus fits plot suggests that the assumptions of equal variance and linearity are valid.

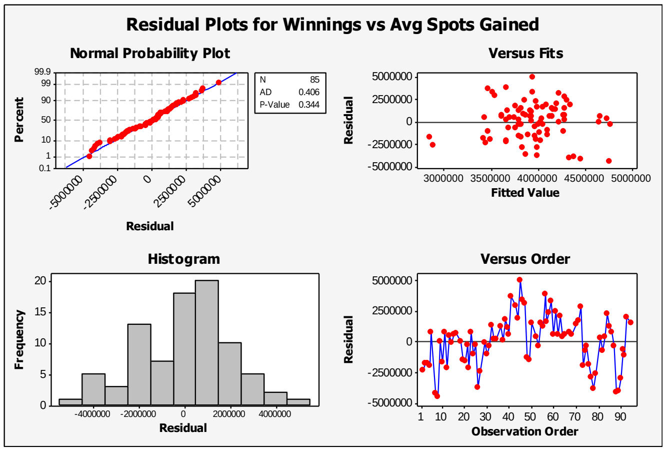


Figure A3 Residual Plots of Winnings vs. Average spots gained

Figure A4 shows the residual plots for the regression of winnings by the number of starts. From looking at all the plots many of the assumptions appear to be violated. We tried multiple transformations on y and x but were unable to fix the assumptions. We are going to run out tests with the original data. We feel the assumptions should not be a big issue because the data set is so large.

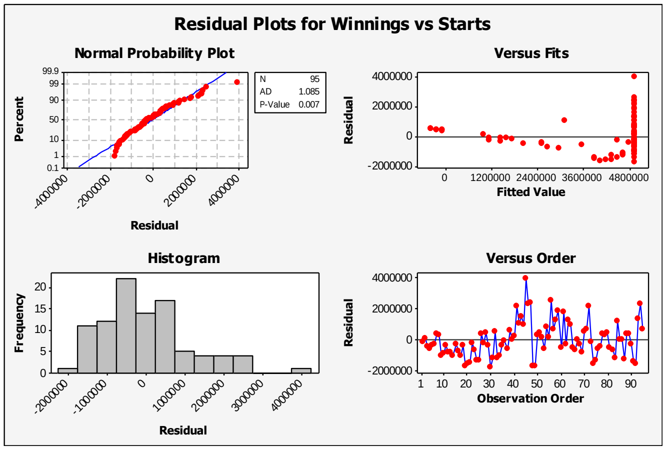


Figure A4 Residual Plots of Winnings vs. Starts

Figure A5 displays the residual plots for the regression of winnings by number of laps led. Normality appears to be violated to we tried a transformation on x. We performed a LN(1+X) transformation because some of the x values are zero. The new residual plots can be shown in Figure A6.

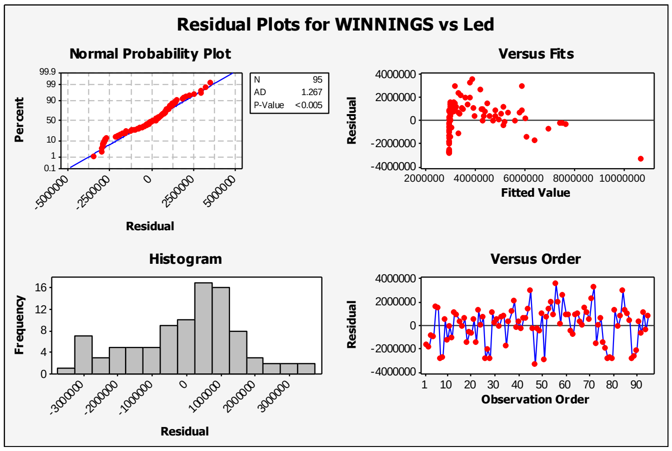


Figure A5 Residual Plots of Winnings vs. Led

Figure A6 is the residual plots for the transformed regression equation of winnings vs. natural log of one plus the number of laps led. All of the assumptions now appear to be valid. Normality because the Anderson Darling p-value of above .01, equal variance and linearity because of the random scatter of points in the versus fits plot, and independence because of the Durbin Watson statistic above 1.

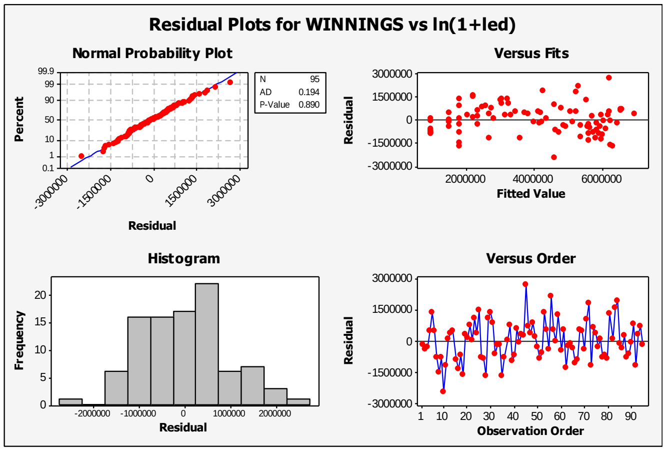


Figure A6 Residual Plots of Winnings vs. LN(1+Led)

Figure A7 contains the residual plots for the regression of winnings by year. The given Anderson Darling p-value of .01 suggests that the data is normal. The Durbin Watson statistic of 0.89 suggests that there may be a problem with independence because the value is less than one. We are going to continue with our analysis even though this assumption may be violated. The random scatter in the versus fits plot suggests that the assumptions of equal variance and linearity are valid.

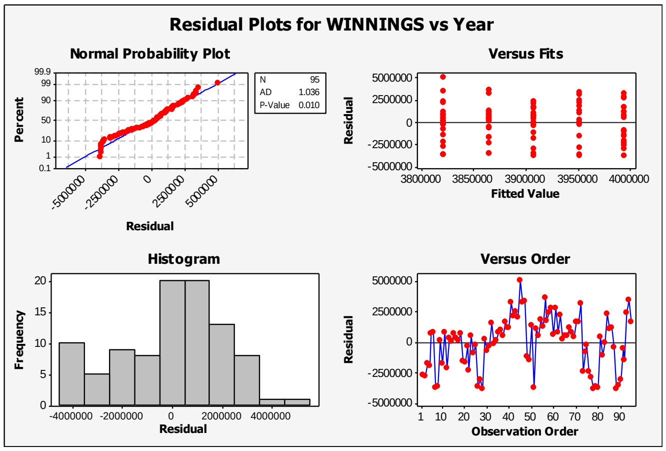


Figure A7 Residual Plots of Winnings vs. Year

Figure A8 is the residual plots for the regression of winnings by make. The given Anderson Darling p-value of .052 suggests that the data is normal. The Durbin Watson statistic of 1.01 suggests independence is valid. Equal variance may be violated but we will continue our analysis because of such a large data set.

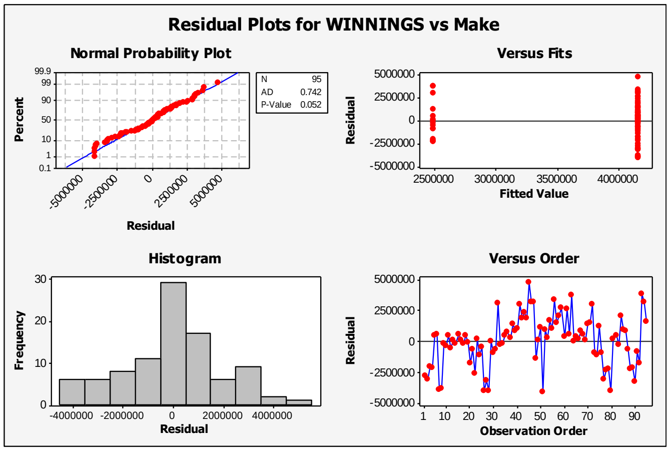


Figure A8 Residual Plots of Winnings vs. Make

MINITAB OUTPUT

Regression Analysis: WINNINGS versus AVG. FIN.

```
The regression equation is
WINNINGS = 8608560 - 212702 AVG. FIN.

Predictor      Coef  SE Coef       T      P
Constant    8608560   333722   25.80  0.000
AVG. FIN.   -212702    14134  -15.05  0.000

S = 1137220   R-Sq = 70.9%   R-Sq(adj) = 70.6%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       1  2.92885E+14  2.92885E+14  226.47  0.000
Residual Error  93  1.20274E+14  1.29327E+12
Total           94  4.13159E+14

Durbin-Watson statistic = 1.45640
```

Regression Analysis: WINNINGS versus MAKE_Toyota

```
The regression equation is
WINNINGS = 4170034 - 1688913 MAKE_Toyota

Predictor        Coef  SE Coef      T      P
Constant      4170034   225143  18.52  0.000
MAKE_Toyota  -1688913   566597  -2.98  0.004

S = 2013740   R-Sq = 8.7%   R-Sq(adj) = 7.7%

Analysis of Variance

Source          DF           SS           MS     F      P
Regression       1  3.60307E+13  3.60307E+13  8.89  0.004
Residual Error  93  3.77129E+14  4.05515E+12
Total           94  4.13159E+14

Durbin-Watson statistic = 1.01396
```

Regression Analysis: WINNINGS versus AVERAGE SPOTS GAINED

```
The regression equation is
WINNINGS = 3983123 - 234096 AVERAGE SPOTS GAINED

Predictor                Coef  SE Coef      T      P
Constant              3983123   221068  18.02  0.000
AVERAGE SPOTS GAINED  -234096   147302  -1.59  0.116

S = 2028633   R-Sq = 3.0%   R-Sq(adj) = 1.8%

Analysis of Variance

Source          DF           SS           MS     F      P
Regression       1  1.03938E+13  1.03938E+13  2.53  0.116
Residual Error  83  3.41574E+14  4.11535E+12
Total           84  3.51968E+14

Durbin-Watson statistic = 0.876602
```

Regression Analysis: WINNINGS versus STARTS

```
The regression equation is
WINNINGS = - 550268 + 152797 STARTS

Predictor     Coef  SE Coef      T      P    VIF
Constant   -550268   310462  -1.77  0.080
STARTS      152797     9899  15.43  0.000  1.000


S = 1116837   R-Sq = 71.9%   R-Sq(adj) = 71.6%

Analysis of Variance

Source          DF           SS           MS       F      P
Regression       1  2.97158E+14  2.97158E+14  238.24  0.000
Residual Error  93  1.16001E+14  1.24732E+12
Total           94  4.13159E+14

Durbin-Watson statistic = 1.13615
```

Regression Analysis: WINNINGS versus LN_Led

```
The regression equation is
WINNINGS = 947034 + 780788 LN_Led

Predictor     Coef  SE Coef      T      P    VIF
Constant    947034   182573   5.19  0.000
LN_Led      780788    40798  19.14  0.000  1.000


S = 948491   R-Sq = 79.7%   R-Sq(adj) = 79.5%

Analysis of Variance

Source          DF           SS           MS       F      P
Regression       1  3.29493E+14  3.29493E+14  366.25  0.000
Residual Error  93  8.36660E+13  8.99635E+11
Total           94  4.13159E+14

Durbin-Watson statistic = 1.53671
```

Regression Analysis: WINNINGS versus YEAR

```
The regression equation is
WINNINGS = - 83443457 + 43502 YEAR


Predictor        Coef    SE Coef      T      P    VIF
Constant    -83443457  309394596  -0.27  0.788
YEAR            43502     154089   0.28  0.778  1.000

S = 2106839   R-Sq = 0.1%   R-Sq(adj) = 0.0%

Analysis of Variance

Source          DF           SS           MS     F      P
Regression       1  3.53779E+11  3.53779E+11  0.08  0.778
Residual Error  93  4.12806E+14  4.43877E+12
Total           94  4.13159E+14

Durbin-Watson statistic = 0.894138
```

SAS CODE

Regression for quantitative variable: Make of Car:

Ford	1 0 0
Dodge	0 1 0
Chevy	0 0 1
Toyota	0 0 0

```
proc reg;
model winnings=make_ford make_dodge make_chevrolet;
run;
```

Before Transformation on Y:

```
proc reg data=car;
model winnings=starts avg_fin ln_led/vif;
run;
proc sgscatter data=car;
matrix winnings starts avg_fin ln_LED;
```

```
run;
proc corr data=car;
var winnings starts avg_fin ln_LED;
run;
data car2;
set car;
x12 = starts*avg_fin;
x13= starts*ln_LED;
x23 = avg_fin*ln_LED;
proc reg;
model winnings=starts avg_fin ln_led x12 x13 x23;
run;
proc reg data=car2;
model winnings=starts avg_fin ln_led x23/r p;
output out=car3 r=residual p=yhat;
proc gplot data=car3;
symbol color=blue value=circle;
plot residual*yhat;
plot residual*starts;
plot residual*avg_fin;
plot residual*ln_LED;
run;
proc transreg data=car2;
model boxcox(winnings)=identity(starts avg_fin ln_LED x23);
run;
```

After Transformation on Y:

```
data car_new;
set car;
x12=starts*avg_fin;
x13=starts*ln_led;
x23 = avg_fin*ln_LED;
trans_winnings=sqrt(winnings);
proc reg;
model trans_winnings= starts avg_fin ln_led x23/r p;
output out=car_new2 r=residual p=yhat;
proc rank data=car_new2 out=test_normal;
var residual;
ranks ranke;
run;
data normal;
set test_normal;
ev = sqrt(0.02274)*Probit((ranke-.375)/(95+.25));
proc corr;
var ev residual;
run;
data bptest;
set car_new2;
ressqr = residual**2;
proc reg;
model ressqr=starts avg_fin ln_LED x23;
run;
quit;
proc reg data=car_new;
model trans_winnings= year starts avg_fin average_spots_gained ln_led x23/ sle=.1
sls=.15 selection=forward; *same for stepwise and backward;
run; quit;
```

Validation:

Calculate MSPR$_L$:

```
data car;
set car;
yhat=(0.59820+(0.04175*Starts)-(0.00606*Avg_Fin)+(0.11633*ln_LED)-
(0.00393*(Avg_fin*ln_LED)))**2;
devsq=(winnings-yhat)**2;
proc means sum;
var devsq;
run;
```

Calculate MSPR$_T$:

```
data car_validation;
set car_validation;
yhat=(0.59820+(0.04175*Starts)-(0.00606*Avg_Fin)+(0.11633*ln_LED)-
(0.00393*(Avg_fin*ln_LED)))**2;
devsq=(winnings-yhat)**2;;
proc means sum;
var devsq;
run;
```

REFERENCES

Data resource: http://www.nascar.com/kyn/nbtn/ (under the statistics section)

Kutner, Nachtsheim and Neter. Feb 1, 2008. Applied Linear Regression Models. New York, NY: McGraw-Hill

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Yun Gao
Enterprise: California State University Long Beach
Address:
City, State ZIP:
Work Phone: (626) 756-0300
Fax:
E-mail: yungao103@gmail.com
Web: