10-year risk for Atherosclerosis Cardiovascular Disease (ASCVD) a Linear Model of Risk

Marcus Migura

Ali Hussain Syed

Steven De La Garza

Vannessa Aguirre

Introduction

The data set chosen has to do with the risk of heart disease in relation to 9 predictors and one predictor that was added by us. When using regression models on the dataset, we should see what variables are most important when it comes to predicting the severity of heart disease risk for a patient. We also want to deduce if any variables are intertwined with one another, by testing whether or not interaction terms produce a better model.

X1: isMale: categorical
X2: isBlack: categorical
X3: isSmoker: categorical
X4: isDiabetic: categorical
X5: isHyptertensive: categorical
X6: Age: numerical
X7: Systolic: numerical
X8: Cholesterol: numerical (Total Cholesterol)
X9: HDL: numerical (Good Cholesterol)
Y0: Risk: numerical (Response Variable)
X22: [(x8-x9)/x9] : LDL to HDL Ratio (Bad/Good Cholesterol)

The objective is to create a relatively uncomplicated multiple regression model that adequately predicts a person's 10-year risk for Atherosclerosis Cardiovascular Disease (ASCVD) Risk Model. While algorithms exist, they are generally larger black box algorithms with more predictors. The additional predictors used by these algorithms include information regarding current treatment and additional test results. The advantage of a less complicated algorithm that uses fewer predictors is the ability to use it as a gross screening tool, that can be used to determine if further testing is needed so that a more accurate risk assessment can be made and so that recommendations can be provided.

The 10-year risk for ASCVD is categorized as:

- Low-risk (<5%),
- Borderline risk (5% to 7.4%)
- Intermediate risk (7.5% to 19.9%), and
- High risk (≥20%).

While everyone desires accuracy at all ranges, the most important range is 5% to 20%. Specifically, we do not wish to underestimate the risk in this range, and will tolerate some amount of over estimation. The model should provide a result close enough to call it a quantitative estimate. A model that predicts high risk for everyone will simply be ignored. If our crude model typically over estimates the risk score by 20% percent would be acceptable. Since our region of interest is Y between 0 and 20, that would be a score of 24.

For the purpose of model building, we will extend that range out to 25%. For risk estimates greater than 25, errors in either direction are more tolerable because the standard of care requires patients with estimates in that range to undergo further testing and evaluation by larger and more complex models.

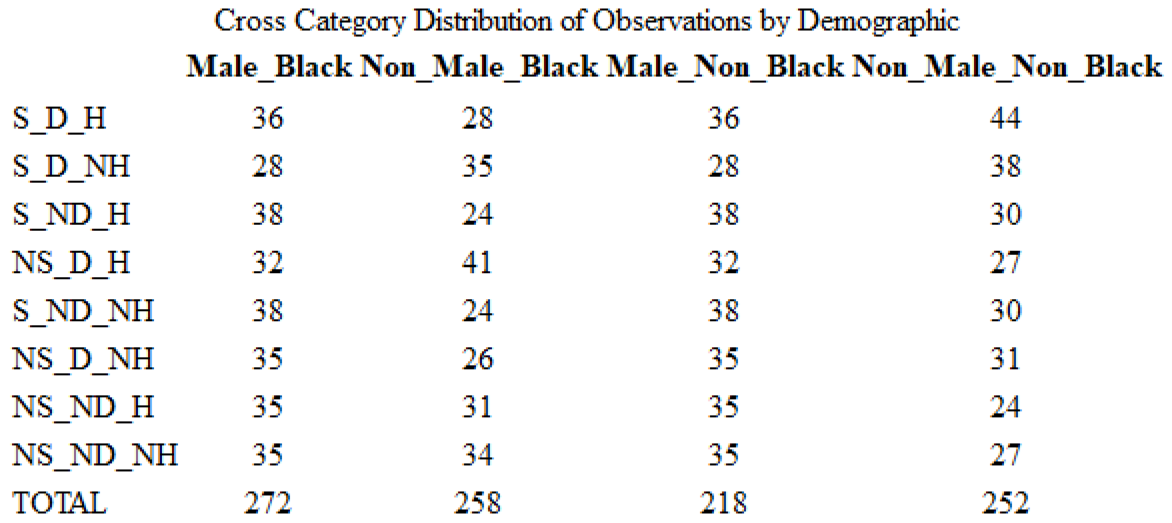
It should be noted that the category "isHypertensive", indicates that the patient is receiving treatment for hypertension. Therefore, it provides information not included in the measure of systolic blood pressure. Also, interaction terms will likely be necessary. If a patient has low systolic values while being treated for hypertension, then that means the treatment is effective in lowering the pressure but should not lower the risk below a patient who had the same systolic value without ever needing treatment, assuming all other predictors were equal. That is if the model adequately reflects clinical expectations. This is captured by the interaction of "isHypertensive" and systolic blood pressure.

The Methodology

Using the original predictors without interaction terms, we conduct an assessment of normality of the residuals. If residuals are not normally distributed then we apply a transformation that maximizes the normality of the residuals. We simply iterated the basic model for exponents, i/100, with i between values of 1 and 100. Additionally we removed gross outliers, by building the model with only true y values greater than 40. All of the models considered were tested against these extreme y values using the prediction function, and none of the models had a problem recognizing the extreme values. Once we have an acceptable transformation, we then begin using interaction terms, and using stepwise regression to remove excess interaction terms. Additionally, we consolidated the x8(Total Cholesterol) and x9(HDL, good cholesterol) predictors, total cholesterol and HDL cholesterol respectively. Which in this case involves computing what is known as the LDL/HDL ratio, represented as x22. Afterwards, we use anova to further remove interaction terms from the step wise reduced model.

Data Analysis

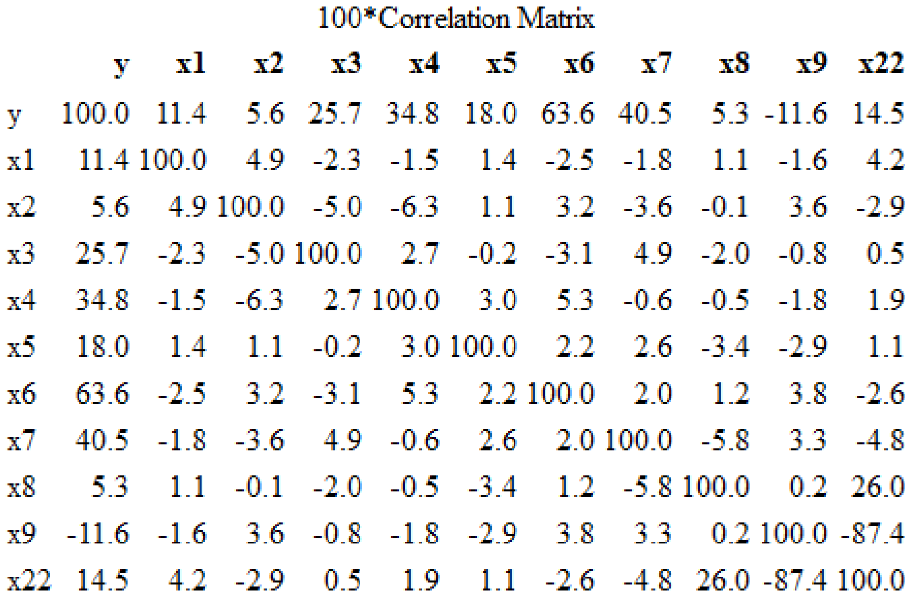
The general breakdown and then correlation is described by the following tables and explanations.



Cross Category Distribution of Observations by Demographic

	Male_Black	Non_Male_Black	Male_Non_Black	Non_Male_Non_Black
S_D_H	36	28	36	44
S_D_NH	28	35	28	38
S_ND_H	38	24	38	30
NS_D_H	32	41	32	27
S_ND_NH	38	24	38	30
NS_D_NH	35	26	35	31
NS_ND_H	35	31	35	24
NS_ND_NH	35	34	35	27
TOTAL	272	258	218	252

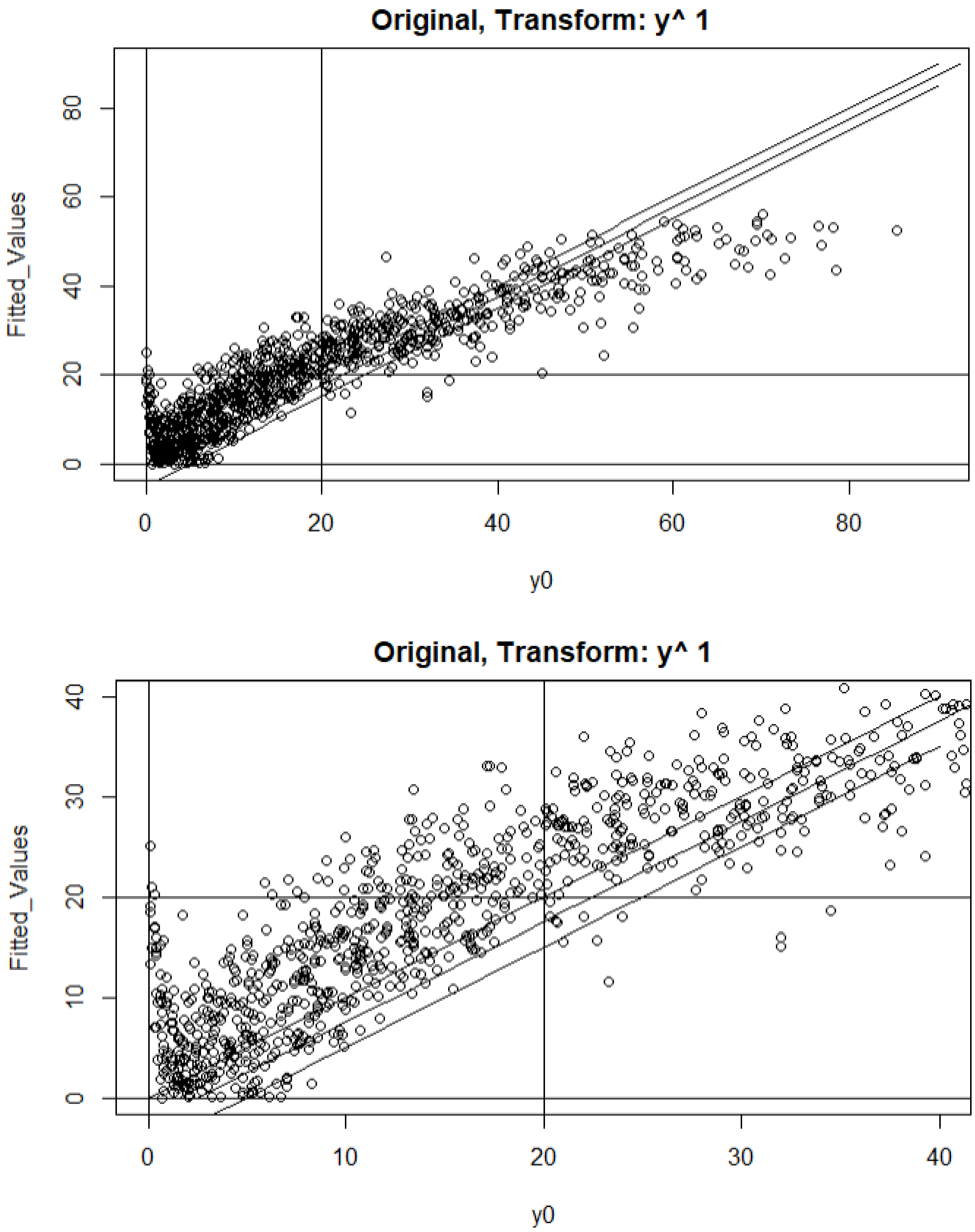
There are 1,000 total observations, with 510 male and 590 non male. A perfectly uniform distribution across all cohorts would be a value of 31.25. This data set appears to have all demographics and cohorts reasonably well represented.



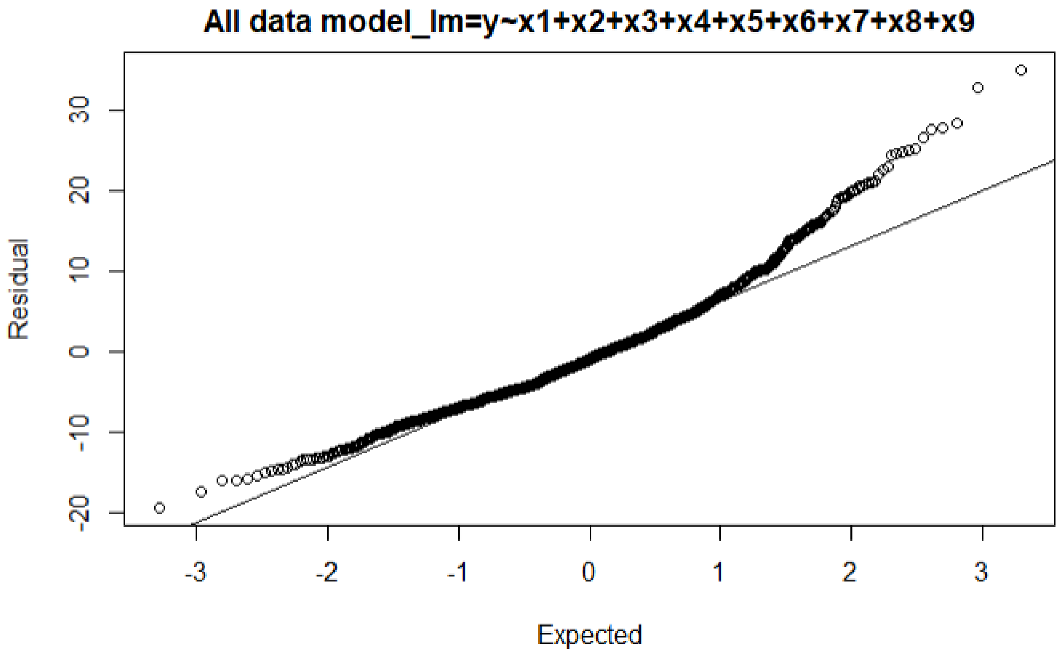
100*Correlation Matrix

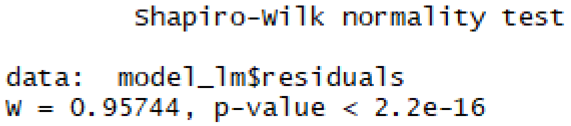
	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x22
y	100.0	11.4	5.6	25.7	34.8	18.0	63.6	40.5	5.3	-11.6	14.5
x1	11.4	100.0	4.9	-2.3	-1.5	1.4	-2.5	-1.8	1.1	-1.6	4.2
x2	5.6	4.9	100.0	-5.0	-6.3	1.1	3.2	-3.6	-0.1	3.6	-2.9
x3	25.7	-2.3	-5.0	100.0	2.7	-0.2	-3.1	4.9	-2.0	-0.8	0.5
x4	34.8	-1.5	-6.3	2.7	100.0	3.0	5.3	-0.6	-0.5	-1.8	1.9
x5	18.0	1.4	1.1	-0.2	3.0	100.0	2.2	2.6	-3.4	-2.9	1.1
x6	63.6	-2.5	3.2	-3.1	5.3	2.2	100.0	2.0	1.2	3.8	-2.6
x7	40.5	-1.8	-3.6	4.9	-0.6	2.6	2.0	100.0	-5.8	3.3	-4.8
x8	5.3	1.1	-0.1	-2.0	-0.5	-3.4	1.2	-5.8	100.0	0.2	26.0
x9	-11.6	-1.6	3.6	-0.8	-1.8	-2.9	3.8	3.3	0.2	100.0	-87.4
x22	14.5	4.2	-2.9	0.5	1.9	1.1	-2.6	-4.8	26.0	-87.4	100.0

This correlation matrix shows reasonably good correlation to the predictor variable. Since x22 is a composite of x8 and x9, we expect it to have high correlation to those predictors. This is also our reasoning for the removal of x8 and x9. Thus we retain only x22 to represent cholesterol.

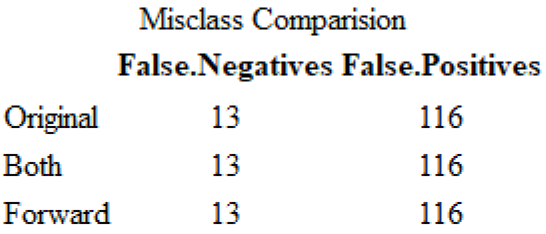


False positives, represented in the upper left quadrant. False negatives, in the lower right quadrant. Also the extreme true y values, may not be useful for the purposes of model building, since they are so far outside of the region of interest.





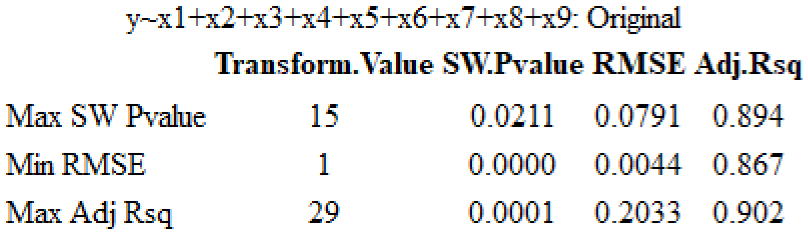
```
        Shapiro-Wilk normality test

data:  model_lm$residuals
W = 0.95744, p-value < 2.2e-16
```



Misclass Comparision

	False.Negatives	False.Positives
Original	13	116
Both	13	116
Forward	13	116

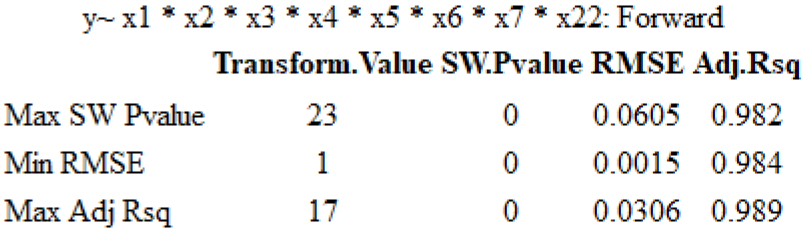
The basic model, containing only the original variables without transformation does yield a reasonable qualitative result as far as minimizing the false negatives. The criteria for false negative was based on if the yhat value was below 20 while true y values was greater than 20.



y~x1+x2+x3+x4+x5+x6+x7+x8+x9: Original

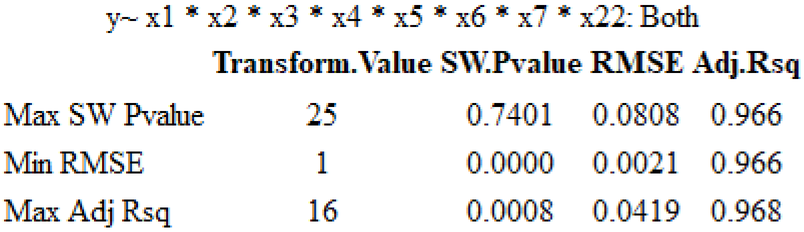
	Transform.Value	SW.Pvalue	RMSE	Adj.Rsq
Max SW Pvalue	15	0.0211	0.0791	0.894
Min RMSE	1	0.0000	0.0044	0.867
Max Adj Rsq	29	0.0001	0.2033	0.902

A transform value of 15 indicates that y was transformed by y^(15/100).
With alpha at 0.01, the y^(15/100) resulted in a normal distribution residuals.



y~ x1 * x2 * x3 * x4 * x5 * x6 * x7 * x22: Forward

	Transform.Value	SW.Pvalue	RMSE	Adj.Rsq
Max SW Pvalue	23	0	0.0605	0.982
Min RMSE	1	0	0.0015	0.984
Max Adj Rsq	17	0	0.0306	0.989



y~ x1 * x2 * x3 * x4 * x5 * x6 * x7 * x22: Both

	Transform.Value	SW.Pvalue	RMSE	Adj.Rsq
Max SW Pvalue	25	0.7401	0.0808	0.966
Min RMSE	1	0.0000	0.0021	0.966
Max Adj Rsq	16	0.0008	0.0419	0.968

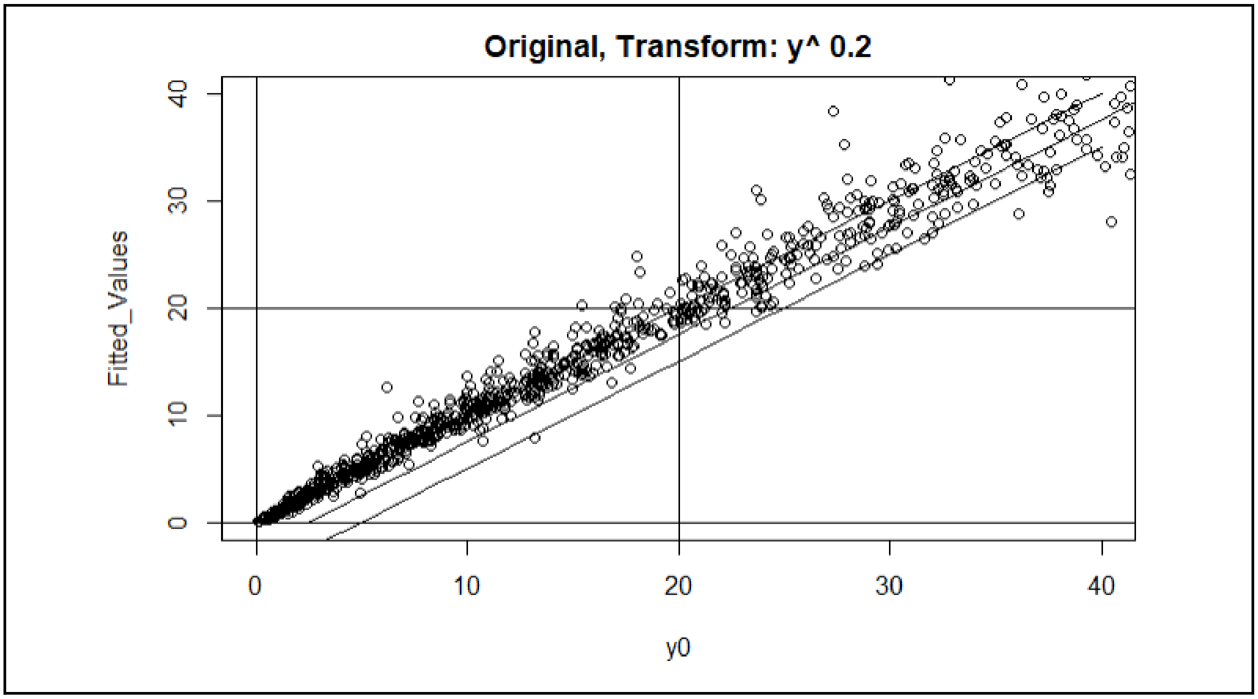
When the extreme Y,those with a risk score greater than 40, are omitted, the residuals normalize. After the models were built they were evaluated by using the predict function on the extreme Y values. The minimum yhat for the extreme values was 30, which is reasonable enough, considering our objective.

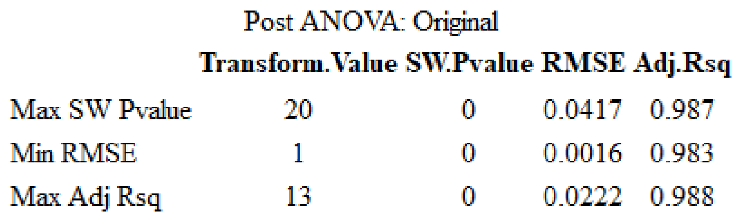
Conclusion

After viewing the results of anova we then removed more interaction terms until it was reduced to the following.

y ~ x6 + x7 + x4 + x3 + x22 + x1 + x2 + x5 + x6:x2 + x6:x1 + x22:x2+ x6:x3 + x7:x1 + x7:x2 + x6:x7 + x2:x5 + x3:x1 + x22:x1 + x4:x2 + x6:x5 + x4:x1 + x1:x5 + x1:x2 + x3:x2 + x7:x5 + x6:x7:x2 + x6:x7:x1 + x6:x22:x1 + x6:x2:x5 + x6:x1:x5 + x22:x1:x2+ x7:x1:x2 + x1:x2:x5 + x6:x1:x2 + x4:x1:x2 + x7:x22:x1 + x6:x3:x2 + x3:x1:x2 + x3:x22:x2 + x6:x7:x1:x2 + x6:x1:x2:x5 + x6:x22:x1:x2

In this version there is no difference between the results of step-wise regression and the original model provided to the step-wise regression. After this model was built the prediction function was used to provide a fitted value, for the extreme y values, those greater than 40. The minimum fitted value was 30. This kind of error is considered acceptable considering the region of interest is less than 40, specifically it is between 0 and 20. The residuals were still reasonably normal for this model. Ultimately, it must be validated to determine performance on new data. Even though the number of predictors is greatly reduced from the model with all possible predictors, it still contains a significant number of predictor interaction terms. Considering the distribution of the original categorical predictors discussed in the introduction, it is not surprising that many still remain.





Post ANOVA: Original

	Transform.Value	SW.Pvalue	RMSE	Adj.Rsq
Max SW Pvalue	20	0	0.0417	0.987
Min RMSE	1	0	0.0016	0.983
Max Adj Rsq	13	0	0.0222	0.988

This reduced model still performed reasonably well. Ultimately it would be reasonable to add 10 percent to the estimate provided by the model. Shown below. Note, in the lower graph, the upper line is 5 points above the true Y line.