2016

Distance of American Football Punts

STAT512 FINAL PROJECT

ASHISH MALIK, SONGYAN LI, PRATIK CHAWLA, MALLORY HUFF

Introduction

Background of the Study

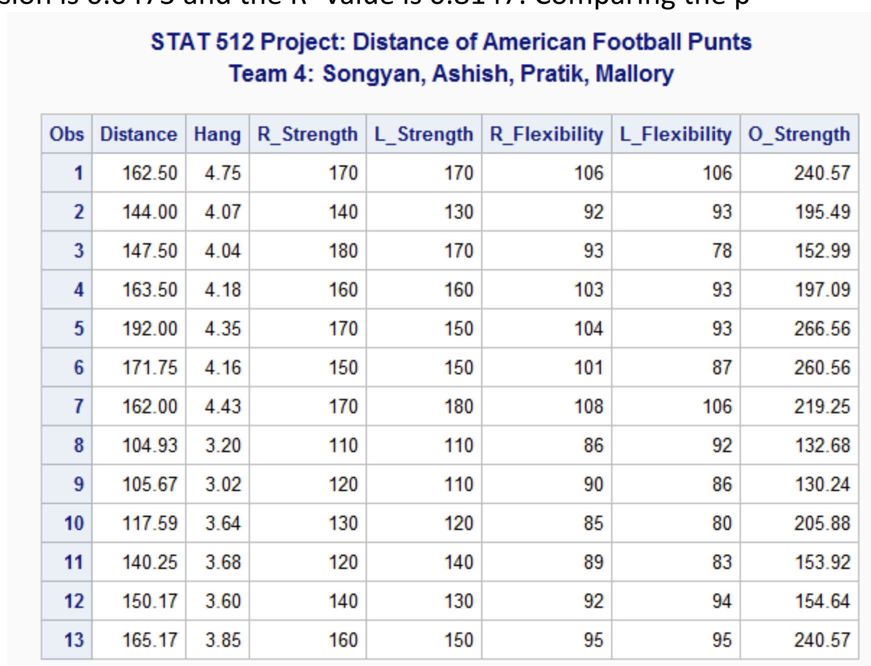
This report is based on a study of the relationship between a punter's punting ability and selected physical performance variables. There are thirteen observations in the data set with each observation taken from one punter – thirteen different punters in total. The response variable of the study is **distance**, which represents the average recorded distance of a punter's ten punts, in feet. The six explanatory/predictor variables and their description can be seen in Table 1.

Table 1. Model Predictor Variables and Their Description

Predictor Variable	Description
Hang	The average hang time of a punter's ten punts (in feet)
R_Strength	A punter's right leg strength (in pounds)
L_Strength	A punter's left leg strength (in pounds)
R_Flexibility	A punter's right leg/hamstring muscle flexibility (in degrees)
L_Flexibility	A punter's left leg/hamstring muscle flexibility (in degrees)
O_Strength	A punter's overall leg strength (in pounds)

Regression of Model with all Six Predictor Variables

All statistical analyses for this project was completed in SAS. Furthermore, a compilation of all thirteen observations can be seen in Figure 1. Running a regression of the full model (using all six variables to predict distance) results in the output shown in Appendix 2. As noticed, the p-value of the regression is 0.0473 and the R^2 value is 0.8147. Comparing the p-value to an assumed alpha level (α = 0.05), the test is significant because 0.0473 < α = 0.05. This implies that at least one of the six predictors is non-zero and is helpful in predicting the distance that a punter punts. The regression also produces a high R^2 value, meaning that a large amount of variation in distance is explained by the model. While these two factors point to a good model, it is noticed that no predictors are individually significant because none of their individual p-values shown at the bottom of the regression results are less than the



STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory

Obs	Distance	Hang	R_Strength	L_Strength	R_Flexibility	L_Flexibility	O_Strength
1	162.50	4.75	170	170	106	106	240.57
2	144.00	4.07	140	130	92	93	195.49
3	147.50	4.04	180	170	93	78	152.99
4	163.50	4.18	160	160	103	93	197.09
5	192.00	4.35	170	150	104	93	266.56
6	171.75	4.16	150	150	101	87	260.56
7	162.00	4.43	170	180	108	106	219.25
8	104.93	3.20	110	110	86	92	132.68
9	105.67	3.02	120	110	90	86	130.24
10	117.59	3.64	130	120	85	80	205.88
11	140.25	3.68	120	140	89	83	153.92
12	150.17	3.60	140	130	92	94	154.64
13	165.17	3.85	160	150	95	95	240.57

Figure 1. Dataset of Thirteen Observations

assumed α value of 0.05. This is a sign of multicollinearity and will be addressed later.

Part 1

Question 1: Piecewise Model

In order to determine which predictor variable to perform a piecewise model with, we generated scatterplots for each individual variable against distance, the response, and plotted a smoothing curve to show the relationship. All six individual scatterplots can be seen in Appendix 3. Hang was the predictor of choice for the piecewise model. This was chosen due to an obvious difference in slope as hang increases, which can be seen from the scatterplot in Figure 2.

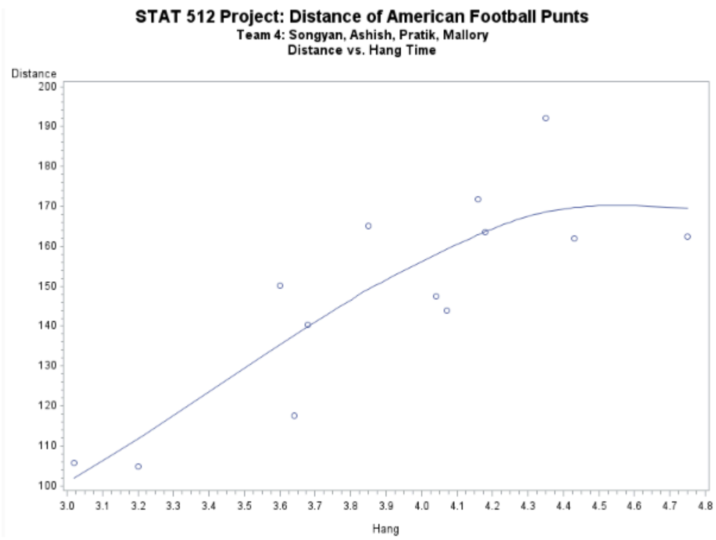


Figure 2. Distance vs. Hang Scatterplot with Smoothing Curve (sm = 70)

As noticed from the scatterplot, there appears to be a change in slope around a hang time of 4.18 seconds. The slope still appears to be slightly positive after 4.18 seconds but it is definitely not as steep of a relationship. It is also convenient that this change in slope corresponds to an already existing data point at 4.18 seconds (and a distance of 163.5 feet). In order to create the piecewise model, the data was first ordered based on the hang time variable. Secondly, the data was split into two pieces using a new variable named "cslope". The new dataset ordered based on hang time and including the "cslope" variable can be seen in Figure 3. The value of "cslope" was based on which piece of the model is being addressed. The first piece – a point that has a hang value less than or equal to 4.18 seconds – has a "cslope" value set to 0. The "cslope" value of the second piece – any point with a hang value greater than 4.18 seconds – is equivalent to the hang value minus 4.18 seconds.

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Distance vs. Hang Time

Obs	Distance	Hang	R_Strength	L_Strength	R_Flexibility	L_Flexibility	O_Strength	cslope
1	105.67	3.02	120	110	90	86	130.24	0.00
2	104.93	3.20	110	110	86	92	132.68	0.00
3	150.17	3.60	140	130	92	94	154.64	0.00
4	117.59	3.64	130	120	85	80	205.88	0.00
5	140.25	3.68	120	140	89	83	153.92	0.00
6	165.17	3.85	160	150	95	95	240.57	0.00
7	147.50	4.04	180	170	93	78	152.99	0.00
8	144.00	4.07	140	130	92	93	195.49	0.00
9	171.75	4.16	150	150	101	87	260.56	0.00
10	163.50	4.18	160	160	103	93	197.09	0.00
11	192.00	4.35	170	150	104	93	266.56	0.17
12	162.00	4.43	170	180	108	106	219.25	0.25
13	162.50	4.75	170	170	106	106	240.57	0.57

Figure 3. Ordered Data by Hang time and Including "Cslope"

In theory, $Y - hat = b_0 + b_1x_1 + b_2x_3$, where b_0 is the intercept, b_1 is the coefficient of hang time, and $x_3 = x_1\ x_2 = "cslope"$ and x_2 is an indicator variable representing 0 when the hang time is less than or equal to 4.18 seconds and 1 when the hang time is greater than 4.18 seconds. $x_3 = "cslope"$ is an explanatory variable that adds a constant to the already existing slope whenever the hang time is greater than 4.18.

$$Y - hat = b_0 + b_1x_1 + b_2x_3$$
$$Y - hat = b_0 + b_1x_1 \ when \ x_2 = 0 \ (x_1 \leq 4.18)$$
$$Y - hat = (b_0 - 4.18b_2) + (b_1 + b_2)x_1 \ when \ x_2 = 1 \ (x_1 > 4.18)$$

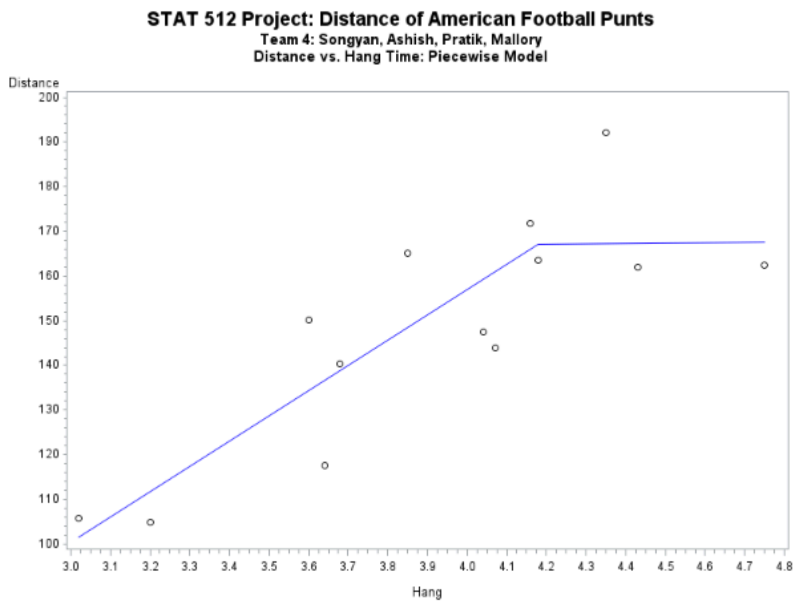


Figure 4. Piecewise Scatterplot of Distance vs. Hang

A scatterplot of the resulting piecewise model can be seen in Figure 4. Running a regression of this piecewise model gives the equations of the two pieces. The equation of the first piece (all hang values less than or equal to 4.18 seconds) is $distance - hat = -69.63871 + 56.65758 * hang$ and the equation of the second piece is given by $distance - hat = 164.71 + 0.5931 * hang$. The slopes of the two different pieces (56.65758 and 0.5931) are obviously not the same. The full regression results can be seen in Appendix 4. This piecewise model is highly significant with a p-value of 0.0012 (< α = 0.05), implying that the slopes of hang are non-zero, as verified by the scatterplot. Additionally, the regression produces an R^2 value of 0.7416. About 74% of the variation in the distance can be explained by the two hang pieces.

Question 2: Extra Sum of Squares

Before conducting calculations related to the extra sum of squares, a “SUM” variable was created as the addition of the R_Strength and R_Flexibility variables. Two regressions were then performed and the results can be seen in Figures 5 & 6. The first regression was run in order to predict the distance response using all of the predictor variables except the “SUM” variable and the two predictor variables used to create it (R_Strength and R_Flexibility). The second regression predicts the distance response using all of the predictor variables including the “SUM” variable (still not including R_Strength and R_Flexibility because they are included in “SUM”).

$$SSM(without \ SUM) = 6331.70324$$
$$SSM \ (with \ SUM) = 6511.37312$$
$$Extra \ Sum \ of \ Squares = SSM(without \ SUM) - SSM(with \ SUM) = 179.66988$$

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Regression with all predictor variables

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6331.70324	1582.92581	7.19	0.0093
Error	8	1761.60523	220.20065		
Corrected Total	12	8093.30848			

Root MSE	14.83916	R-Square	0.7823
Dependent Mean	148.23308	Adj R-Sq	0.6735
Coeff Var	10.01069		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	15.74208	49.80100	0.32	0.7600
Hang	1	3.55879	25.47261	0.14	0.8923
L_Strength	1	0.46956	0.40167	1.17	0.2760
L_Flexibility	1	-0.06969	0.59261	-0.12	0.9093
O_Strength	1	0.29233	0.14850	1.97	0.0845

Figure 5. Regression without SUM

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Regression with all predictor variables including the sum

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6511.37312	1302.27462	5.76	0.0200
Error	7	1581.93535	225.99076		
Corrected Total	12	8093.30848			

Root MSE	15.03299	R-Square	0.8045
Dependent Mean	148.23308	Adj R-Sq	0.6649
Coeff Var	10.14145		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.95631	56.51017	-0.12	0.9055
Hang	1	2.25646	25.84664	0.09	0.9329
L_Strength	1	0.10492	0.57691	0.18	0.8608
L_Flexibility	1	-0.11027	0.60207	-0.18	0.8599
O_Strength	1	0.24199	0.16069	1.51	0.1758
SUM	1	0.38554	0.43239	0.89	0.4022

Figure 6. Regression with SUM

From the regression outputs above, the F-statistic can be calculated. This value is the general linear test statistic for testing the null hypothesis that the coefficient of the "SUM" variable is zero in the model with all predictors (other than R_Strength and R_Flexibility). The degrees of freedom for this F-statistic is 1 and the F-statistic itself is 0.79503.

$$dfm(F-R) = DFM(F) - DFM(R) = 5 - 4 = 1$$

$$F = \frac{\frac{SSM(F-R)}{dfm\,(F-R)}}{MSE(F)} = \frac{\frac{179.66988}{1}}{225.99076} = 0.79503$$

To verify that this test statistic is correct, an individual test for the "SUM" coefficient was performed. The null hypothesis of this test states that the coefficient of "SUM" is zero while the alternate hypothesis states that the coefficient is non-zero. The results of this test can be seen in

Test t1 Results for Dependent Variable Distance				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	179.66988	0.80	0.4022
Denominator	7	225.99076		

Figure 7. Individual Test of "SUM"

Figure 7. The F-statistic is 0.80 and the degrees of freedom is 1. Additionally, the p-value of the test is 0.4022. Because 0.4022 > α = 0.05, the conclusion is to fail to reject the null hypothesis.

Comparing the degrees of freedom and the F-statistic of the individual test to the same calculated values above, it is noticed that the are equivalent. The individual p-value of "SUM" in Figure 6 is 0.4022 which is also equivalent to the p-value of the individual test determining if the "SUM" variable is zero or non-zero. One can also notice that $F = t^2 = 0.89^2 \approx 0.8$. Due to the conclusion based on the p-value, there is evidence that the coefficient of "SUM" is in fact zero and is not helpful in predicting the distance of punts when L_Strength, L_Flexibility, Hang, and O_Strength are present.

Question 3: Type I & Type II Sums of Squares

To evaluate the Type I and Type II Sums of Squares, a regression of the model with all six predictors (not including "SUM") was run and Type I and Type II sums of squares were generated. The order the predictors were placed in the model is as follows: Hang, L_Strength, L_Flexibility, O_Strength, R_Strength, and R_Flexibility. The full regression results can be found in Appendix 5 but the specific Type I and Type II Sums of Squares Values are listed in Table 2.

Table 2. Ordered Predictor Variables and Their Type I and Type II SS

Predictor Variable (In Order)	Type I SS	Type II SS
Hang	5426.73476	2.29989
L_Strength	45.56266	0.94680
L_Flexibility	6.10052	62.69707
O_Strength	853.30530	367.36136
R_Strength	103.91050	78.13926
R_Flexibility	157.67347	157.67347

$$Sum\ of\ Type\ I\ SS: 5426.73476 + 45.56266 + 6.10052 + 853.30530 + 103.91050 + 157.67347 = 6593.28721$$

$$Sum\ of\ Type\ II\ SS: 2.29989 + 0.94680 + 62.69707 + 367.36136 + 78.13926 + 157.67347 = 669.11785$$

The Type I sum of squares for the predictor variables adds up to the model sum of squares in the full regression output because in each instance, one extra SS for a variable is added to the current sum of squares calculated from all other variables already in the equation. No two individual SS is being counted twice because Type I SS is based on the previous variables in the model, not all other variables.

$$Type\ I\ SS: SSM(Full) = SSM(Hang) + SSM\ (L\_Strength|Hang) + SSM(L\_Flexibility|\ Hang, L\_Strength) + SSM\ (O\_Strength|Hang, L\_Strength, L\_Flexibility) + SSM(R\_Strength|\ Hang, L\_Strength, L\_Flexibility, O\_Strength) + SSM(R\_Flexibility|Hang, L\_Strength, L\_Flexibility, O\_Strength, R\_Strength)$$

The Type I and Type II SS are equivalent for the final predictor in the model – R_Flexibility. Both values are represented by the following because R_Flexibility is the last predictor: $SSM(R\_Flexibility|Hang, L\_Strength, L\_Flexibility, O\_Strength, R\_Strength)$ This is due to the fact that Type II SS is based on the extra sum of squares when all other predictors are in the model, regardless of order. Type I SS is a sequential sum of squares and therefore all SS for the other predictors is included in the model when the last predictor is added, making them equivalent.

(Also notice there is a significant difference between the Type I and Type II SS of both Hang and O_Strength variables. This implies that there is a multicollinearity problem within the dataset. For example, when Hang is the only variable in the model it accounts for 5426.73476 of the 6593.28721 sum of squares in the model. When all other predictors are in the model, the extra sum of squares for the Hang variable is only 2.29989. Multicollinearity will be addressed later.)

Question 4: Various Regressions with Different Predictor Variables

Different combinations of predictor variables were used to predict the distance of a punter's punt. The "SUM" variable was included as a predictor variable option. Table 3 describes the different combinations of predictor variables chosen and the R^2 values generated from the individual regression of each particular model.

Table 3. Different Models and R-squared Values

Predictor Variables in the Model	R^2 Value
Hang	0.6705
R_Strength	0.6264
L_Strength	0.5536
R_Flexibility	0.6502
L_Flexibility	0.1663
SUM	0.6916
O_Strength	0.6339
R_Flexibility and R_Strength	0.7198
L_Flexibility and L_Strength	0.5641
R_Strength, O_Strength, L_Strength, SUM	0.8069
R_Flexibility, L_Flexibility, SUM	0.7322
Hang, R_Flexibility, R_Strength	0.7392
Hang, R_Strength, L_Strength, R_Flexibility, L_Flexibility, O_Strength	0.8147

Generally, the value of R^2 increases with an increased number of predictors. This is not the case in some instances, such as the value of R^2 corresponding to the model with L_Flexibility and L_Strength not being higher than some models with only one predictor. However, this is intuitive when thinking about football and commonalities. Many punters are right-footed and not left-footed so physical performance relating to the left leg might not be as helpful in predicting distance.

Part 2

Question 1: Are Any Transformations Necessary?

In order to determine if there was a better model for predicting distance other than the model including all six predictor variables as given, different transformation options were checked and compared with the original variables. First, all predictor variables were plotted against the response variable of distance and a smoothing curve of weight 85 was used to evaluate the relationship. These six scatterplots with smoothing curves can be seen in Appendix 6. At first glance, most of the individual scatterplots appears to be fairly linear. Two variables that look like they could cause a slight linearity problem are R_Strength and L_Strength. The individual residual plots of the six predictor variables in Figure 8 also show no major issues in the constant variance or linearity assumptions. There appears to be no pattern or "megaphone" shape within the plots; the pattern appears to be random. No obvious outliers are present either.

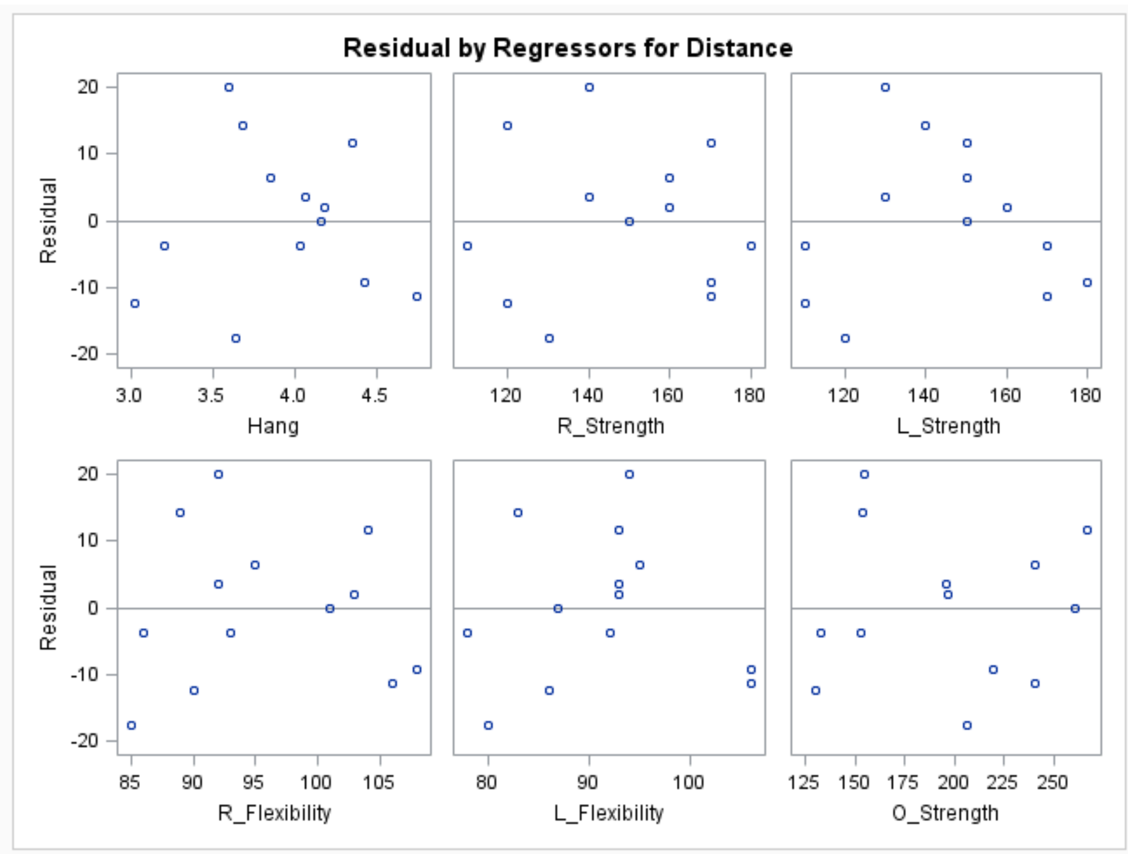


Figure 8. Residual Plots of Predictor Variables

To check normality of the given dataset, a QQplot and histogram were generated (see Figures 9 & 10). From these plots, it was determined that the normality assumption was not violated. The points on the QQplot follow a relatively linear pattern and the histogram shows the typical bell curve of a normal relationship.

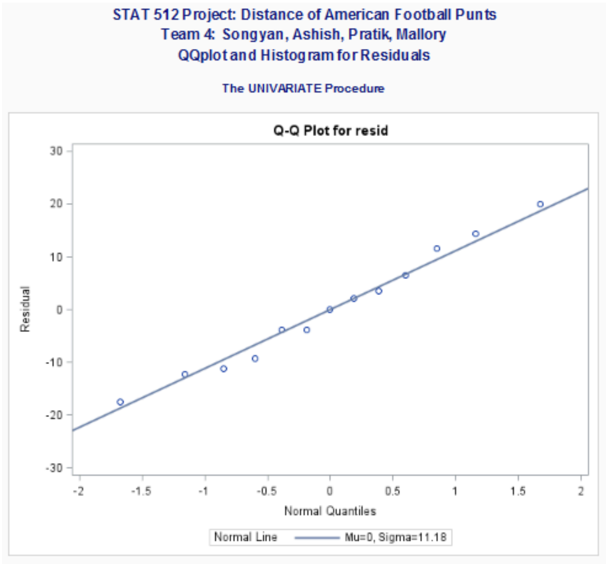


Figure 9. QQplot of the Data

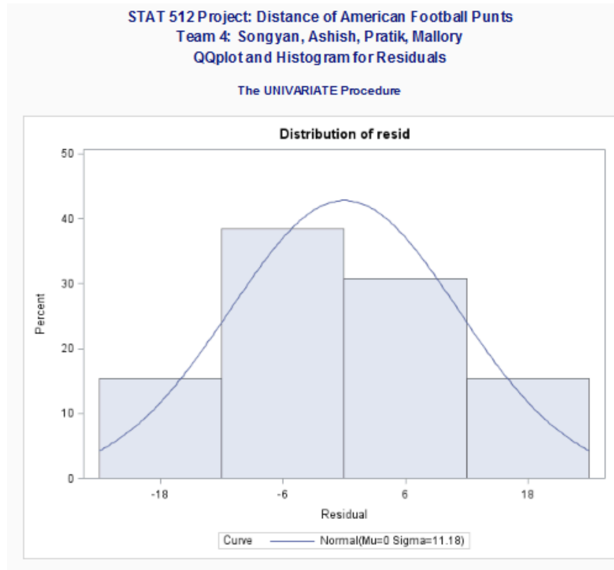


Figure 10. Histogram of the Data

Based on the above plots, there appeared to be no violated assumptions. To check if different transformations would improve the model, the predictor variables were addressed first. As mentioned earlier, R_Strength and L_Strength were the two variables that appeared to be the least linear when plotted against distance. Multiple different transformations were used for both R_Strength and L_Strength to see if any improvements could be made. No transformation seemed to result in any beneficial difference. An example of this can be seen in Appendix 7, which compares the scatterplot of the L_Strength log transformation to the original scatterplot using an untransformed L_Strength variable. Not much – if any – change is seen as a result.

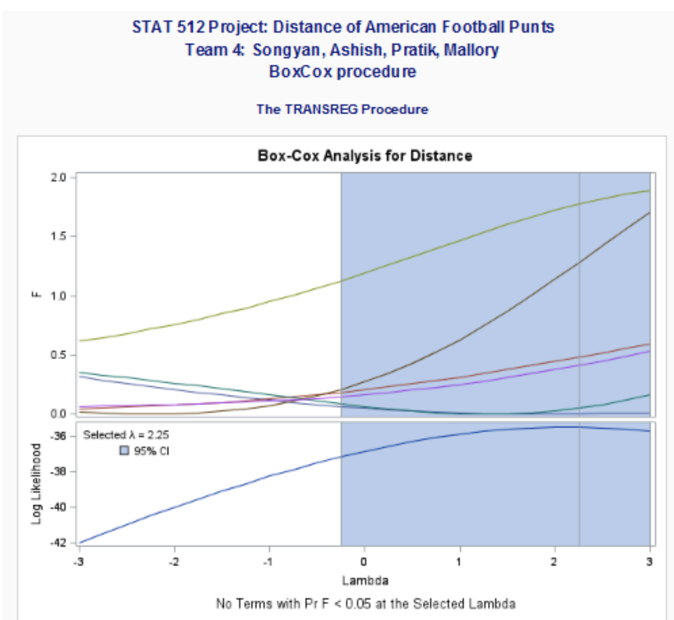


Figure 11. Box-Cox Results

A Box-Cox analysis was then performed to determine the best transformation that could be used for the response variable. The Box-Cox results can be seen in Figure 11, with the selected $\lambda = 2.25$. The response variable of Distance was transformed with this λ value to Distance$^{2.25}$. The transformed response was plotted against the Hang predictor variable and compared to the original scatterplot of Distance and Hang in Appendix 8. Like the attempted transformations in X, this transformation in the Y showed no obvious differences in the plots. A regression for the new model with a response of Distance$^{2.25}$ is found in Appendix 9. The p-values and R^2

values of the transformed and untransformed regressions are given in table 4.

Table 4. Comparison of Regression Results for Models using Transformed Y and No Transformation in Y

	Transformed	No Transformation
p-value	0.0424	0.0473
R^2	0.8221	0.8147

As noticed, both p-values are significant ($< \alpha = 0.05$) and the R^2 value resulting from the regression with the transformed response is greater than the R^2 resulting from the regression with no transformation, but only slightly. Because of this very slight difference in R^2 and no obvious change in the scatterplots between the Hang predictor and the response, it was decided that it was unnecessary to transform the response variable.

Therefore, no transformations were made and the original full model was kept for further analysis.

Question 2: Select Best Subset of Predictor Variables Using C_p Criterion

A best subset of predictor variables was then chosen using the C_p criterion. The one best subset option for each number of predictors in the model was generated and the results are given in Figure 12. All options except for the full model of six predictors appears feasible because their C_p values are less than their respective number of unknowns (p, the number of predictors +1).

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Best model for each model number based on Cp criterion

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

R-Square Selection Method

Number of Observations Read	13
Number of Observations Used	13

Number in Model	R-Square	C(p)	Parameter Estimates						
			Intercept	Hang	R_Strength	L_Strength	R_Flexibility	L_Flexibility	O_Strength
1	0.6705	1.6661	-22.32579	43.50138	.	.	.	.	.
2	0.7845	-0.0243	12.76759	.	0.55632	.	.	.	0.27169
3	0.8058	1.2857	-35.24876	.	0.39144	.	0.85600	.	0.22303
4	0.8138	3.0270	-33.29989	.	0.32558	.	1.33547	-0.40658	0.21789
5	0.8145	5.0038	-33.28882	3.50255	0.29314	.	1.27913	-0.42395	0.20780
6	0.8147	7.0000	-31.26259	2.60809	0.27589	0.03803	1.24223	-0.41339	0.21354

Figure 12. Best Subsets of Predictor Variables Based on Cp Criterion

From the above results, it was determined that the best model would be given by the two predictor variables of R_Strength and O_Strength. Initially, at first glance, it appeared that the model with three predictor variables would be best. Upon further thought, however, a model with two predictor variables was deemed better due to simplicity and the fact that this two-predictor model has the lowest C_p value. Additionally, it only has a slightly lower R^2 value than the best model with three predictor variables.

Question 3: Select Best Subset of Predictor Variables Using Stepwise Option

The stepwise selection option was then used to determine the best combination of predictors and thus, the best model. The results from Figure 13 show that the best model determined through stepwise selection is a two-predictor model using predictor variables of Hang and O_Strength.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Hang		1	0.6705	0.6705	1.6661	22.39	0.0006
2	O_Strength		2	0.0733	0.7438	1.2928	2.86	0.1216

Figure 13. Best Subset of Predictor Variables Using Stepwise Selection

Question 4: Validate Assumptions of "Best" Model

To determine which of these two "best" models is actually the best, regressions were run using the two different subsets of variables to predict the response. The regression results of the two models are compared in Figures 14 & 15.

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Distance vs. Right Leg Strengh and Overall Leg Strength

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6349.36563	3174.68281	18.20	0.0005
Error	10	1743.94285	174.39428		
Corrected Total	12	8093.30848			

Root MSE	13.20584	R-Square	0.7845
Dependent Mean	148.23308	Adj R-Sq	0.7414
Coeff Var	8.90884		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.76759	24.99257	0.51	0.6205
R_Strength	1	0.55632	0.21043	2.64	0.0246
O_Strength	1	0.27169	0.10030	2.71	0.0220

Figure 14. Regression Results of "Best" Cp Model

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Distance vs. Hang Time and Overall Leg Strength

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6020.08016	3010.04008	14.52	0.0011
Error	10	2073.22831	207.32283		
Corrected Total	12	8093.30848			

Root MSE	14.39871	R-Square	0.7438
Dependent Mean	148.23308	Adj R-Sq	0.6926
Coeff Var	9.71356		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.30038	35.80193	-0.04	0.9717
Hang	1	26.89846	12.98516	2.07	0.0651
O_Strength	1	0.22464	0.13278	1.69	0.1216

Figure 15. Regression Results of "Best" Stepwise Model

As recognized from the results, both have p-values of less than α = 0.05, implying that the models are significant and at least one of the predictors is non-zero and helpful in predicting the distance of a punter's punt. Upon further analysis, however, it appears that both predictors in the "best" C_p model are individually significant with p-values less than α = 0.05 (and thus have non-zero coefficients) while only one of the predictors in the "best" stepwise model is individually significant. The R^2 value of the "best" C_p model at 0.7845 is also slightly greater than the R^2 value of the "best" stepwise model at 0.7438. For these two reasons, it was determined that the overall best model based on the six original predictor variables is the "best" model based on the C_p criterion. R_Strength and O_Strength is the combination of predictor variables that best predicts the response of Distance.

To further ensure that this best model makes sense, a correlation matrix was generated and given in Appendix 10. There appears to be a high correlation between the Hang variable and the two selected predictor variables. This observation is beneficial because Hang has proven to be a good predictor of Distance throughout the analysis (Type I SS, being included in the "best" model based on stepwise selection, etc.). R_Strength is also highly correlated with other variables like L_Strength that one would think could be important when determining punt distance.

Listed in Appendix 11 are the residual plots, QQplot, and histogram generated using this new, best model. From these plots it was determined that no assumptions were violated. The data is independent, there appears to be no pattern in the residual plots, the QQplot appears to be approximately linear, and the histogram relatively follows a bell shaped curve. Since there is no pattern or megaphone shape in the residual plots, the constant variance assumption is not violated. The QQplot and histogram give proof that the data is relatively normal and linearity is not violated through the residual plots. No outliers are obvious, either.

Question 5: "Best" Model Prediction and Regression Diagnostics

To further analyze the best selected model, a regression was run and other diagnostics were included. The regression results are given in Figure 14 but other diagnostics are shown in Appendix 12. How to interpret each of the individual regression diagnostics is given in table 5.

Table 5. Regression Diagnostics Interpretation

Diagnostic	Critical Value	How to Interpret	Observations
Student Residual	+3 or -3	If less than -3 or greater than +3	No outliers
Studentized Deleted Residual	$t(n-p-1, 2\alpha/n) = t(13-3-1, 0.05/23) = 4.211$	If greater than	No outliers
Cook's D	$F_{p,n-p}(0.5) = F_{3,10}(0.5) = 0.845$	If greater than	No points with a lot of influence
Hat Diag H	2p/n = 2(3)/13 = 0.461	If greater than	Observation #7 could have influence
DFFITS & DFBETAS	1 because small dataset	If greater than	No influential points

Comparing all of the diagnostic values with their critical values, it appears that the model does not include any outliers and only one observation could potentially have a large amount of influence due to the fact that it is greater than twice as far away from the center of the X's. This potentially influential point is the observation at (180, 147.50) in the scatterplot of Distance and R_Strength (Figure 16). From the plot, it does not appear to be concerning and thus should be left alone. If any outliers or influential points were to exist, weighted linear regression could be used to "fix" them and better model the data.

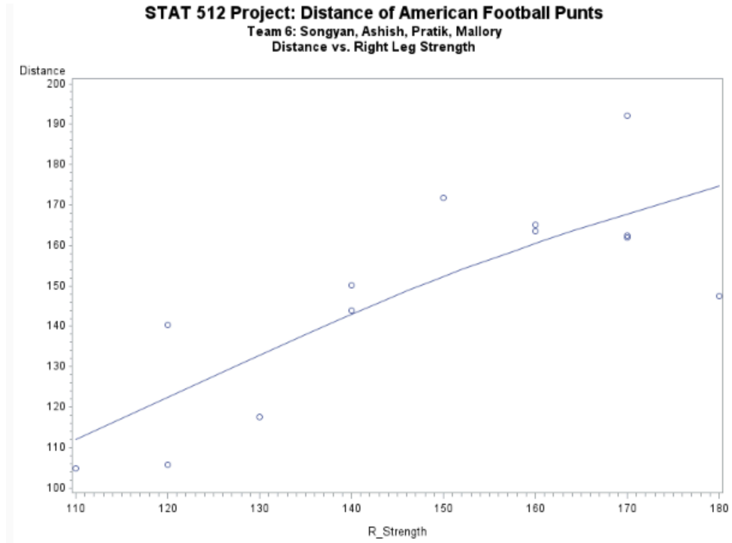


Figure 16. Distance vs. R_Strength

The Variance Inflation Factor (VIF) and/or Tolerance of the model can also be examined for any potential problems in the model. From Appendix 12, the VIF value is 1.58202 and the Tolerance value is 0.6310. Because these two values are recipricals, only one has to be evaluated. The VIF value is significantly less than the critical value of 10. This implies that the squared multiple correlation, R_k^2, is small and the predictors in the model are not well predicted by the other predictos. Thus, there is no multicollinearity present in this determined best model.

Finally, the partial residual plots in Figures 17 & 18 both appear to follow the line with non-zero slope more than the horizontal line. Both predictors of R_Strength and O_Strength appear to be useful in the model.

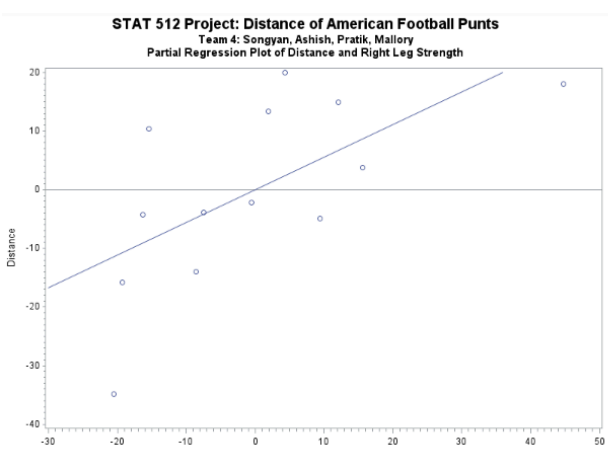


Figure 17. Partial Residual Plot of R_Strength

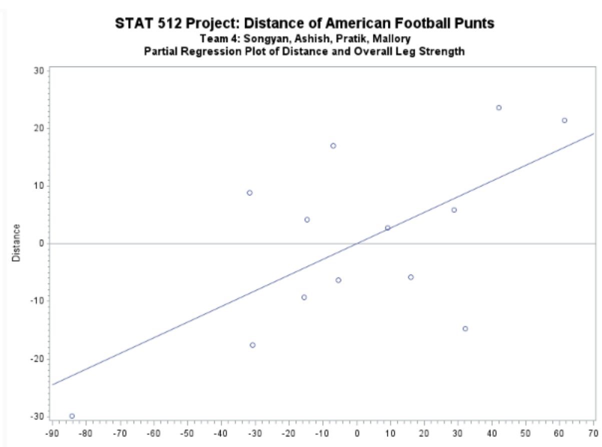


Figure 18. Partial Residual Plot of O_Strength

Question 6: “Best” Model Inferences

Finally, the equation of the regression model that best predicts the respone is given by

$$Distance = 12.76759 + 0.55632 * R\_Strength + 0.27169 * O\_Strength$$

Additionally, 90% Confidence Intervals for the mean of the response variable, distance, and 90% Prediction Intervals for individual observations are given in Figure 19 (and generated with regression results in Appendix 13).

Obs	90% CL Mean		90% CL Predict	
1	102.8377	126.9827	88.1029	141.7175
2	96.4940	123.5259	82.5223	137.4976
3	123.5639	141.7675	107.0584	158.2729
4	130.6736	151.3741	114.9468	167.1010
5	110.5637	132.1239	95.0931	147.5945
6	157.9047	176.3717	141.4839	192.7925
7	135.1064	173.8336	123.6831	185.2570
8	136.5364	150.9919	118.7616	168.7667
9	153.9948	180.0174	139.7631	194.2491
10	147.2508	163.3996	130.0649	180.5855
11	167.5937	191.9314	152.9117	206.6134
12	157.3822	176.4358	141.1476	192.6703
13	162.7878	182.6149	146.7945	198.6083

Figure 19. 90% Confidence Intervals for the Mean of the Reponse and 90% Prediction Intervals for Individual Observations

Finally, the 90% Confidence Intervals for the Regression Coefficients are given below (also generated with regression results in Appendix 13). Notice, both confidence intervals do not contain a zero, thus verifying that both individual tests for R_Strength and O_Strength are significant.

$$90\%\ CI\ for\ \beta_1(coefficient\ for\ R\_Strength): [0.17493, 0.93771]$$
$$90\%\ CI\ for\ \beta_2(coefficient\ for\ O\_Strength): [0.08990, 0.45348]$$

Appendix

Appendix 1 – SAS Code

```
title1 'STAT 512 Project: Distance of American Football Punts';
title2 'Team 4: Songyan, Ashish, Pratik, Mallory';


*Insert the dataset;
data punting;
		input Distance Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength;
		cards;
162.5	4.75	170	170	106	106	240.57
144	4.07	140	130	92	93	195.49
147.5	4.04	180	170	93	78	152.99
163.5	4.18	160	160	103	93	197.09
192	4.35	170	150	104	93	266.56
171.75	4.16	150	150	101	87	260.56
162	4.43	170	180	108	106	219.25
104.93	3.2	110	110	86	92	132.68
105.67	3.02	120	110	90	86	130.24
117.59	3.64	130	120	85	80	205.88
140.25	3.68	120	140	89	83	153.92
150.17	3.6	140	130	92	94	154.64
165.17	3.85	160	150	95	95	240.57
;
*Print the dataset to ensure correctness;
proc print data=punting;
run;
*Run the regression with all predictor variables;
proc reg data=punting;
model Distance = Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength;
run;


*Part I;
*Question 1;
*Sort the data set from shortest hang time to longest;
proc sort data=punting;
  by Hang;
		*Print the dataset to ensure correctness;
		proc print data=punting;
		run;
*Plot the graph of distance vs. hang time;
title3 'Distance vs. Hang Time';
proc gplot data=punting;
symbol1 v=circle i=sm70;
		plot Distance*Hang;
*Create an additional variable cslope depending if the hang time is less or greater than 4.18;
data punting;
 set punting;
		if Hang le 4.18
		  then cslope=0;
		if Hang gt 4.18
		  then cslope=Hang-4.18;
*Print the dataset to ensure correctness;
proc print data=punting;
*Run the regression based on the predictors of hang time and cslope;
title3 'Distance vs. Hang Time and Cslope';
proc reg data=punting;
  model Distance=Hang cslope;
		*Save the output and predicted values;
		output out=punting1 p=Distancehat;
*Print output data to ensure correctness;
proc print data=punting1;


*Plot data with fitted values to obtain piecewise model;
title3 'Distance vs. Hang Time: Piecewise Model';
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=blue;
proc sort data=punting1; by Hang;
proc gplot data=punting1;
  plot (Distance Distancehat)* Hang/overlay;
```

```
*TEST IF SAME LINE;

*Question2;
*Create a variable to be the sum of right leg strength and flexibility;
title3 ' ';
data punting;
set punting;
SUM = R_Strength + R_Flexibility;
*Print dataset to ensure correctness;
proc print data=punting;

*Run the regression for all variables and for all variables including sum;
proc reg data=punting;
title3 'Regression with all predictor variables';
model Distance = Hang  L_Strength L_Flexibility O_Strength;
proc reg data=punting;
title3 'Regression with all predictor variables including the sum';
model Distance = Hang  L_Strength L_Flexibility O_Strength SUM;
t1: test SUM=0;

*Question 3;
title3 'Regression with Type I and II sums of squares';
proc reg data=punting;
model Distance = Hang  L_Strength L_Flexibility O_Strength R_Strength R_Flexibility/ SS1 SS2;

*Question 4;
title3 'Regressions with multiple different models';
proc reg data=punting;
model Distance= Hang;
model Distance= R_Strength;
model Distance= L_Strength;
model Distance= R_Flexibility;
model Distance= L_Flexibility;
model Distance= SUM;
model Distance= O_Strength;
model Distance= R_Strength O_Strength L_Strength SUM;
model Distance= R_Flexibility L_Flexibility SUM;
model Distance= R_Flexibility R_Strength;
model Distance= L_Flexibility L_Strength;
model Distance= Hang R_Strength R_Flexibility;
model Distance= Hang L_Strength L_Flexibility O_Strength R_Strength R_Flexibility;

*Part2;
*Question1;
*Run regression with all 6 variables;
title3 'Regression with all predictor variables';
proc reg data=punting;
model Distance = Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength;
output out=ash r=resid;
run;

*Generate plots for distance vs. all predictor variables individually;
title3 'Distance vs. Hang Time';
proc sort data=punting;
by Hang;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*Hang;
run;

title3 'Distance vs. Right Leg Flexibility';
proc sort data=punting;
by R_Flexibility;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*R_Flexibility;
run;

title3 'Distance vs. Left Leg Flexibility';
proc sort data=punting;
```

```
by L_Flexibility;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*L_Flexibility;
run;

title3 'Distance vs. Right Leg Strength';
proc sort data=punting;
by R_Strength;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*R_Strength;
run;

title3 'Distance vs. Left Leg Strength';
proc sort data=punting;
by L_Strength;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*L_Strength;
run;

title3 'Distance vs. Overall Leg Strength';
proc sort data=punting;
by O_Strength;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance*O_Strength;
run;

*Try to transform L_strength;
data punting;
set punting;
L_Str_new=log10(L_Strength);
proc print data=punting;
run;

title3 'Distance vs. log(Left Leg Strength)';
proc sort data= punting;
by L_Str_new;
symbol1 v=circle i=sm85;
proc gplot data=punting;
plot Distance * L_Str_new;
run;

*Generate histogram and QQplot for the residuals;
title3 'QQplot and Histogram for Residuals';
proc univariate data=ash plot normal;
var resid;
qqplot resid /normal (L=1 mu=est sigma=est);
histogram /normal (L=1 mu=est sigma=est);
run;

*Check BoxCox for potential transformations in Y;
title3 'BoxCox procedure';
proc transreg data=punting;
model boxcox(Distance) = identity(Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength);
run;

*Transform Distance;
title3 'Dataset with Transformed Distance, Y^2.25';
data punting1;
set punting;
Distance1= Distance*Distance*sqrt(sqrt(Distance));
proc print data=punting1;
run;

*Run regression with transformed Y;
title3 'Regression with Transformed Response Variable';
```

```
proc reg data=punting1;
model Distance1 = Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength;
output out=ash1 r=resid1;
run;

*Check relationship between transformed reponse and predictors;
title3 'Transformed Distance vs. Hang Time';
proc sort data= punting1;
by Hang;
symbol1 v=circle i=sm85;
proc gplot data=punting1;
plot Distance1 * Hang;
run;
*Decision: No transformations necessary;

*Question2;
*Selection of Best Model based on Cp Criterion;
title3 'Best model for each model number based on Cp criterion';
proc reg data=punting1;
model Distance = Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength/selection= rsquare cp b best=1;
run;

*Question3;
*Selection of Best Model based on Stepwise;
title3 'Best model based on Stepwise selection';
proc reg data=punting1;
model Distance = Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength/selection= stepwise;
run;

*Question4;
*Check regression for "best" Cp model;
title3 'Distance vs. Right Leg Strength and Overall Leg Strength';
proc reg data=punting1;
model Distance= R_Strength O_Strength;
output out=ash2 r=resid2;
run;

*Check regression for "best" Stepwise model;
title3 'Distance vs. Hang and Overall Leg Strength';
proc reg data=punting1;
model Distance= Hang O_Strength;
output out=ash3 r=resid3;
run;

*Generate histogram and QQplot;
title3 'QQplot and histogram of best model residuals';
proc univariate data=ash3 plot normal;
var resid2;
qqplot resid2/normal(L=1 mu=est sigma=est);
histogram resid2/normal(L=1 mu=est sigma=est);
run;

*Question5;
*Generate regression diagnostics;
title3 'Regression Diagnostics';
proc reg data=punting1;
model Distance = R_Strength O_Strength/r influence;
output out=ash2 r=resid2;
run;

proc reg data=punting1;
model Distance = R_Strength R_Flexibility L_Flexibility O_Strength/tol vif;
output out=ash2 r=resid2;
run;

*Generate Partial Regression Plots;
title3 'Partial Regression Plot of Distance and Right Leg Strength';
proc reg data=punting;
model Distance R_Strength = O_Strength;
```

```
output out=song r=mallory pratik;
symbol1 v=circle i=rl;
axis1 label=('R_Strength');
axis2 label=(angle=90 'Distance');
proc gplot data=song;
plot mallory*pratik/haxis=axis1 vaxis=axis2  vref=0;
run;

title3 'Partial Regression Plot of Distance and Overall Leg Strength';
proc reg data=punting;
model Distance O_Strength = R_Strength;
output out=song2 r=mallory2 pratik2;
symbol1 v=circle i=rl;
axis1 label=('O_Strength');
axis2 label=(angle=90 'Distance');
proc gplot data=song2;
plot mallory2*pratik2/haxis=axis1 vaxis=axis2  vref=0;
run;

title3 'Partial Regression Plot of Right Leg Strength and Overall Leg Strength';
proc reg data=punting;
model O_Strength R_Strength = Distance;
output out=song1 r=mallory1 pratik1;
symbol1 v=circle i=rl;
axis1 label=('R_Strength');
axis2 label=(angle=90 'O_Strength');
proc gplot data=song1;
plot mallory1*pratik1/haxis=axis1 vaxis=axis2  vref=0;
Run;

*Generate correlation coefficients between predictor variables;
title3 'Correlation Coefficients';
proc corr data = punting noprob;

*Question6;
*Generate 90% confidence intervals and prediction intervals;
title3 'Best Model Inferences';
proc reg data=punting1 alpha=0.1;
model Distance=  R_Strength O_Strength/ clb cli clm;
run;
```

Appendix 2 – Regression Output with all Six Predictors

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

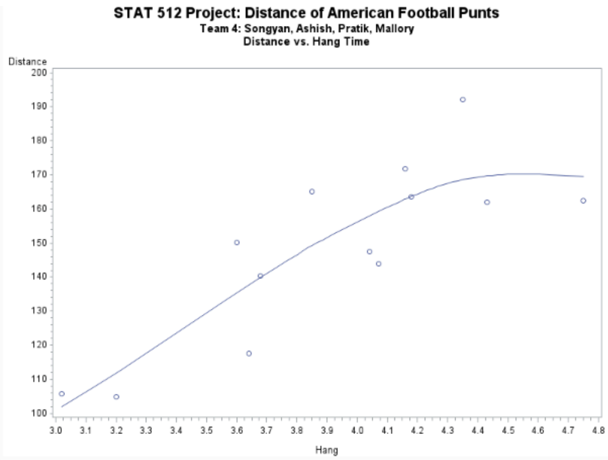
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6593.28721	1098.88120	4.40	0.0473
Error	6	1500.02126	250.00354		
Corrected Total	12	8093.30848			

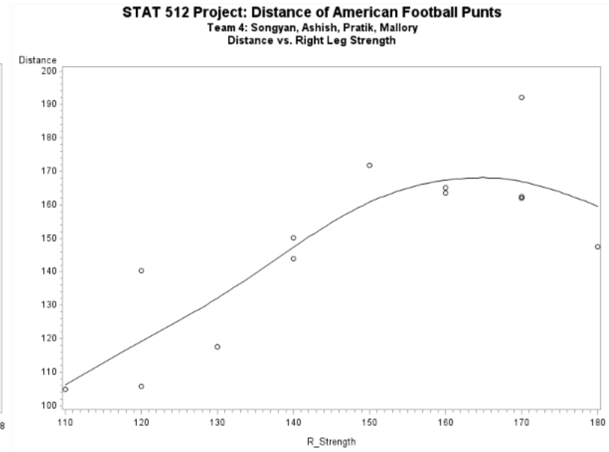
Root MSE	15.81150	R-Square	0.8147
Dependent Mean	148.23308	Adj R-Sq	0.6293
Coeff Var	10.66665		

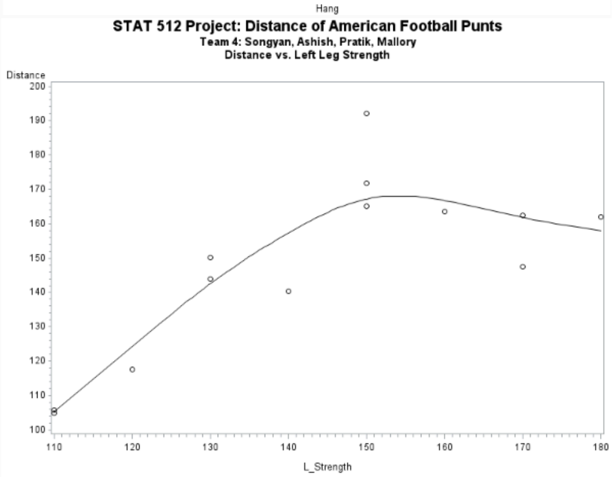
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-31.28259	73.04680	-0.43	0.6836
Hang	1	2.60809	27.19209	0.10	0.9267
R_Strength	1	0.27589	0.49348	0.56	0.5964
L_Strength	1	0.03803	0.61793	0.06	0.9529
R_Flexibility	1	1.24223	1.56422	0.79	0.4574
L_Flexibility	1	-0.41339	0.82548	-0.50	0.6344
O_Strength	1	0.21354	0.17616	1.21	0.2710

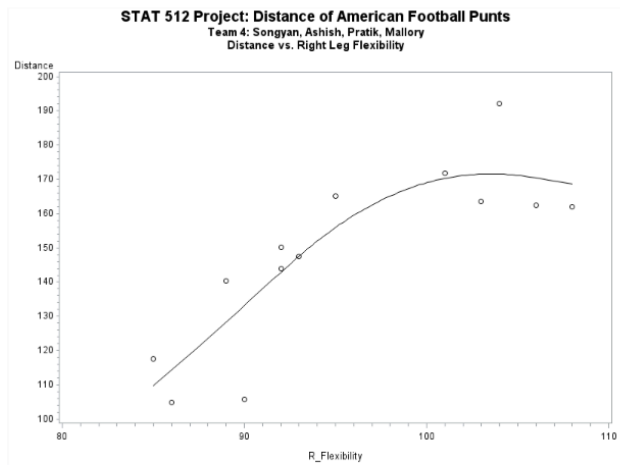
p-value	0.0473
R^2	0.8147

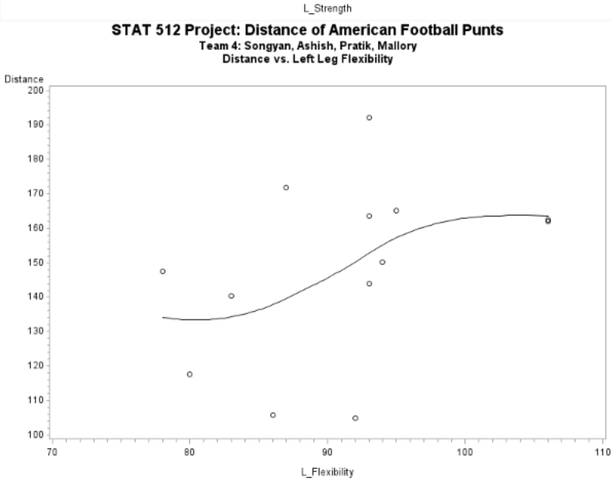
Appendix 3 – Scatterplots of the Individual Variables against Distance using Smoothing of Weight 70

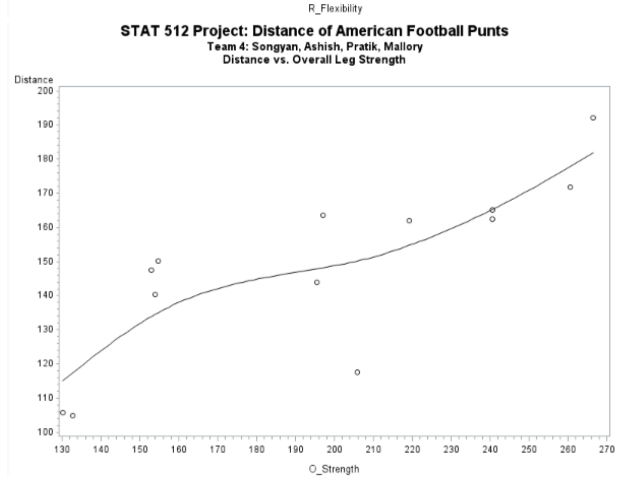












Appendix 4 – Regression Output of Piecewise Model

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Distance vs. Hang Time and Cslope

Obs	Distance	Hang	R_Strength	L_Strength	R_Flexibility	L_Flexibility	O_Strength	cslope	Distancehat
1	105.67	3.02	120	110	90	86	130.24	0.00	101.467
2	104.93	3.20	110	110	86	92	132.68	0.00	111.666
3	150.17	3.60	140	130	92	94	154.64	0.00	134.329
4	117.59	3.64	130	120	85	80	205.88	0.00	136.595
5	140.25	3.68	120	140	89	83	153.92	0.00	138.861
6	165.17	3.85	160	150	95	95	240.57	0.00	148.493
7	147.50	4.04	180	170	93	78	152.99	0.00	159.258
8	144.00	4.07	140	130	92	93	195.49	0.00	160.958
9	171.75	4.16	150	150	101	87	260.56	0.00	166.057
10	163.50	4.18	160	160	103	93	197.09	0.00	167.190
11	192.00	4.35	170	150	104	93	266.56	0.17	167.291
12	162.00	4.43	170	180	108	106	219.25	0.25	167.338
13	162.50	4.75	170	170	106	106	240.57	0.57	167.528

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Distance vs. Hang Time and Cslope

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6001.92814	3000.96407	14.35	0.0012
Error	10	2091.38034	209.13803		
Corrected Total	12	8093.30848			

Root MSE	14.46161	R-Square	0.7416
Dependent Mean	148.23308	Adj R-Sq	0.6899
Coeff Var	9.75599		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-69.63871	44.17124	-1.58	0.1460
Hang	1	56.65758	11.65600	4.86	0.0007
cslope	1	-56.06448	33.80628	-1.66	0.1282

p-value	0.0012
R^2	0.7416

Appendix 5 – Type I and Type II Sum of Squares Analysis

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Regressions with multiple different models

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

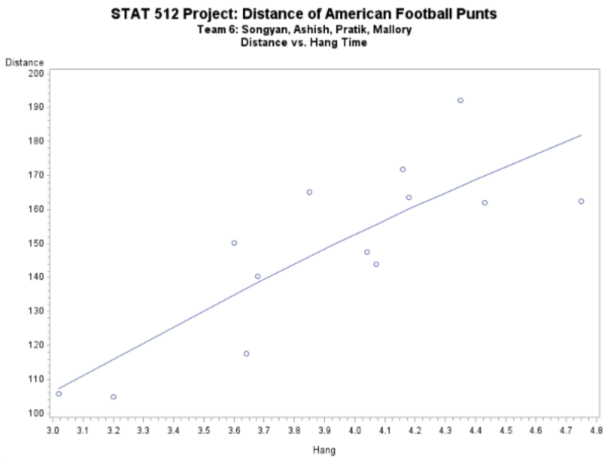
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6593.28721	1098.88120	4.40	0.0473
Error	6	1500.02126	250.00354		
Corrected Total	12	8093.30848			

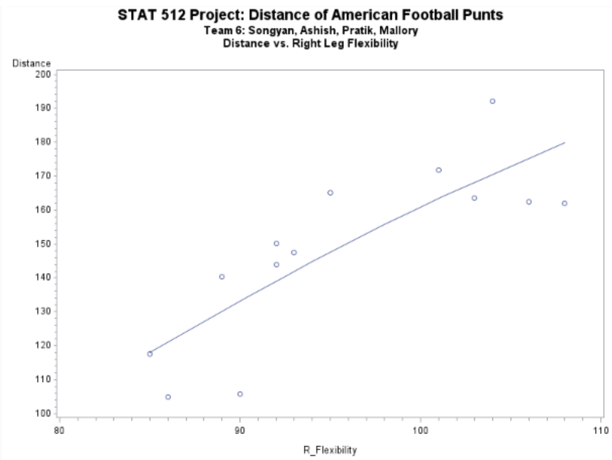
Root MSE	15.81150	R-Square	0.8147
Dependent Mean	148.23308	Adj R-Sq	0.6293
Coeff Var	10.66665		

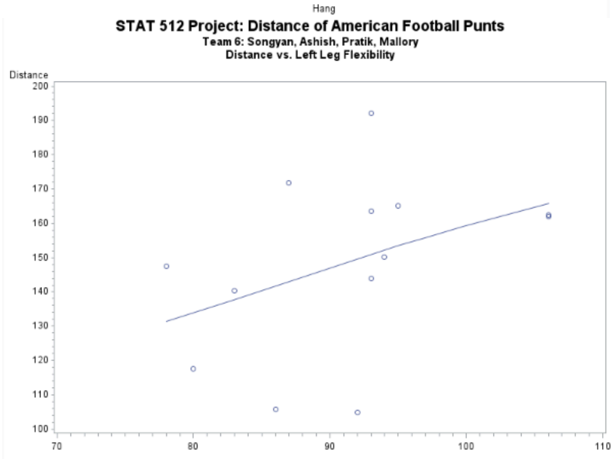
Parameter Estimates

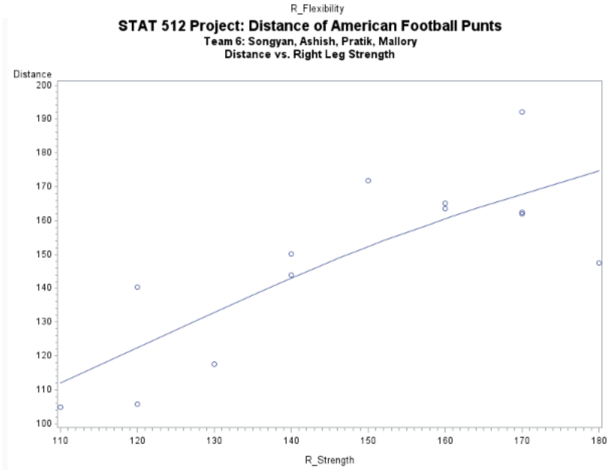
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-31.26259	73.04680	-0.43	0.6836	285650	45.79242
Hang	1	2.60809	27.19209	0.10	0.9267	5426.73476	2.29989
L_Strength	1	0.03803	0.61793	0.06	0.9529	45.56266	0.94680
L_Flexibility	1	-0.41339	0.82548	-0.50	0.6344	6.10052	62.69707
O_Strength	1	0.21354	0.17616	1.21	0.2710	853.30530	367.36136
R_Strength	1	0.27589	0.49348	0.56	0.5964	103.91050	78.13926
R_Flexibility	1	1.24223	1.56422	0.79	0.4574	157.67347	157.67347

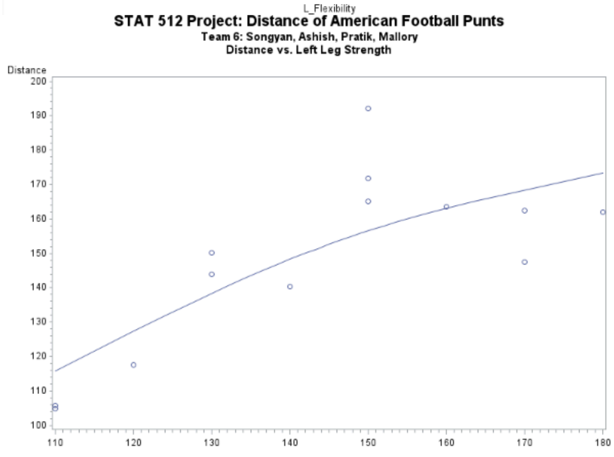
Appendix 6 – Scatterplots of the Individual Variables against Distance using Smoothing of Weight 85

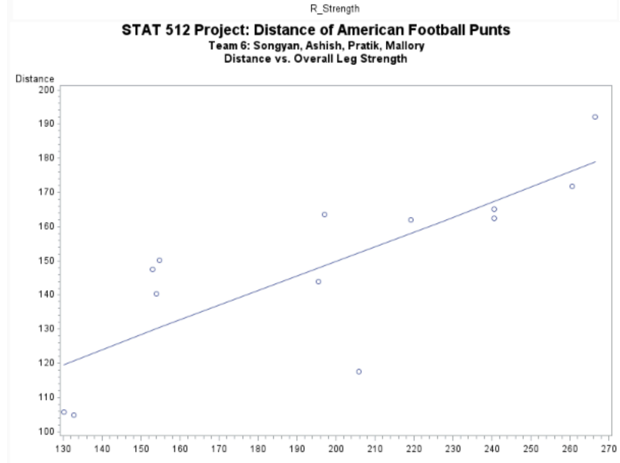






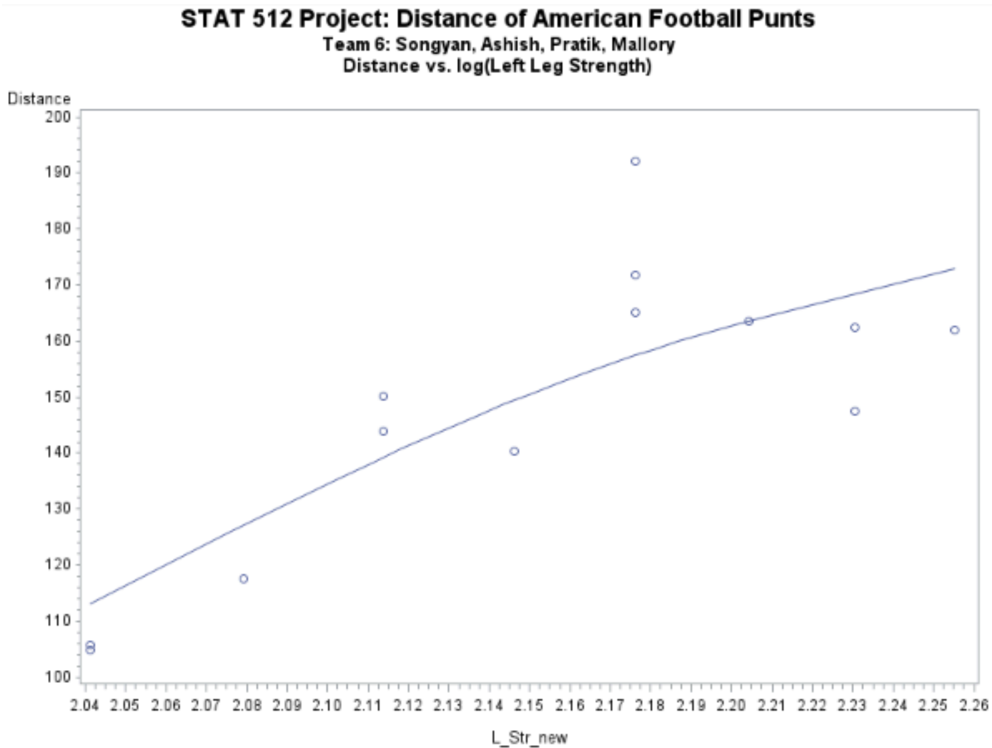




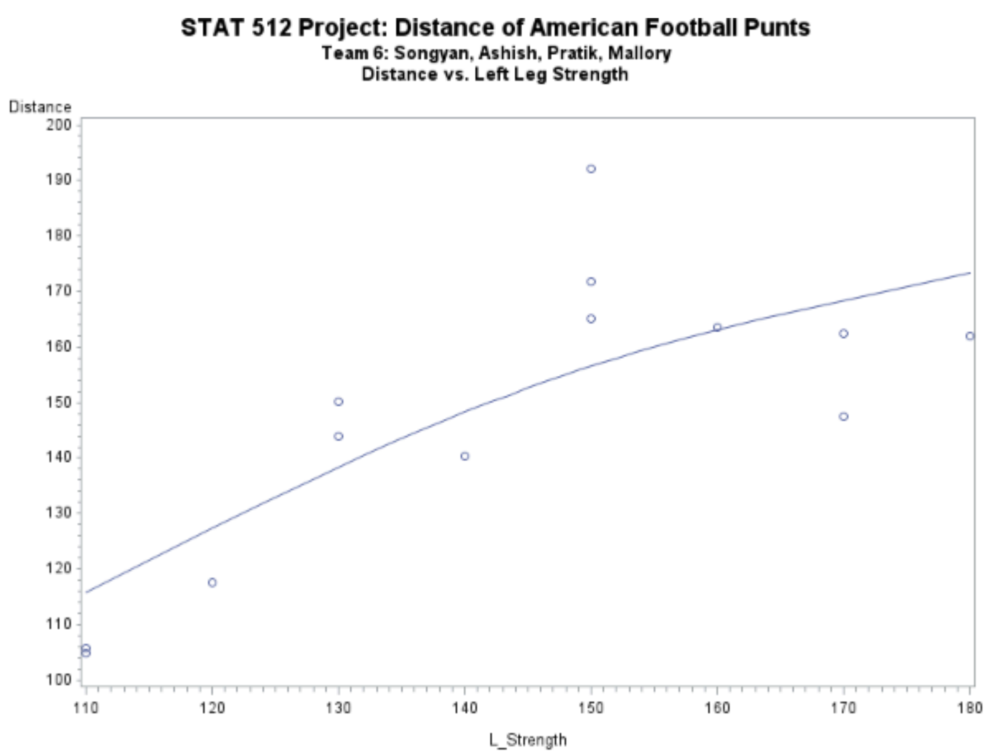


Appendix 7 – Comparison of a Transformed and Untransformed L_Strength Variable

Log Transformation:

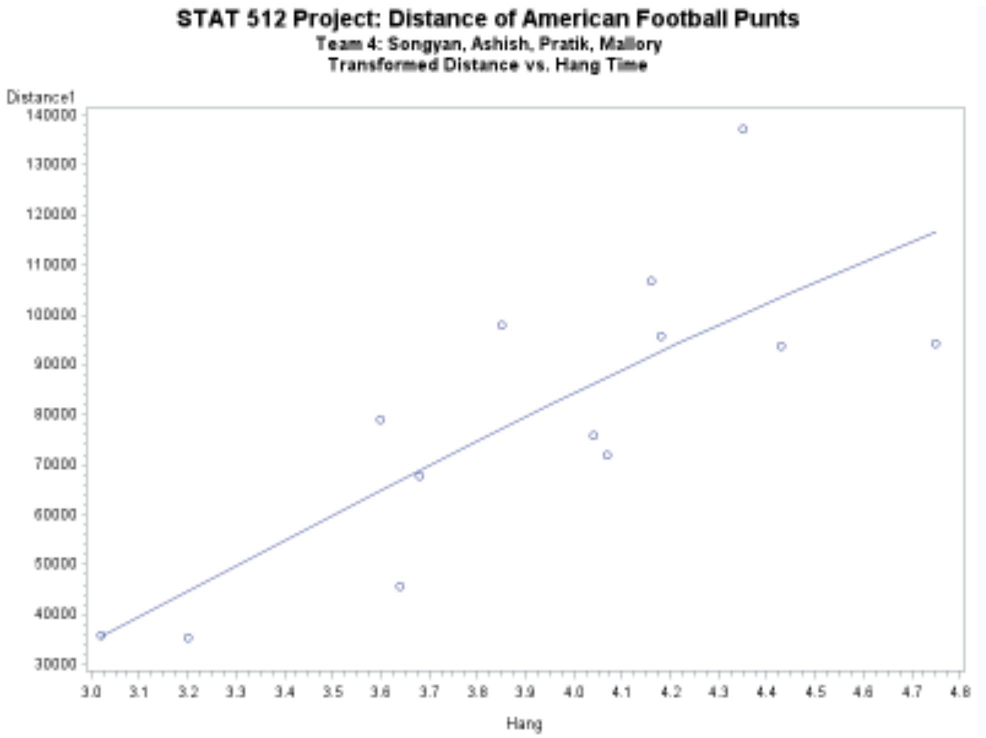


No Transformation:

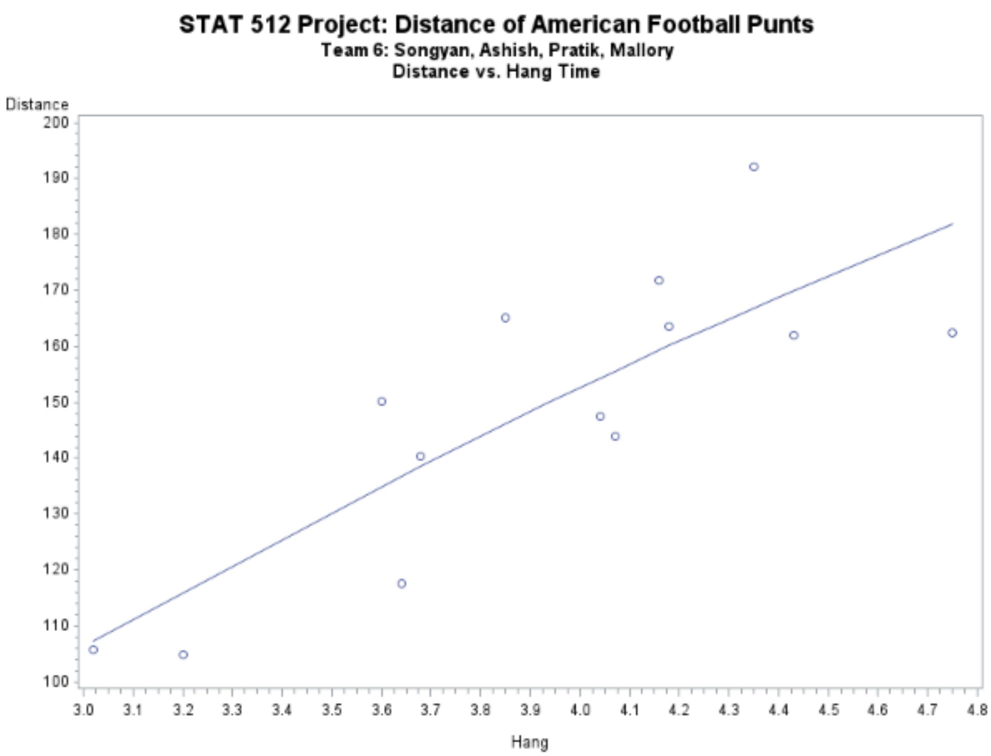


Appendix 8 – Comparison of the Transformed and Untransformed Distance Response

Transformation using λ = 2.25:



No Transformation:



Appendix 9 – Regression Using Model with Transformed Distance

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Regression with Transformed Response Variable

The REG Procedure
Model: MODEL1
Dependent Variable: Distance1

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	8485671232	1414278539	4.62	0.0424
Error	6	1836722952	306120492		
Corrected Total	12	10322394184			

Root MSE	17496	R-Square	0.8221
Dependent Mean	79705	Adj R-Sq	0.6441
Coeff Var	21.95123		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-134216	80830	-1.66	0.1479
Hang	1	-1326.58417	30090	-0.04	0.9663
R_Strength	1	379.46547	546.06674	0.69	0.5131
L_Strength	1	-150.99132	683.77629	-0.22	0.8326
R_Flexibility	1	1957.52512	1730.89353	1.13	0.3013
L_Flexibility	1	-586.35728	913.43893	-0.64	0.5447
O_Strength	1	259.81147	194.93286	1.33	0.2310

p-value	0.0424
R^2	0.8221

Appendix 10 – Correlation Matrix

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Correlation Coefficients

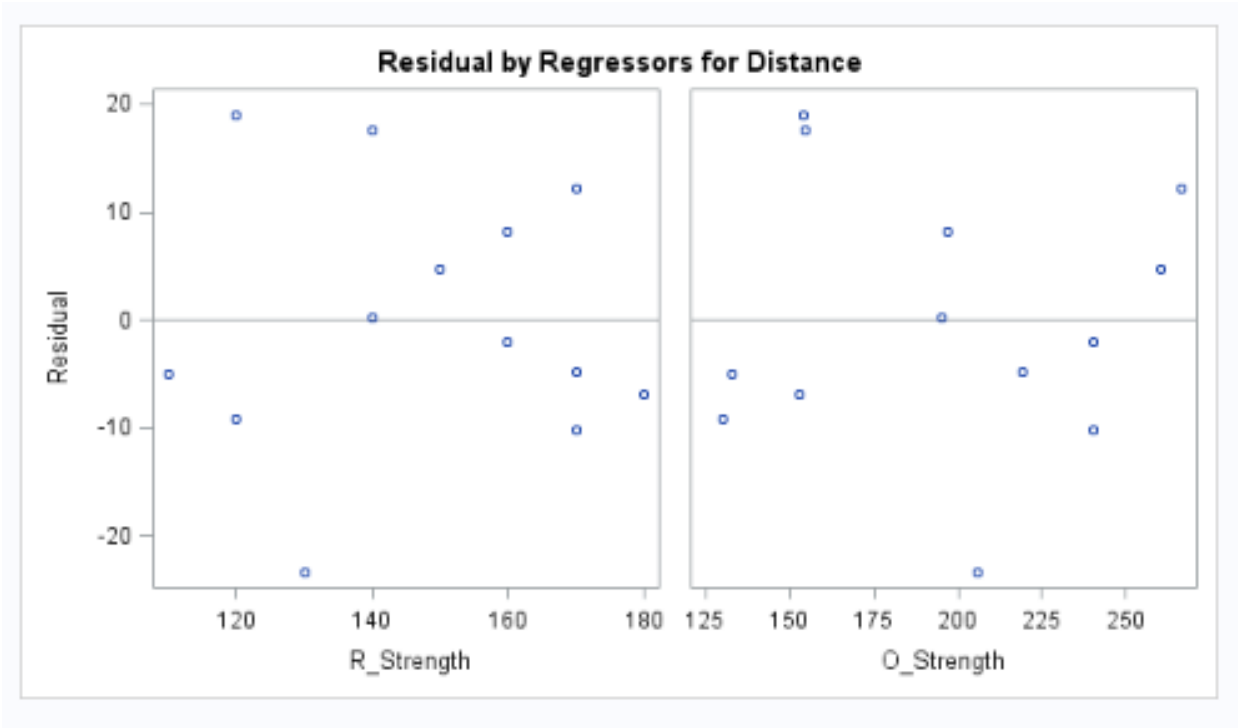
The CORR Procedure

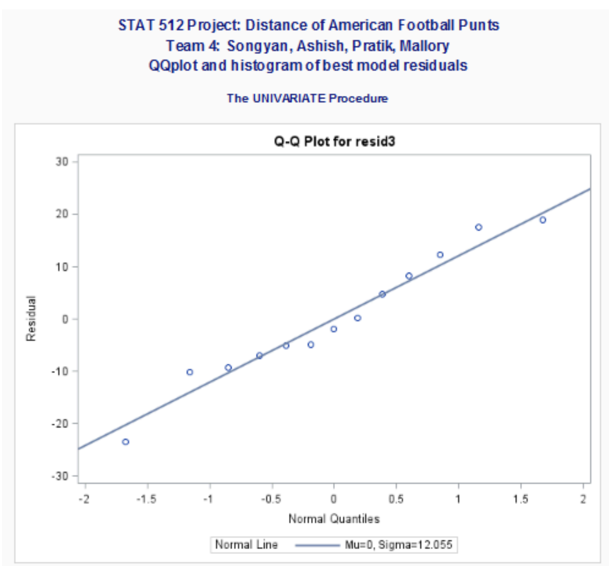
8 Variables	Distance Hang R_Strength L_Strength R_Flexibility L_Flexibility O_Strength cslope

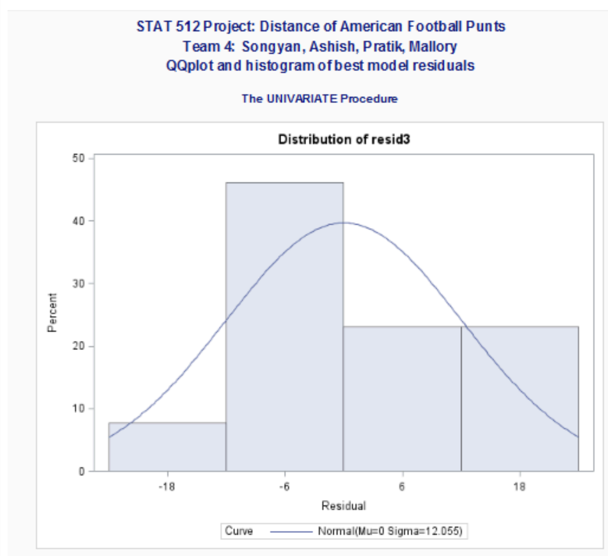
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Distance	13	148.23308	25.97003	1927	104.93000	192.00000
Hang	13	3.92077	0.48885	50.97000	3.02000	4.75000
R_Strength	13	147.69231	22.78664	1920	110.00000	180.00000
L_Strength	13	143.84615	22.92686	1870	110.00000	180.00000
R_Flexibility	13	95.69231	7.79299	1244	85.00000	108.00000
L_Flexibility	13	91.23077	8.57471	1186	78.00000	106.00000
O_Strength	13	196.18769	47.80506	2550	130.24000	266.56000
cslope	13	0.07615	0.16855	0.99000	0	0.57000

Pearson Correlation Coefficients, N = 13								
	Distance	Hang	R_Strength	L_Strength	R_Flexibility	L_Flexibility	O_Strength	cslope
Distance	1.00000	0.81885	0.79147	0.74403	0.80633	0.40774	0.79619	0.36199
Hang	0.81885	1.00000	0.83207	0.86221	0.84508	0.53275	0.75580	0.68060
R_Strength	0.79147	0.83207	1.00000	0.89572	0.77468	0.35695	0.60654	0.47918
L_Strength	0.74403	0.86221	0.89572	1.00000	0.81407	0.42324	0.52308	0.53896
R_Flexibility	0.80633	0.84508	0.77468	0.81407	1.00000	0.68954	0.69028	0.65757
L_Flexibility	0.40774	0.53275	0.35695	0.42324	0.68954	1.00000	0.40812	0.71564
O_Strength	0.79619	0.75580	0.60654	0.52308	0.69028	0.40812	1.00000	0.44500

Appendix 11 – Residual Plots, QQplot, and Histogram for Best Determined Model







Appendix 12 – Regression Diagnostics

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Regression Diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS R_Strength	DFBETAS O_Strength
1	105.6700	114.9102	6.6608	-9.2402	11.403	-0.810	| *| |	0.075	-0.7953	0.2544	1.5006	-0.4646	-0.3640	0.1266	0.2148
2	104.9300	110.0099	7.4572	-5.0799	10.899	-0.466	| | |	0.034	-0.4471	0.3189	1.8855	-0.3059	-0.2799	0.1669	0.0640
3	150.1700	132.6657	5.0218	17.5043	12.214	1.433	| |** |	0.116	1.5253	0.1446	0.8046	0.6271	0.2391	0.1135	-0.3978
4	117.5900	141.0239	5.7106	-23.4339	11.907	-1.968	| ***| |	0.297	-2.3852	0.1870	0.3881	-1.1439	-0.6723	0.8639	-0.6471
5	140.2500	121.3438	5.9478	18.9062	11.791	1.604	| |*** |	0.218	1.7649	0.2029	0.7055	0.8903	0.7695	-0.4874	-0.1055
6	165.1700	167.1382	5.0944	-1.9682	12.184	-0.162	| | |	0.002	-0.1535	0.1488	1.5990	-0.0642	0.0201	0.0014	-0.0363
7	147.5000	154.4700	10.6836	-6.9700	7.762	-0.898	| *| |	0.509	-0.8884	0.6545	3.0852	-1.2227	0.5181	-1.0789	0.9678
8	144.0000	143.7642	3.9878	0.2358	12.589	0.0187	| | |	0.000	0.0178	0.0912	1.5092	0.0056	0.0025	-0.0022	0.0013
9	171.7500	167.0061	7.1788	4.7439	11.084	0.428	| | |	0.026	0.4098	0.2955	1.8421	0.2654	-0.0018	-0.1268	0.2278
10	163.5000	155.3252	4.4549	8.1748	12.432	0.658	| |* |	0.019	0.6378	0.1138	1.3557	0.2285	-0.0744	0.1300	-0.0759
11	192.0000	179.7625	6.7140	12.2375	11.372	1.076	| |** |	0.135	1.0857	0.2585	1.2788	0.6410	-0.3143	0.0394	0.4021
12	162.0000	166.9090	5.2563	-4.9090	12.115	-0.405	| | |	0.010	-0.3876	0.1584	1.5510	-0.1682	0.1003	-0.1053	0.0171
13	162.5000	172.7014	5.4697	-10.2014	12.020	-0.849	| *| |	0.050	-0.8358	0.1716	1.3231	-0.3803	0.2229	-0.1387	-0.1116

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	12.76759	24.99257	0.51	0.6205	.	0
R_Strength	1	0.55632	0.21043	2.64	0.0246	0.63210	1.58202
O_Strength	1	0.27169	0.10030	2.71	0.0220	0.63210	1.58202

Appendix 13 – Final Regression Results including Confidence Intervals and Prediction Intervals

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Best Model Inferences

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	105.6700	114.9102	6.6608	102.8377	126.9827	88.1029	141.7175	-9.2402
2	104.9300	110.0099	7.4572	96.4940	123.5259	82.5223	137.4976	-5.0799
3	150.1700	132.6657	5.0218	123.5639	141.7675	107.0584	158.2729	17.5043
4	117.5900	141.0239	5.7106	130.6736	151.3741	114.9468	167.1010	-23.4339
5	140.2500	121.3438	5.9478	110.5637	132.1239	95.0931	147.5945	18.9062
6	165.1700	167.1382	5.0944	157.9047	176.3717	141.4839	192.7925	-1.9682
7	147.5000	154.4700	10.6836	135.1064	173.8336	123.6831	185.2570	-6.9700
8	144.0000	143.7642	3.9878	136.5364	150.9919	118.7616	168.7667	0.2358
9	171.7500	167.0061	7.1788	153.9948	180.0174	139.7631	194.2491	4.7439
10	163.5000	155.3252	4.4549	147.2508	163.3995	130.0649	180.5855	8.1748
11	192.0000	179.7625	6.7140	167.5937	191.9314	152.9117	206.6134	12.2375
12	162.0000	166.9090	5.2563	157.3822	176.4358	141.1476	192.6703	-4.9090
13	162.5000	172.7014	5.4697	162.7878	182.6149	146.7945	198.6083	-10.2014

STAT 512 Project: Distance of American Football Punts
Team 4: Songyan, Ashish, Pratik, Mallory
Best Model Inferences

The REG Procedure
Model: MODEL1
Dependent Variable: Distance

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6349.36563	3174.68281	18.20	0.0005
Error	10	1743.94285	174.39428		
Corrected Total	12	8093.30848			

Root MSE	13.20584	R-Square	0.7845
Dependent Mean	148.23308	Adj R-Sq	0.7414
Coeff Var	8.90884		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	12.76759	24.99257	0.51	0.6205	-32.53047	58.06566
R_Strength	1	0.55632	0.21043	2.64	0.0246	0.17493	0.93771
O_Strength	1	0.27169	0.10030	2.71	0.0220	0.08990	0.45348