STA6235: spring 2023 Exam 3 (SAS)

Name: Raymond Fleming

Article II. University of West Florida (UWF) Honor Code The University of West Florida's Student Code of Academic Conduct is guided by the following Honor Code: ***As Argonauts, we act with integrity. We do not lie, cheat, steal or tolerate those who do.***

Signed:

Instructions

1. Stay calm. Do not panic.

2. Submit your report to canvas. It is OK to also email me your report to ramin@uwf.edu.

3. It is allowed to use the text and your own notes. You also may use the recorded lectures and notes that I saved at canvas.

4. It is allowed to Google for SAS commands. All links used must be listed.

Use alpha=0.05 in all parts unless it is stated to use a specific alpha value.

5. It is NOT allowed to share information with another person in this exam.

6. A single file is used for your "report". There will be clear and extensive instructions given to you on the exam sheet.

7. For each problem below, give the SAS program ("editor") without showing all data, and list the relevant SAS results. The organization of your report is very important. I expect to see for each part (in each problem) that you show: the problem; SAS program; SAS results; discussion of the results. Do not separate these parts. Clarity is very important in the report write-up. Do not have appendices that I must go through. Do not have multiple files.

8. Submit the completed exam as ONE file (pdf is preferred; or WORD) to canvas. I will have a folder ready for it then. Submit the report by 11:59 PM CST on Monday (April 24).

```
data;
input id group $ distance length rel_dist tailbeat;
rel_dist2= sqrt(rel_dist)/length;
tailbeat2=sqrt(tailbeat)/length;
rel_dist3= (rel_dist**.0625)/length;
tailbeat3=(tailbeat**.0625)/length;
datalines;
1 A 3.3 2.4 1.4 0.44
2 A 3.9 2.3 1.69 0.45
3 A 3 2.1 1.43 0.46
4 A 3.9 2.5 1.54 0.46
5 A 4.1 1.9 2.11 0.47
6 A 3.6 1.9 1.86 0.48
7 A 2.7 1.4 1.89 0.49
8 A 3.7 2.1 1.72 0.49
9 A 2.9 2 1.42 0.49
10 A 2.5 1.8 1.39 0.5
11 A 1.8 2.3 0.78 0.5
12 A 3.2 2.2 1.45 0.5
13 A 2.5 2.5 1.01 0.51
14 A 4.8 2.4 1.98 0.52
15 A 3.6 2.2 1.65 0.53
16 A 2.7 2 1.38 0.53
17 A 3.3 1.9 1.76 0.53
18 A 4 2.1 1.89 0.54
19 A 1.7 2 0.84 0.54
20 A 3.4 2.5 1.38 0.54
21 A 3.4 2.2 1.55 0.54
22 A 2.7 1.9 1.39 0.54
23 A 4.4 2.6 1.72 0.55
24 A 3.7 2 1.88 0.55
25 A 1.2 1.8 0.68 0.56
26 A 4.7 2.2 2.14 0.56
27 A 2.6 2 1.33 0.56
28 A 4.4 2.2 1.98 0.56
29 A 4.5 2.2 1.99 0.56
30 A 4.7 1.8 2.54 0.56
31 A 3.7 1.9 1.95 0.57
32 A 2.4 2.2 1.11 0.57
33 A 3.6 2.6 1.36 0.57
34 A 4.8 2.1 2.32 0.57
35 A 4.1 2.3 1.84 0.58
36 A 3.3 2 1.61 0.58
37 A 2.4 2.2 1.07 0.58
38 A 2.7 1.8 1.51 0.58
39 A 3.2 2.1 1.57 0.58
40 A 2.2 2.1 1.05 0.58
41 A 3.3 1.8 1.8 0.58
42 A 3.7 2.5 1.46 0.59
```

43 A 1.7 1.9 0.86 0.59
44 A 4.7 1.3 3.64 0.59
45 A 2.3 1.8 1.28 0.59
46 A 4.3 2.2 1.95 0.59
47 A 4.1 1.8 2.24 0.59
48 A 2.9 1.8 1.57 0.59
49 A 1.4 2.1 0.69 0.59
50 A 2.3 2.3 1.01 0.59
51 A 2.5 1.6 1.56 0.59
52 A 1.7 2 0.82 0.59
53 A 4 2.3 1.73 0.6
54 A 4.4 1.6 2.78 0.6
55 A 4.3 2.1 2.06 0.6
56 A 4 2.1 1.91 0.61
57 A 4 1.8 2.22 0.61
58 A 5 2 2.54 0.61
59 A 4.9 1.9 2.57 0.61
60 A 3.6 2.1 1.7 0.61
61 A 4.3 1.2 3.42 0.61
62 A 4.2 2.1 2 0.61
63 A 3.5 2 1.76 0.64
64 A 1.8 2 0.94 0.64
65 A 3 1.9 1.57 0.64
66 A 4.3 1.7 2.45 0.64
67 A 2.4 2 1.18 0.64
68 A 2.2 2 1.1 0.65
69 A 3.3 2 1.65 0.65
70 A 2.7 1.8 1.46 0.65
71 A 3 1.9 1.56 0.65
72 A 3.5 1.7 2.01 0.65
73 A 3.7 2.2 1.66 0.65
74 A 3.5 1.9 1.82 0.65
75 A 3.8 1.7 2.23 0.67
76 A 4.4 1.9 2.27 0.67
77 A 3.7 1.9 1.93 0.67
78 A 1.3 1.9 0.71 0.68
79 A 3.9 2 1.94 0.68
80 A 2.5 2 1.24 0.68
81 A 4.8 1.9 2.52 0.68
82 A 3.1 1.9 1.61 0.68
83 A 2.7 1.9 1.39 0.68
84 A 4 1.7 2.3 0.7
85 A 3.8 2.1 1.82 0.7
86 A 2.4 1.7 1.4 0.71
87 A 3 1.6 1.83 0.71
88 A 4.3 2.1 2.09 0.71
89 A 3.1 2.4 1.29 0.73
90 A 2.4 2.1 1.12 0.73
91 A 2.5 2 1.28 0.75

92 A 3.2 1.4 2.2 0.75
93 A 3.5 1.9 1.9 0.75
94 A 4.5 2 2.24 0.75
95 A 4.8 2.6 1.89 0.75
96 A 3.6 1.6 2.3 0.75
97 A 2.7 2.2 1.23 0.75
98 A 2.7 2.3 1.18 0.77
99 A 3.3 2.2 1.5 0.79
100 A 1.9 2.4 0.81 0.79
101 A 2.3 2.4 0.95 0.79
102 A 2.6 1.9 1.39 0.83
103 A 4.8 2.1 2.28 0.83
104 A 3 1.7 1.8 0.86
105 A 4.5 2 2.28 0.86
106 A 4 2.4 1.68 0.86
107 A 2.1 2.3 0.89 0.86
108 A 2 1.9 1.05 0.88
109 A 4.5 1.8 2.44 0.91
110 A 4 2.4 1.65 0.91
111 A 2.4 2.5 0.94 0.94
112 A 2.5 2.3 1.09 0.94
113 A 4.3 2.7 1.63 1
114 A 2.6 2.4 1.11 1
115 A 3 2 1.5 1.2
116 B 2.3 2.4 0.95 0.45
117 B 4 2.6 1.55 0.45
118 B 4.3 2.5 1.68 0.46
119 B 4.3 2.2 1.97 0.46
120 B 3.6 2 1.76 0.47
121 B 4.4 2.1 2.12 0.48
122 B 3.5 1.9 1.84 0.51
123 B 5 2 2.54 0.51
124 B 4.7 1.8 2.54 0.52
125 B 4.6 2.3 2.01 0.52
126 B 3.7 1.8 2.09 0.52
127 B 4.5 1.7 2.58 0.54
128 B 4.5 2.1 2.15 0.54
129 B 2.8 2.2 1.29 0.55
130 B 2.9 2.1 1.39 0.55
131 B 2.4 1.9 1.24 0.56
132 B 3.5 2.3 1.55 0.56
133 B 3.9 1.8 2.2 0.56
134 B 2.3 2.3 1 0.56
135 B 1.4 1.9 0.76 0.56
136 B 4.3 1.8 2.38 0.56
137 B 2.6 2 1.28 0.56
138 B 4.1 2.3 1.76 0.57
139 B 4.5 1.9 2.34 0.57
140 B 1.9 2.1 0.91 0.57

```
141 B 2.6 2 1.34 0.57
142 B 4.1 2.1 1.94 0.57
143 B 1.5 2.3 0.68 0.58
144 B 2.5 1.7 1.42 0.58
145 B 2.5 2.4 1.08 0.58
146 B 3 2 1.52 0.59
147 B 2.4 2 1.19 0.59
148 B 2.1 2.1 0.97 0.59
149 B 2.1 1.8 1.21 0.59
150 B 2.7 2 1.35 0.59
;

/* part A */
proc reg; model distance=tailbeat /cli p; where group eq 'A';run; /* staring group */
proc reg; model distance=tailbeat /cli p; where group eq 'B';run; /* interrupted eye contact group */
```

The REG Procedure
Model: FullEyeContact
Dependent Variable: distance

Number of Observations Read	115
Number of Observations Used	115

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22585	0.22585	0.26	0.6089
Error	113	96.94597	0.85793		
Corrected Total	114	97.17183			

Root MSE	0.92624	R-Square	0.0023
Dependent Mean	3.32435	Adj R-Sq	-0.0065
Coeff Var	27.86244		

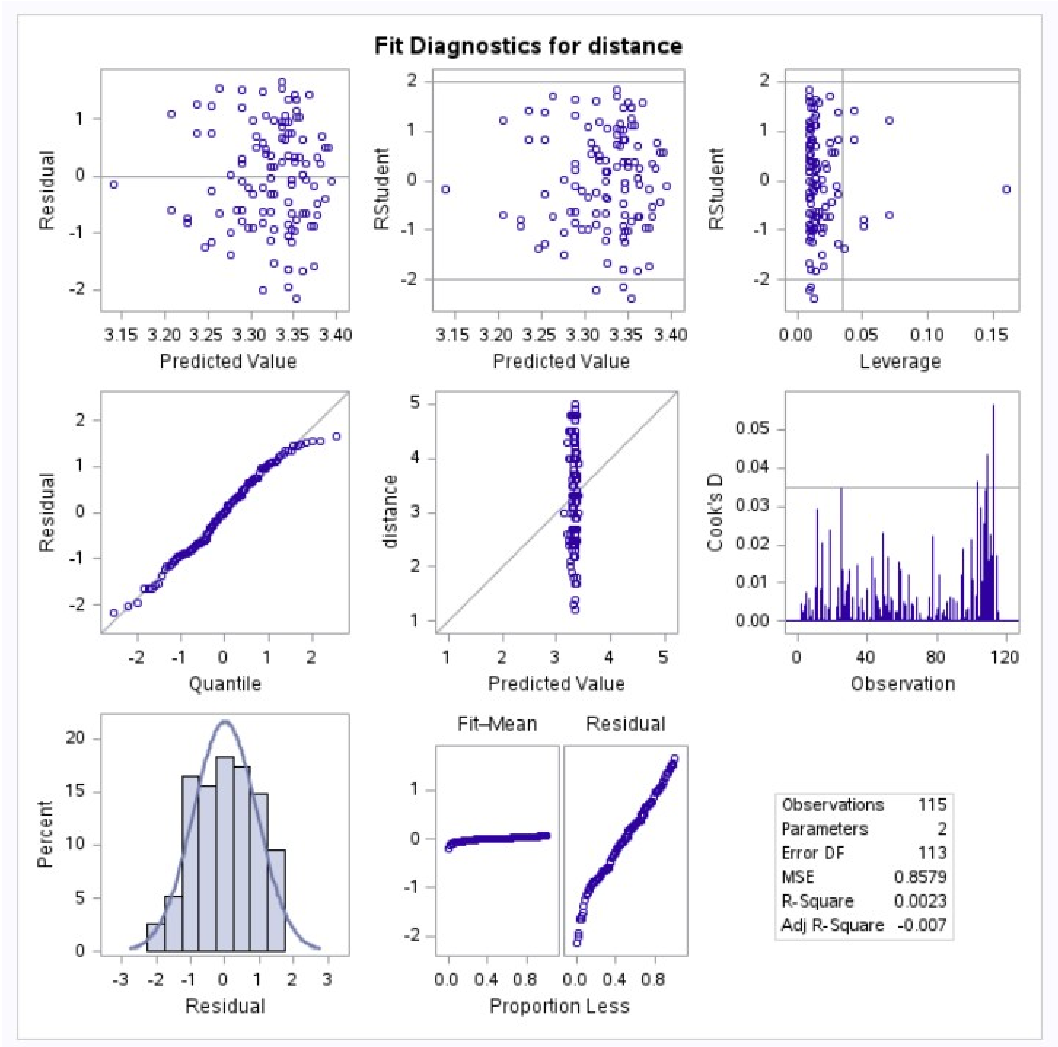
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.54141	0.43178	8.20	<.0001
tailbeat	1	-0.33502	0.65295	-0.51	0.6089

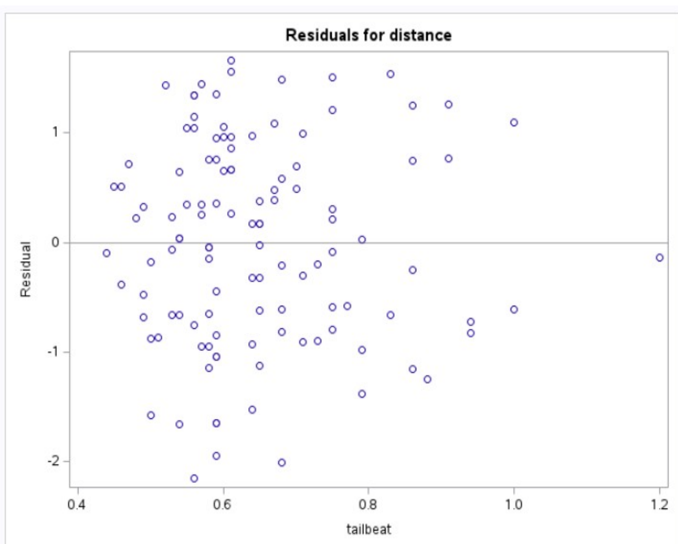
The REG Procedure
Model: FullEyeContact
Dependent Variable: distance

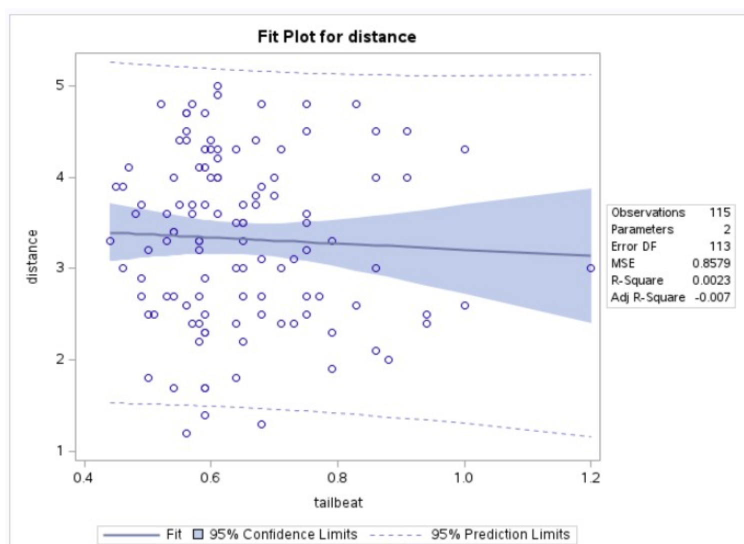
Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	3.3	3.3940	0.1609	1.5315	5.2565	-0.0940
2	3.9	3.3907	0.1554	1.5299	5.2514	0.5093
3	3.0	3.3873	0.1500	1.5283	5.2463	-0.3873
4	3.9	3.3873	0.1500	1.5283	5.2463	0.5127
5	4.1	3.3840	0.1448	1.5266	5.2413	0.7160
6	3.6	3.3806	0.1396	1.5248	5.2364	0.2194
7	2.7	3.3773	0.1345	1.5229	5.2316	-0.6773
8	3.7	3.3773	0.1345	1.5229	5.2316	0.3227
9	2.9	3.3773	0.1345	1.5229	5.2316	-0.4773
10	2.5	3.3739	0.1296	1.5210	5.2268	-0.8739
11	1.8	3.3739	0.1296	1.5210	5.2268	-1.5739
12	3.2	3.3739	0.1296	1.5210	5.2268	-0.1739
13	2.5	3.3706	0.1248	1.5189	5.2222	-0.8706
14	4.8	3.3672	0.1201	1.5168	5.2176	1.4328
15	3.6	3.3639	0.1157	1.5145	5.2132	0.2361
16	2.7	3.3639	0.1157	1.5145	5.2132	-0.6639
17	3.3	3.3639	0.1157	1.5145	5.2132	-0.0639
18	4.0	3.3605	0.1115	1.5122	5.2088	0.6395
19	1.7	3.3605	0.1115	1.5122	5.2088	-1.6605
20	3.4	3.3605	0.1115	1.5122	5.2088	0.0395
21	3.4	3.3605	0.1115	1.5122	5.2088	0.0395
22	2.7	3.3605	0.1115	1.5122	5.2088	-0.6605
23	4.4	3.3572	0.1075	1.5098	5.2045	1.0428
24	3.7	3.3572	0.1075	1.5098	5.2045	0.3428
25	1.2	3.3538	0.1037	1.5073	5.2003	-2.1538
26	4.7	3.3538	0.1037	1.5073	5.2003	1.3462
27	2.6	3.3538	0.1037	1.5073	5.2003	-0.7538
28	4.4	3.3538	0.1037	1.5073	5.2003	1.0462
29	4.5	3.3538	0.1037	1.5073	5.2003	1.1462

Sum of Residuals	0
Sum of Squared Residuals	96.94597
Predicted Residual SS (PRESS)	100.09438


Fit Diagnostics for distance
Residual
Predicted Value
RStudent
Leverage
Quantile
distance
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 115
Parameters 2
Error DF 113
MSE 0.8579
R-Square 0.0023
Adj R-Square -0.007


Residuals for distance
Residual
tailbeat


Fit Plot for distance
distance
tailbeat
Observations 115
Parameters 2
Error DF 113
MSE 0.8579
R-Square 0.0023
Adj R-Square -0.007
Fit
95% Confidence Limits
95% Prediction Limits

For the full, uninterrupted eye contact group, the coefficient of determination (R^2) was found to be 0.002, with the adjusted version found to be -0.0065. The intercept estimate for this model is 3.54 with a standard error of .432. This indicates that at zero tailbeat, the model would predict a distance of 3.54, with a 95% likelihood that the distance somewhere between 2.69 and 4.39 assuming this is a normal distribution. The t-test shows that the intercept is significant at p = less than 0.0001. This indicates that there is sufficient evidence to suggest that the intercept is not zero.

The slope of the model was found to be -0.33502 with a standard error of 0.65295, indicating that there is a 95% likelihood that the true slope is between -1.615 and .945. In the case that the slope is truly negative, this would indicate that tailbeat frequency decreases as distance increases. However, in this case a 95% confidence interval includes zero, therefore there is not sufficient evidence to indicate that tailbeat frequency decreases with distance. The t-test shows that the slope is not significant at p=.6089, which is greater than the cutoff of 0.05 for alpha=0.05. This indicates that there is not sufficient evidence to suggest that the slope is not zero.

The overall model was not found to be significant at alpha=0.05 with a p-value of 0.6089. The root mean square error was found to be 0.92624 which is a mean square error of 0.8579. PRESS was found to be 100.09438. Given that the coefficient of determination is near zero, and that this model is not significant would indicate that this model is a poor fit for the data.

There is either a better model which could be found, or it is possible that distance from diver and tailbeat are in no way correlated. PRESS is not high enough to be of great concern, but given the other fit indicators it can likely be much lower with a better model (assuming there is a better model in this case).

The REG Procedure
Model: InterruptedEyeContact
Dependent Variable: distance

Number of Observations Read	35
Number of Observations Used	35

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.12265	8.12265	9.63	0.0039
Error	33	27.83735	0.84356		
Corrected Total	34	35.96000			

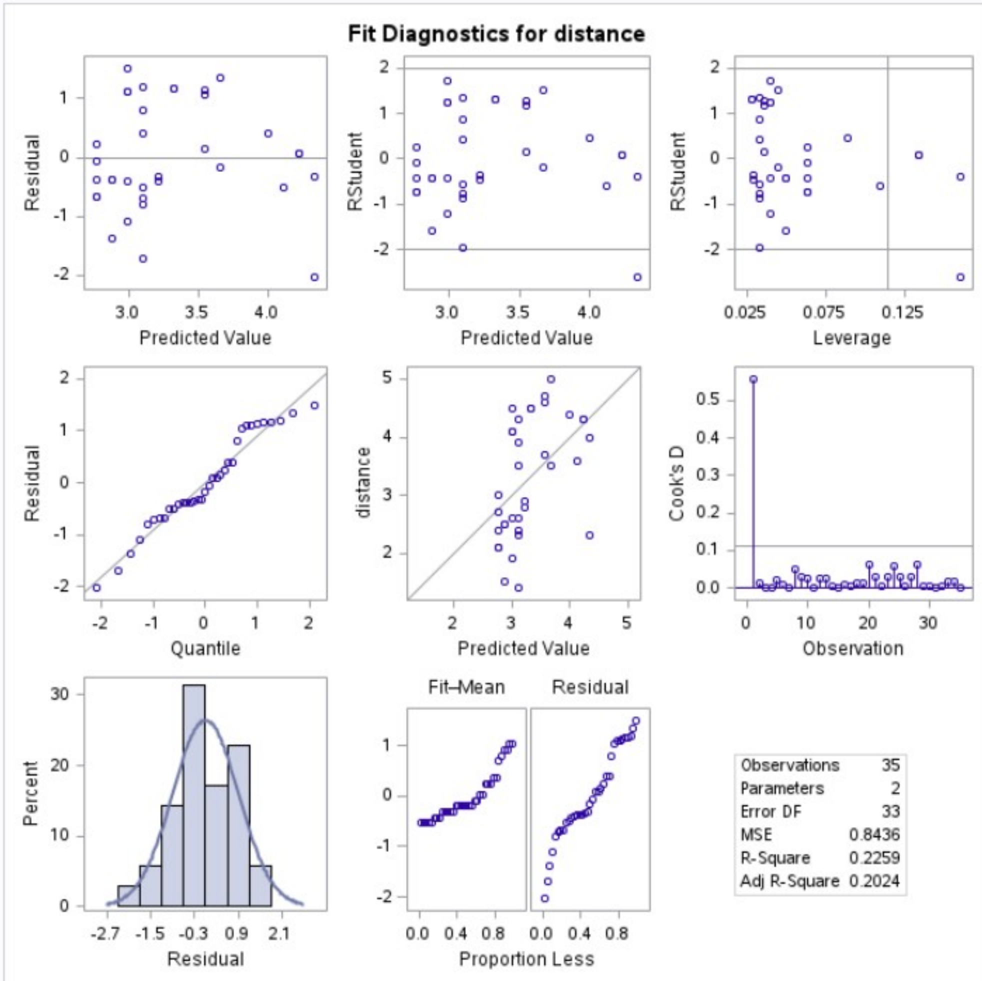
Root MSE	0.91845	R-Square	0.2259
Dependent Mean	3.30000	Adj R-Sq	0.2024
Coeff Var	27.83191		

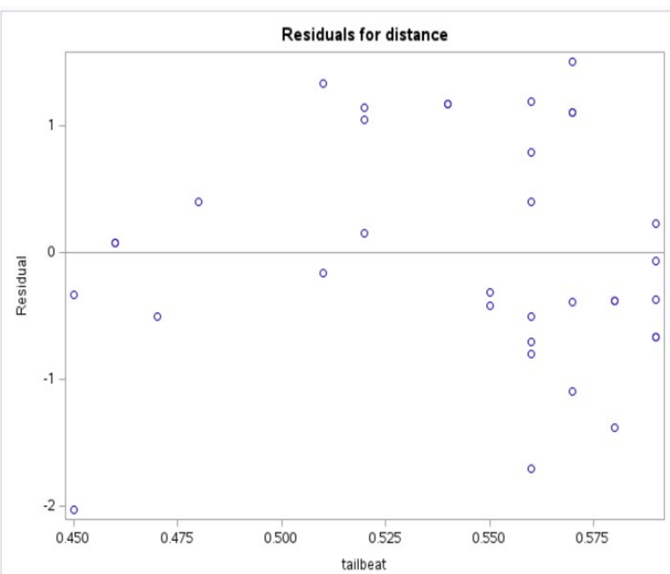
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.36206	1.95973	4.78	<.0001
tailbeat	1	-11.17283	3.60057	-3.10	0.0039

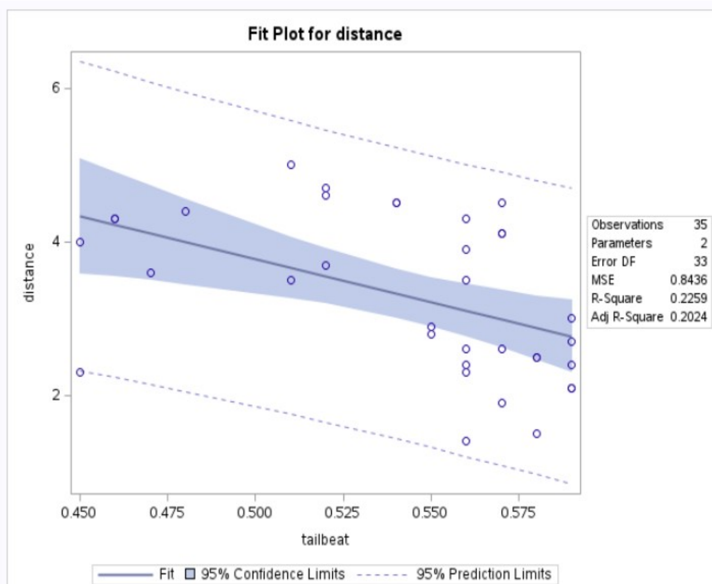
The REG Procedure
Model: InterruptedEyeContact
Dependent Variable: distance

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	2.3	4.3343	0.3677	2.3215	6.3471	-2.0343
2	4.0	4.3343	0.3677	2.3215	6.3471	-0.3343
3	4.3	4.2226	0.3354	2.2333	6.2119	0.0774
4	4.3	4.2226	0.3354	2.2333	6.2119	0.0774
5	3.6	4.1108	0.3039	2.1426	6.0791	-0.5108
6	4.4	3.9991	0.2736	2.0493	5.9489	0.4009
7	3.5	3.6639	0.1946	1.7538	5.5740	-0.1639
8	5.0	3.6639	0.1946	1.7538	5.5740	1.3361
9	4.7	3.5522	0.1752	1.6499	5.4545	1.1478
10	4.6	3.5522	0.1752	1.6499	5.4545	1.0478
11	3.7	3.5522	0.1752	1.6499	5.4545	0.1478
12	4.5	3.3287	0.1555	1.4335	5.2239	1.1713
13	4.5	3.3287	0.1555	1.4335	5.2239	1.1713
14	2.8	3.2170	0.1575	1.3211	5.1129	-0.4170
15	2.9	3.2170	0.1575	1.3211	5.1129	-0.3170
16	2.4	3.1053	0.1675	1.2059	5.0047	-0.7053
17	3.5	3.1053	0.1675	1.2059	5.0047	0.3947
18	3.9	3.1053	0.1675	1.2059	5.0047	0.7947
19	2.3	3.1053	0.1675	1.2059	5.0047	-0.8053
20	1.4	3.1053	0.1675	1.2059	5.0047	-1.7053
21	4.3	3.1053	0.1675	1.2059	5.0047	1.1947
22	2.6	3.1053	0.1675	1.2059	5.0047	-0.5053
23	4.1	2.9935	0.1840	1.0878	4.8993	1.1065
24	4.5	2.9935	0.1840	1.0878	4.8993	1.5065
25	1.9	2.9935	0.1840	1.0878	4.8993	-1.0935
26	2.6	2.9935	0.1840	1.0878	4.8993	-0.3935
27	4.1	2.9935	0.1840	1.0878	4.8993	1.1065
28	1.5	2.8818	0.2056	0.9670	4.7967	-1.3818
29	2.5	2.8818	0.2056	0.9670	4.7967	-0.3818

Sum of Residuals	0
Sum of Squared Residuals	27.83735
Predicted Residual SS (PRESS)	31.62756


Fit Diagnostics for distance
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
distance
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 35
Parameters 2
Error DF 33
MSE 0.8436
R-Square 0.2259
Adj R-Square 0.2024


Residuals for distance
Residual
tailbeat


Fit Plot for distance
distance
tailbeat
Observations 35
Parameters 2
Error DF 33
MSE 0.8436
R-Square 0.2259
Adj R-Square 0.2024
Fit
95% Confidence Limits
95% Prediction Limits

For the interrupted eye contact group, the coefficient of determination (R^2) was found to be 0.2259, with the adjusted version found to be 0.2024 The intercept estimate for this model is 9.362 with a standard error of 1.95973. This indicates that at zero tailbeat, the model would predict a distance of 9.362, with a 95% likelihood that the distance somewhere between 5.521 and 13.203 assuming this is a normal distribution.   The t-test on the intercept was found to be significant at p = less than 0.0001. This indicates that there is sufficient evidence to suggest that the intercept is not zero.

The slope of the model was found to be -11.17283 with a standard error of 3.600, indicating that there is a 95% likelihood that the true slope is between -18.229 and -4.117. This would indicate that there is a high likelihood that tailbeat frequency decreases with distance, but the overall range of how much it decreases is quite large. The t-test on the slope was found to be significant at p=0.0039. This indicates that there is sufficient evidence to suggest that the slope is not zero.

The overall model was found to be significant at alpha=0.05 with a p-value of 0.0039.   The root mean square error was found to be 0.91845 which is a mean square error of 0.84355.   PRESS was found to be 31.628.   This model is a much better fit than the model found for the uninterrupted eye contact group. However, given the lower R^2 value, it is likely that a transform will allow for a better model. The quantile-quantile plot appears to indicate this as well.

```
/* part B data unchanged from part A*/
proc reg; FullEyeContactTransform1: model rel_dist2=tailbeat2 /cli p; where group eq 'A';run;
proc reg; InterruptedEyeContactTransform1: model rel_dist2=tailbeat2 /cli p; where group eq 'B';run;
```

The REG Procedure
Model: FullEyeContactTransform1
Dependent Variable: rel_dist2

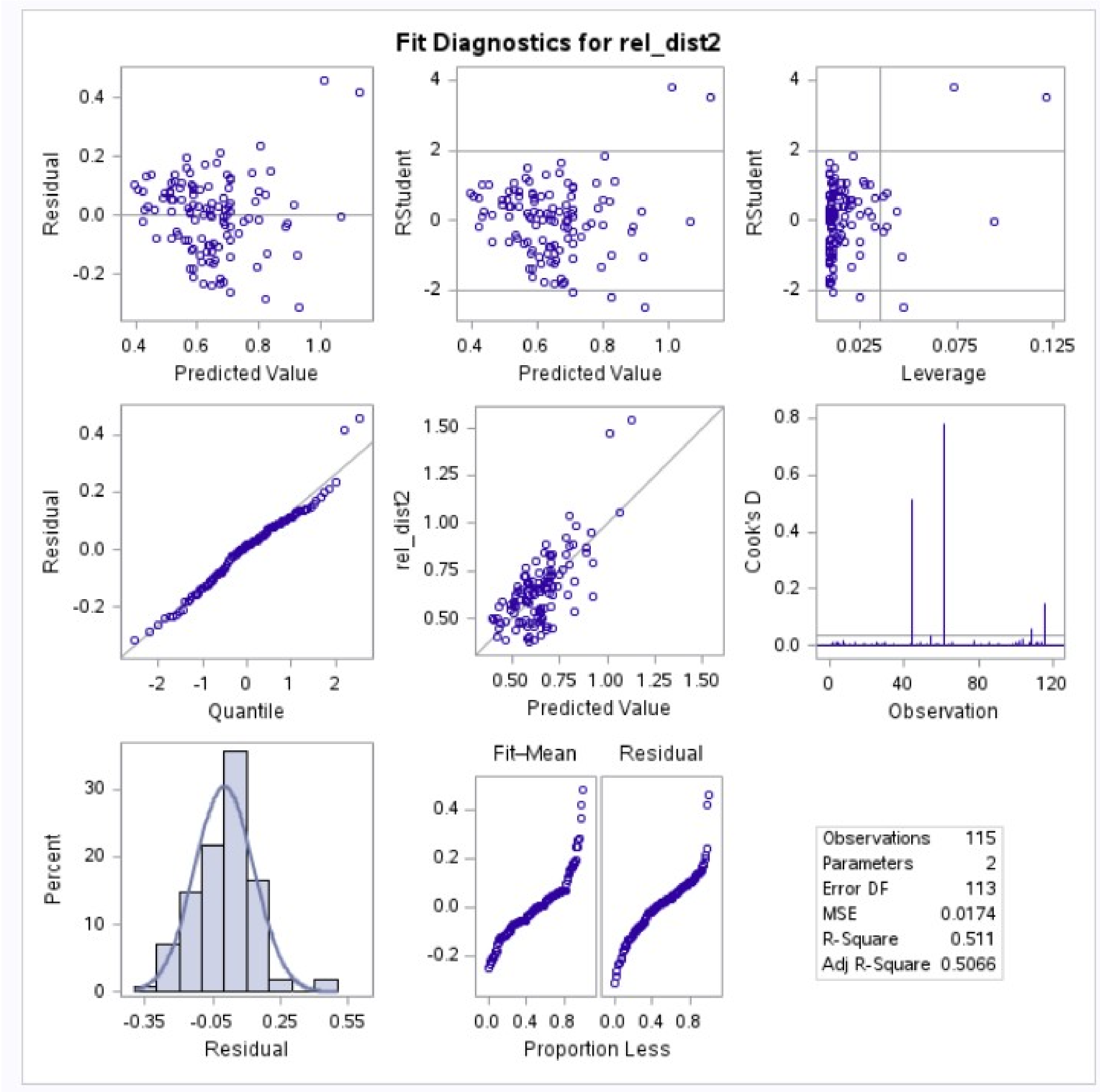
Number of Observations Read	115
Number of Observations Used	115

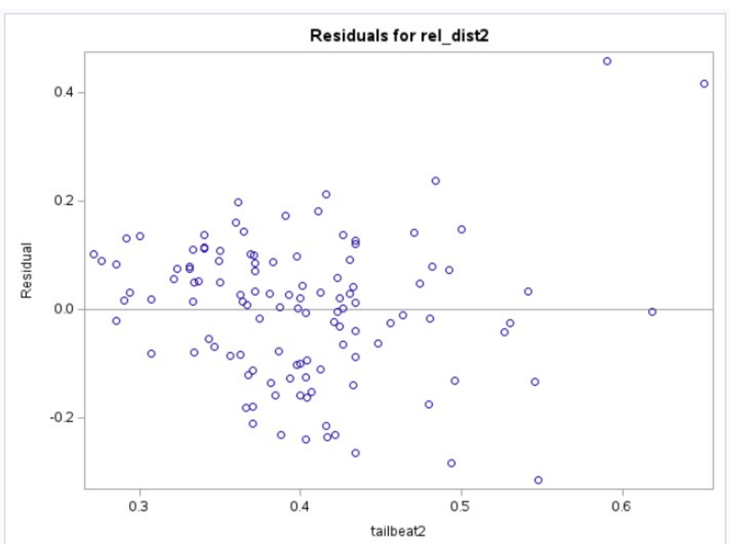
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.05389	2.05389	118.06	<.0001
Error	113	1.96579	0.01740		
Corrected Total	114	4.01968			

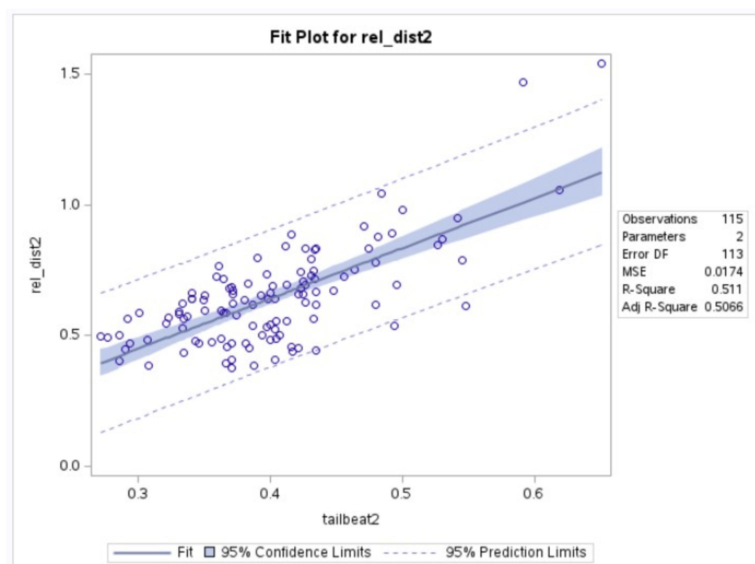
Root MSE	0.13190	R-Square	0.5110
Dependent Mean	0.64453	Adj R-Sq	0.5066
Coeff Var	20.46380		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.12708	0.07207	-1.76	0.0806
tailbeat2	1	1.92457	0.17712	10.87	<.0001

Sum of Residuals	0
Sum of Squared Residuals	1.96579
Predicted Residual SS (PRESS)	2.10364


Fit Diagnostics for rel_dist2
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
rel_dist2
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 115
Parameters 2
Error DF 113
MSE 0.0174
R-Square 0.511
Adj R-Square 0.5066


Residuals for rel_dist2
Residual
tailbeat2


Fit Plot for rel_dist2
rel_dist2
tailbeat2
Observations 115
Parameters 2
Error DF 113
MSE 0.0174
R-Square 0.511
Adj R-Square 0.5066
Fit
95% Confidence Limits
95% Prediction Limits

For the full, uninterrupted eye contact group when transformed with rel_dist2=sqrt(rel_dist)/length and tailbeat2=sqrt(tailbeat)/length, the coefficient of determination (R^2) was found to be 0.5110, with the adjusted version found to be .5066. The intercept estimate for this model is -.12708 with a standard error of .07207. This indicates that at zero tailbeat, the model would predict a distance of -.12708, with a 95% likelihood that the distance somewhere between -.268 and .0142 assuming this is a normal distribution. The intercept was not found to be significant at alpha = 0.05 with a p-value of 0.0806. This indicates that there is not sufficient evidence to suggest that the intercept is not zero.

The slope of the model was found to be 1.92457 with a standard error of 0.17712, indicating that there is a 95% likelihood that the true slope is between 1.577 and 2.272. This indicates that as tailbeat frequency increases, the distance also increases. The t-test of the slope was found to be significant with a p-value of less than 0.0001 indicating that there is evidence to suggest that the slope is not zero.

The overall model was found to be significant at alpha=0.05 with a p-value of less than 0.0001. The root mean square error was found to be 0.13190 which is a mean square error of 0.0174. PRESS was found to be 2.10364.

This is a much better fitting model than that found in part a. However, the transform can likely be adjusted to provide a better fit.

The REG Procedure
Model: InterruptedEyeContactTransform1
Dependent Variable: rel_dist2

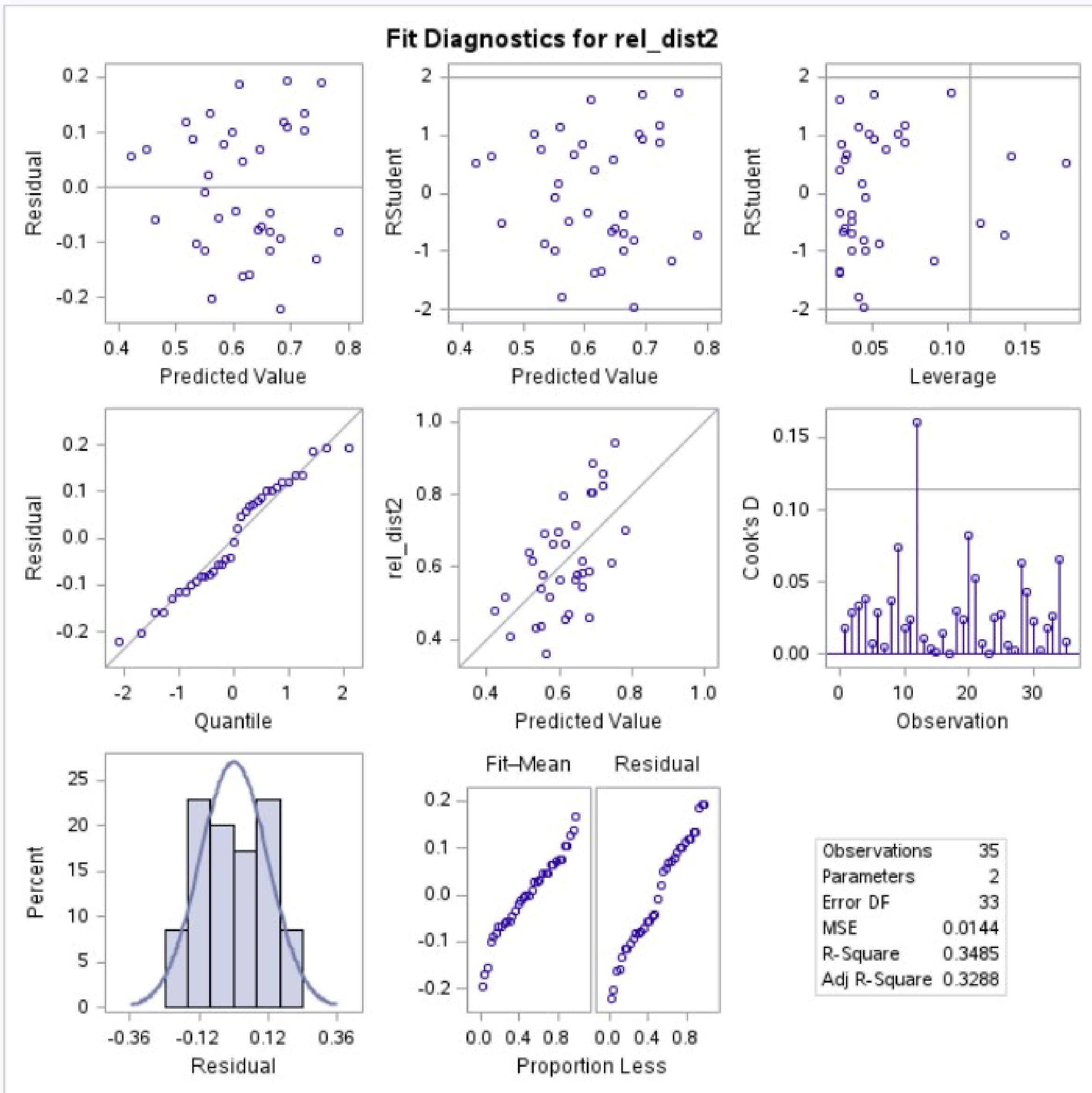
Number of Observations Read	35
Number of Observations Used	35

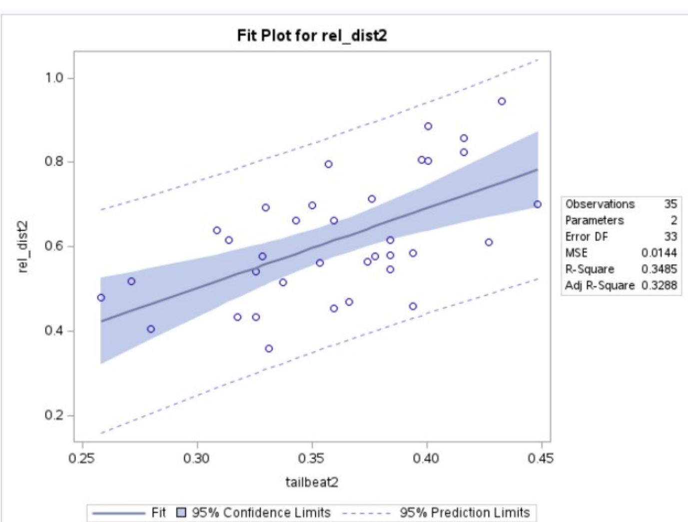
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.25394	0.25394	17.65	0.0002
Error	33	0.47470	0.01438		
Corrected Total	34	0.72864			

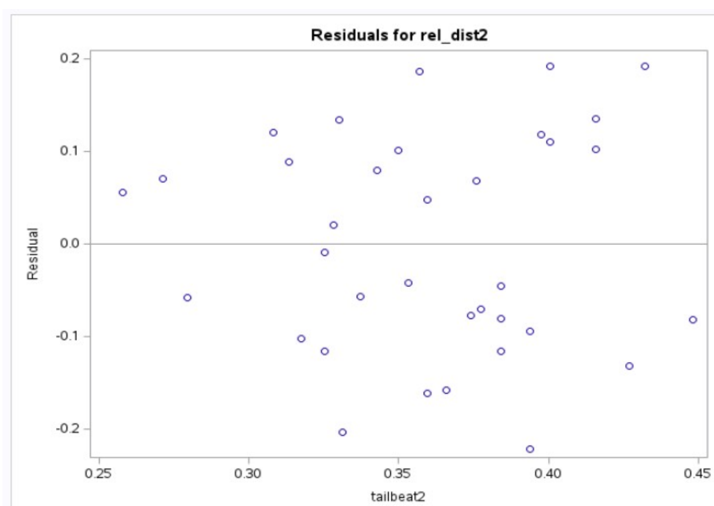
Root MSE	0.11994	R-Square	0.3485
Dependent Mean	0.61703	Adj R-Sq	0.3288
Coeff Var	19.43773		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.06633	0.16390	-0.40	0.6883
tailbeat2	1	1.89563	0.45117	4.20	0.0002

Sum of Residuals	0
Sum of Squared Residuals	0.47470
Predicted Residual SS (PRESS)	0.53108


Fit Diagnostics for rel_dist2
Residual
Predicted Value
RStudent
Leverage
Quantile
rel_dist2
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 35
Parameters 2
Error DF 33
MSE 0.0144
R-Square 0.3485
Adj R-Square 0.3288


Fit Plot for rel_dist2
rel_dist2
tailbeat2
Observations 35
Parameters 2
Error DF 33
MSE 0.0144
R-Square 0.3485
Adj R-Square 0.3288
Fit
95% Confidence Limits
95% Prediction Limits


Residuals for rel_dist2
Residual
tailbeat2

For the interrupted eye contact group when transformed with rel_dist2=sqrt(rel_dist)/length and tailbeat2=sqrt(tailbeat)/length, the coefficient of determination (R^2) was found to be 0.3485, with the adjusted version found to be 0.3288 The intercept estimate for this model is -0.06633 with a standard error of 0.16390. This indicates that at zero tailbeat, the model would predict a distance of -0.06633, with a 95% likelihood that the distance somewhere between -.388 and 0.255 assuming this is a normal distribution.   The t-test on the intercept indicates that there is not sufficient evidence to suggest that the true intercept is not zero. The p-value was 0.6883 which is greater than the 0.05 cutoff for alpha=0.05.

The slope of the model was found to be 1.89563 with a standard error of 0.45117, indicating that there is a 95% likelihood that the true slope is between 1.011 and 2.780. This indicates that tailbeat frequency likely increases with distance. The t-test on the slope indicates that there is sufficient evidence to suggest that the true slope is not zero. The p-value is 0.0002 which is less than the 0.05 cutoff for alpha=0.05.

The overall model was found to be significant at alpha=0.05 with a p-value of 0.0002.   The root mean square error was found to be 0.11994 which is a mean square error of 0.01439.   PRESS was found to be 0.53108.   This model is a much better fit than the model found for the interrupted eye group without the data transform.

```
/* part c data unchanged from part A*/
proc reg; FullEyeContactTransform2: model rel_dist3=tailbeat3 /cli p; where group eq 'A';run;
proc reg; InterruptedEyeContactTransform2: model rel_dist3=tailbeat3 /cli p; where group eq 'B';run;
```

The REG Procedure
Model: FullEyeContactTransform2
Dependent Variable: rel_dist3

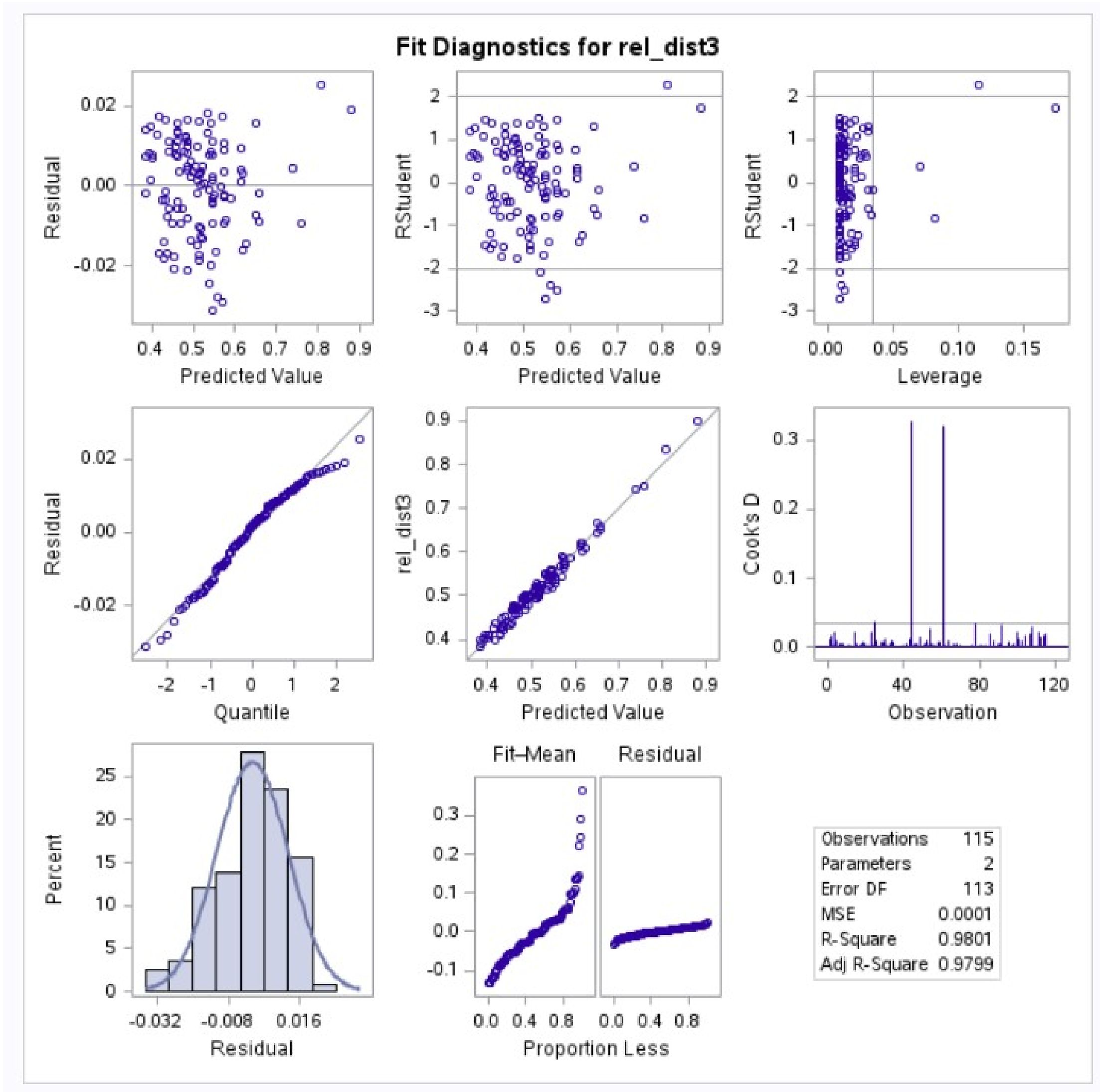
Number of Observations Read	115
Number of Observations Used	115

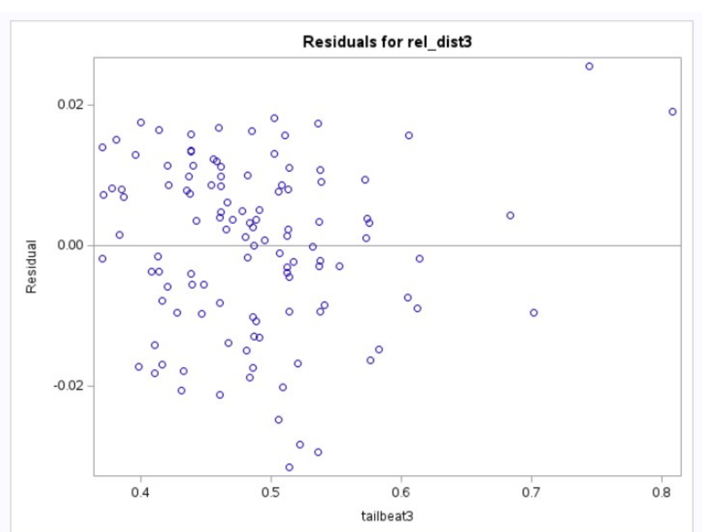
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.80544	0.80544	5570.61	<.0001
Error	113	0.01634	0.00014459		
Corrected Total	114	0.82178			

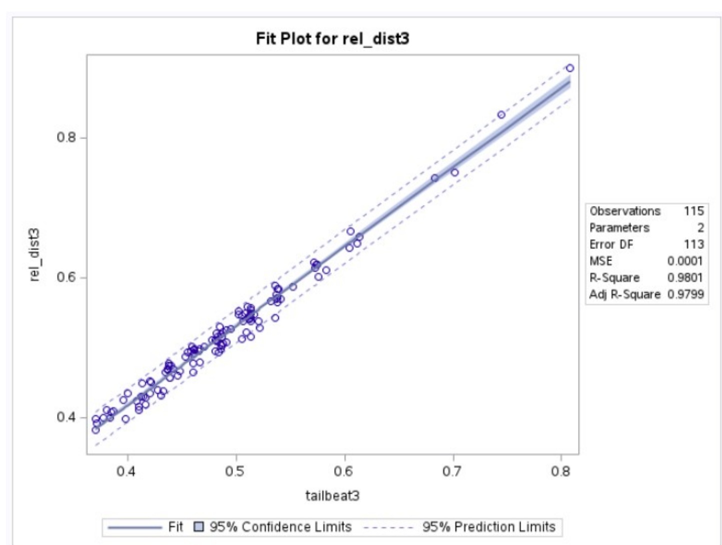
Root MSE	0.01202	R-Square	0.9801
Dependent Mean	0.51613	Adj R-Sq	0.9799
Coeff Var	2.32975		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.03700	0.00750	-4.94	<.0001
tailbeat3	1	1.13595	0.01522	74.64	<.0001

Sum of Residuals	0
Sum of Squared Residuals	0.01634
Predicted Residual SS (PRESS)	0.01713


Fit Diagnostics for rel_dist3
Residual
Predicted Value
RStudent
Leverage
Quantile
rel_dist3
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 115
Parameters 2
Error DF 113
MSE 0.0001
R-Square 0.9801
Adj R-Square 0.9799


Residuals for rel_dist3
Residual
tailbeat3


Fit Plot for rel_dist3
rel_dist3
tailbeat3
Observations 115
Parameters 2
Error DF 113
MSE 0.0001
R-Square 0.9801
Adj R-Square 0.9799
Fit
95% Confidence Limits
95% Prediction Limits

A transformation was developed by first transforming rel_dist increasing the value of the demoninator of the exponent until there was little change in the model, then the same process was repeated to transform tailbeat. This gave a better result than independently transforming either variable.

For the full, uninterrupted eye contact group when transformed with rel_dist3=(rel_dist**.0625)/length; and tailbeat3=(tailbeat**.0625)/length;, the coefficient of determination (R^2) was found to be 0.9801, with the adjusted version found to be 0.9799. The intercept estimate for this model is -0.03700 with a standard error of .00750. This indicates that at zero tailbeat, the model would predict a distance of -0.03700, with a 95% likelihood that the distance somewhere between -0.052 and -0.022 assuming this is a normal distribution. The intercept was found to be significant at alpha = 0.05 with a p-value of less than 0.0001. This indicates that there is sufficient evidence to suggest that the intercept is not zero.

The slope of the model was found to be 1.13595 with a standard error of 0.01522, indicating that there is a 95% likelihood that the true slope is between 1.106 and 1.166. This indicates that as tailbeat frequency increases, the distance also increases. The t-test of the slope was found to be significant with a p-value of less than 0.0001 indicating that there is evidence to suggest that the slope is not zero.

The overall model was found to be significant at alpha=0.05 with a p-value of less than 0.0001. The root mean square error was found to be 0.01202 which is a mean square error of 0.0001. PRESS was found to be .01713.

This is a much better fit model than found in parts a or b. However, it is possible that a transform somewhere between this one ($1 / 2^4$) and the square root version in Part B may be better for prediction. While this is a single regressor, it is possible that this transformation has generated an overfit of this dataset, which would make it a possibly poor predictor for other datasets, even if it models this dataset well.

The REG Procedure
Model: InterruptedEyeContactTransform2
Dependent Variable: rel_dist3

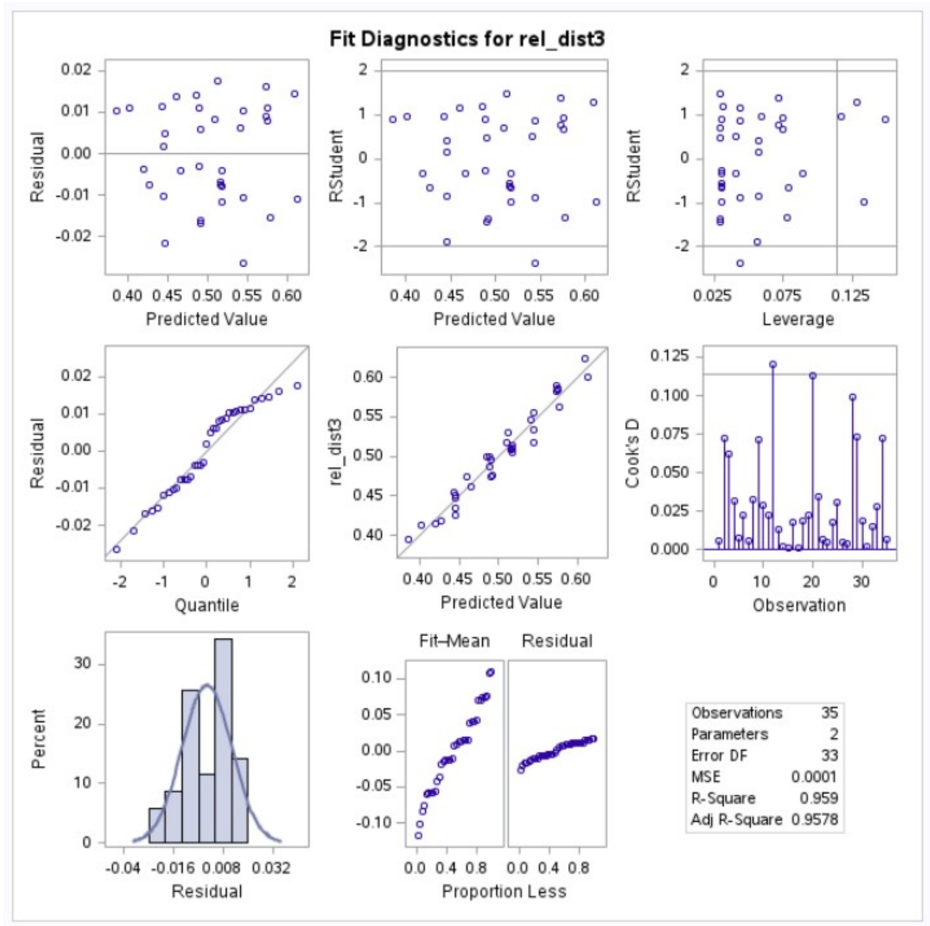
Number of Observations Read	35
Number of Observations Used	35

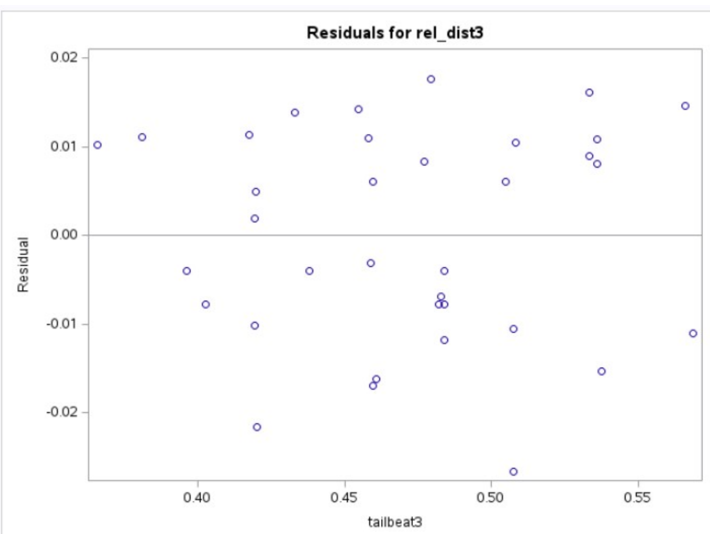
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.11500	0.11500	771.80	<.0001
Error	33	0.00492	0.00014901		
Corrected Total	34	0.11992			

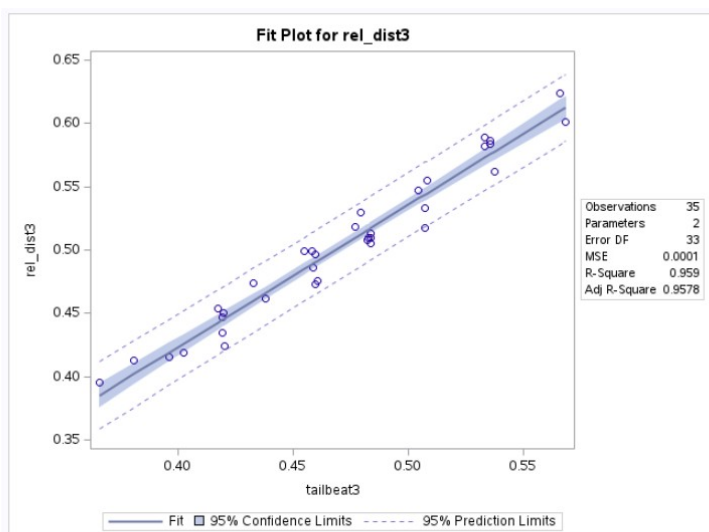
Root MSE	0.01221	R-Square	0.9590
Dependent Mean	0.50266	Adj R-Sq	0.9578
Coeff Var	2.42846		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.02527	0.01911	-1.32	0.1953
tailbeat3	1	1.12149	0.04037	27.78	<.0001

Sum of Residuals	0
Sum of Squared Residuals	0.00492
Predicted Residual SS (PRESS)	0.00554


Fit Diagnostics for rel_dist3
Residual
Predicted Value
RStudent
Leverage
Quantile
rel_dist3
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 35
Parameters 2
Error DF 33
MSE 0.0001
R-Square 0.959
Adj R-Square 0.9578


Residuals for rel_dist3
Residual
tailbeat3


Fit Plot for rel_dist3
rel_dist3
tailbeat3
Observations 35
Parameters 2
Error DF 33
MSE 0.0001
R-Square 0.959
Adj R-Square 0.9578
Fit
95% Confidence Limits
95% Prediction Limits

For the interrupted eye contact group when transformed with rel_dist3= (rel_dist**.0625)/length; and tailbeat3=(tailbeat**.0625)/length;, the coefficient of determination (R^2) was found to be 0.9590 with the adjusted version found to be 0.9578. The intercept estimate for this model is -.02527 with a standard error of .01911. This indicates that at zero tailbeat, the model would predict a distance of -0.0257, with a 95% likelihood that the distance somewhere between -.0627 and 0.012 assuming this is a normal distribution. The intercept was found to be not significant at alpha = 0.05 with a p-value of 0.1953. This indicates that there is not sufficient evidence to suggest that the intercept is not zero.

The slope of the model was found to be 1.12149 with a standard error of 0.04037, indicating that there is a 95% likelihood that the true slope is between 1.042 and 1.201. This indicates that as tailbeat frequency increases, the distance also increases. The t-test of the slope was found to be significant with a p-value of less than 0.0001 indicating that there is evidence to suggest that the slope is not zero.

The overall model was found to be significant at alpha=0.05 with a p-value of less than 0.0001. The root mean square error was found to be 0.01221 which is a mean square error of 0.0001. PRESS was found to be .00554.

This is a much better fit model than found in parts a or b. However, it is possible that a transform somewhere between this one ($1 / 2^4$) and the square root version in Part B may be better for prediction. While this is a single regressor, it is possible that this transformation has generated an overfit of this dataset, which would make it a possibly poor predictor for other datasets, even if it models this dataset well.

```
data;
input obs y x1-x10;

/*
  y=pounds of steam used monthly
  x1=pounds of fatty acid in storage per month
  x2=pounds of crude glycerine made
  x3= average wind velocity (miles/hour)
  x4= calendar days per month
  x5= operating days per month
  x6= days below 32F
  x7= average atmospheric temperature
  x8= (average wind velocity)**2
  x9=number of startups ;
  x10=(average wind velocity)**3

*/

datalines;

1 10.98 5.20 0.61 7.4 31 20 22 35.3 54.8 4 33.428
2 11.13 5.12 0.64 8.0 29 20 25 29.7 64.0 5 40.960
3 12.51 6.19 0.78 7.4 31 23 17 30.8 54.8 4 42.744
4 8.40 3.89 0.49 7.5 30 20 22 58.8 56.3 4 27.587
5 9.27 6.28 0.84 5.5 31 21 0 61.4 30.3 5 25.452
6 8.73 5.76 0.74 8.9 30 22 0 71.3 79.2 4 58.608
7 6.36 3.45 0.42 4.1 31 11 0 74.4 16.8 2 7.056
8 8.50 6.57 0.87 4.1 31 23 0 76.7 16.8 5 14.616
9 7.82 5.69 0.75 4.1 30 21 0 70.7 16.8 4 12.600
10 9.14 6.14 0.76 4.5 31 20 0 57.5 20.3 5 15.428
11 8.24 4.84 0.65 10.3 30 20 11 46.4 106.1 4 68.965
12 12.19 4.88 0.62 6.9 31 21 12 28.9 47.6 4 29.512
13 11.88 6.03 0.79 6.6 31 21 25 28.1 43.6 5 34.444
14 9.57 4.55 0.60 7.3 28 19 18 39.1 53.3 5 31.980
15 10.94 5.71 0.70 8.1 31 23 5 46.8 65.6 4 45.920
16 9.58 5.67 0.74 8.4 30 20 7 48.5 70.6 4 52.244
17 10.09 6.72 0.85 6.1 31 22 0 59.3 37.2 6 31.620
18 8.11 4.95 0.67 4.9 30 22 0 70.0 24.0 4 16.080
19 6.83 4.62 0.45 4.6 31 11 0 70.0 21.2 3 9.540
20 8.88 6.60 0.95 3.7 31 23 0 74.5 13.7 4 13.015
21 7.68 5.01 0.64 4.7 30 20 0 72.1 22.1 4 14.144
22 8.47 5.68 0.75 5.3 31 21 1 58.1 28.1 6 21.075
23 8.86 5.28 0.70 6.2 30 20 14 44.6 38.4 4 26.880
;
```

```
/* part 1 */
proc reg; model y=x1-x10/influence; output out=cooksDData cookd=cookd; run;  /* obs 11 has very
high x3,x8,x10 */
proc print data=cooksDData;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance

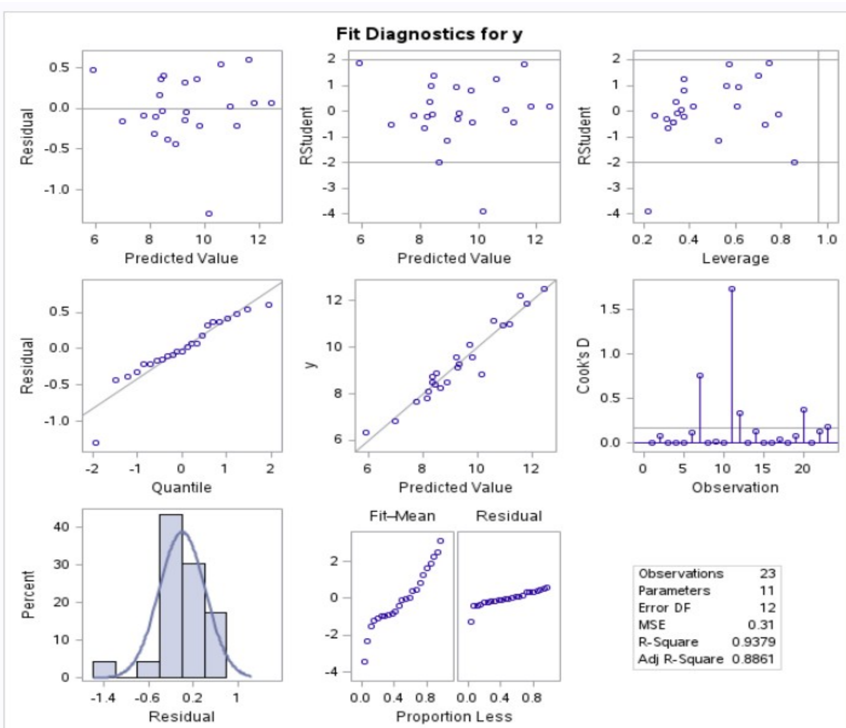
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	56.18481	5.61848	18.12	<.0001
Error	12	3.72046	0.31004		
Corrected Total	22	59.90526			

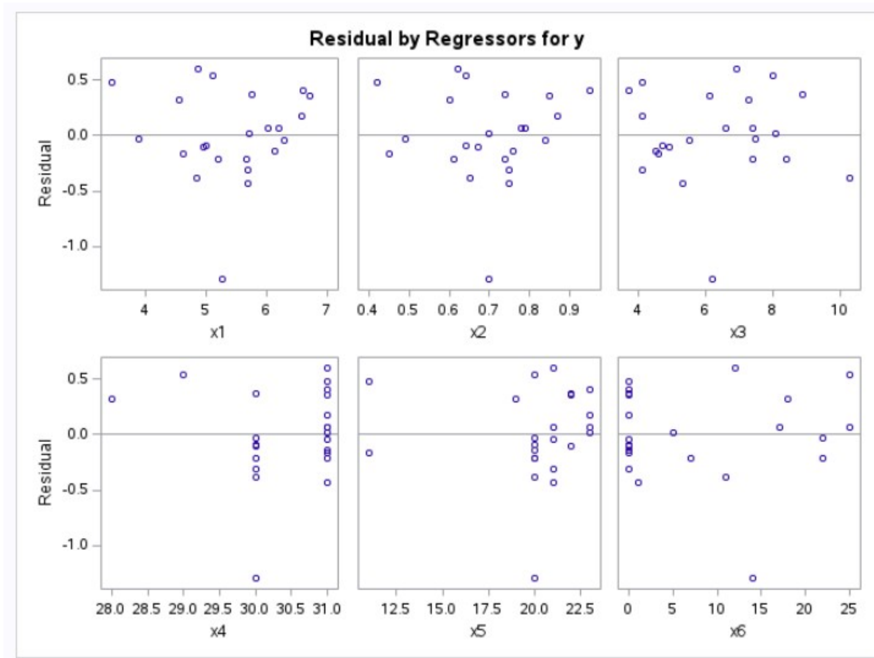
Root MSE	0.55681	R-Square	0.9379
Dependent Mean	9.31130	Adj R-Sq	0.8861
Coeff Var	5.97994		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.85758	7.34348	-0.12	0.9090
x1	1	0.88875	0.58383	1.52	0.1538
x2	1	-2.85220	5.46655	-0.52	0.6113
x3	1	1.06891	0.93474	1.14	0.2751
x4	1	0.19613	0.20487	0.96	0.3573
x5	1	0.18529	0.09514	1.95	0.0752
x6	1	0.00047413	0.03168	0.01	0.9883
x7	1	-0.07462	0.01678	-4.45	0.0008
x8	1	-0.07922	0.05789	-1.37	0.1963
x9	1	-0.35651	0.23593	-1.51	0.1566
x10	1	-0.00409	0.09196	-0.04	0.9652

Sum of Residuals	0
Sum of Squared Residuals	3.72046
Predicted Residual SS (PRESS)	21.23980





Obs	obs	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	cookd
1	1	10.98	5.20	0.61	7.4	31	20	22	35.3	54.8	4	33.428	0.00959
2	2	11.13	5.12	0.64	8.0	29	20	25	29.7	64.0	5	40.960	0.08143
3	3	12.51	6.19	0.78	7.4	31	23	17	30.8	54.8	4	42.744	0.00182
4	4	8.40	3.89	0.49	7.5	30	20	22	58.8	56.3	4	27.587	0.00532
5	5	9.27	6.28	0.84	5.5	31	21	0	61.4	30.3	5	25.452	0.00037
6	6	8.73	5.76	0.74	8.9	30	22	0	71.3	79.2	4	58.608	0.11246
7	7	6.36	3.45	0.42	4.1	31	11	0	74.4	16.8	2	7.056	0.76319
8	8	8.50	6.57	0.87	4.1	31	23	0	76.7	16.8	5	14.616	0.00672
9	9	7.82	5.69	0.75	4.1	30	21	0	70.7	16.8	4	12.600	0.01825
10	10	9.14	6.14	0.76	4.5	31	20	0	57.5	20.3	5	15.428	0.00364
11	11	8.24	4.84	0.65	10.3	30	20	11	46.4	106.1	4	68.965	1.73420
12	12	12.19	4.88	0.62	6.9	31	21	12	28.9	47.6	4	29.512	0.33530
13	13	11.88	6.03	0.79	6.6	31	21	25	28.1	43.6	5	34.444	0.00535
14	14	9.57	4.55	0.60	7.3	28	19	18	39.1	53.3	5	31.980	0.12534
15	15	10.94	5.71	0.70	8.1	31	23	5	46.8	65.6	4	45.920	0.00010
16	16	9.58	5.67	0.74	8.4	30	20	7	48.5	70.6	4	52.244	0.00941
17	17	10.09	6.72	0.85	6.1	31	22	0	59.3	37.2	6	31.620	0.03714
18	18	8.11	4.95	0.67	4.9	30	22	0	70.0	24.0	4	16.080	0.00291
19	19	6.83	4.62	0.45	4.6	31	11	0	70.0	21.2	3	9.540	0.07405
20	20	8.88	6.60	0.95	3.7	31	23	0	74.5	13.7	4	13.015	0.37552
21	21	7.68	5.01	0.64	4.7	30	20	0	72.1	22.1	4	14.144	0.00101
22	22	8.47	5.68	0.75	5.3	31	21	1	58.1	28.1	6	21.075	0.12557
23	23	8.86	5.28	0.70	6.2	30	20	14	44.6	38.4	4	26.880	0.17790

h_{ii} cutoff: > 2p/n = 2*11/23 = 0.95652

|$DFFITS_i$| cutoff: > 2 sqrt(p/n) = 2*sqrt(11/23)=1.38313

|$DFBETAS_{j(i)}$| cutoff: > 2/sqrt(n) = 2/sqrt(23) = 0.41703

COVRAT < 1-3p/n = 1-3*11/23 = -0.434783
COVRAT > 1+3p/n =1=3*11/23= 2.434783
Note: from the textbook on page 219, the lower bound is only appropriate if n > 3p, in this case 23 < 33 so the lower bound will not be used.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	-0.2123	-0.4494	0.3278	3.1716	-0.3139	0.1173	-0.1302	0.0462	-0.0284	-0.0862	-0.0186	-0.0780	-0.0269	-0.0654	0.0658	0.0894
2	0.5421	1.2583	0.3723	0.9451	0.9691	0.5277	0.2952	-0.2799	-0.3106	-0.4814	0.0308	0.2981	-0.0765	0.2035	0.2171	0.1242
3	0.0715	0.1609	0.4150	4.3377	0.1355	0.0190	0.0339	-0.0548	-0.0302	0.0010	0.0562	0.0375	-0.0173	-0.0340	-0.0379	0.0591
4	-0.0326	-0.1210	0.7856	11.9698	-0.2317	0.1204	0.0281	-0.0360	-0.0732	-0.0716	-0.0048	-0.0750	-0.1319	0.0076	-0.0104	0.0647
5	-0.0394	-0.0840	0.3478	3.9648	-0.0613	0.0296	0.0217	-0.0445	-0.0433	-0.0115	0.0409	0.0228	-0.0023	0.0181	0.0074	0.0268
6	0.3657	0.9876	0.5587	2.3178	1.1111	-0.0244	0.1121	-0.1975	0.1499	-0.1224	0.1192	0.1792	0.6024	-0.4527	-0.0889	0.3263
7	0.4730	1.8504	0.7466	0.5215	3.1766	-0.0843	-1.5458	0.3661	-0.3353	0.9523	-0.5247	0.3723	0.1787	-0.7427	0.1036	1.0122
8	0.1722	0.3663	0.3384	3.4448	0.2619	-0.0199	0.0821	-0.0672	-0.0907	0.0228	0.0646	0.1056	0.1038	0.0839	0.0539	0.0135
9	-0.3137	-0.6599	0.3052	2.4455	-0.4374	-0.2951	-0.1517	0.1756	0.2115	0.2393	-0.1964	-0.0220	0.0611	-0.1215	0.0729	-0.0822
10	-0.1432	-0.2950	0.2980	3.4014	-0.1922	-0.0460	-0.0850	0.0610	0.0754	0.0176	-0.0202	0.0567	0.0897	-0.1040	-0.0483	0.0233
11	-0.3792	-2.0066	0.8558	0.5840	-4.8875	-0.3513	0.6994	-0.1961	2.2556	-0.7925	0.1003	0.2747	0.8648	-2.9868	-1.1646	0.0667
12	0.6005	1.8011	0.5744	0.3565	2.0924	-0.3132	-0.4934	0.5116	0.3961	0.4372	0.1850	-1.3600	-1.5309	0.3005	-0.4524	-0.7323
13	0.0680	0.1869	0.6078	6.4128	0.2327	-0.0239	-0.0505	-0.0268	-0.0648	0.0784	0.0026	0.1588	0.0460	-0.0681	0.0759	0.1266
14	0.3246	0.9303	0.6117	2.9151	1.1676	0.6278	-0.0498	0.2479	0.2285	-0.8155	-0.3136	-0.2849	-0.2790	0.0055	-0.0002	-0.2611
15	0.0199	0.0427	0.3608	4.0666	0.0321	-0.0051	0.0097	-0.0070	0.0048	0.0028	0.0123	-0.0141	-0.0083	0.0006	-0.0105	-0.0050
16	-0.2113	-0.4466	0.3262	3.1718	-0.3108	-0.0187	0.0445	-0.1259	-0.1421	0.0796	0.1535	0.0815	0.0091	0.1140	0.1229	0.0265
17	0.3630	0.8131	0.3752	2.1933	0.6301	-0.2015	0.0549	0.0033	0.1747	0.1011	-0.1453	-0.0593	0.1376	-0.1783	0.2699	-0.0019
18	-0.1023	-0.2226	0.3727	3.9514	-0.1716	-0.0470	0.0412	0.0474	0.0298	0.0222	-0.1138	0.0094	-0.0004	0.0469	-0.0091	-0.0655
19	-0.1588	-0.5322	0.7300	7.2927	-0.8751	-0.0453	-0.4928	0.1720	-0.0702	0.1182	0.2981	0.0682	-0.0317	-0.1436	0.1373	0.2024
20	0.4085	1.3861	0.6984	1.4695	2.1090	-0.3067	-0.2752	1.2296	0.3442	-0.0159	-0.7101	-0.1402	-0.0382	0.7002	-0.9843	-1.0077
21	-0.0885	-0.1759	0.2487	3.3610	-0.1012	-0.0458	-0.0120	0.0538	0.0324	0.0315	-0.0638	-0.0072	-0.0106	0.0146	-0.0117	-0.0418
22	-0.4315	-1.1357	0.5231	1.6128	-1.1894	0.3186	0.5471	-0.0157	0.1602	-0.5607	-0.0659	0.1081	0.0462	0.1056	-0.9713	-0.2655
23	-1.2961	-3.8875	0.2198	0.0002	-2.0636	-0.1540	0.5701	-1.0410	-1.0996	0.5796	0.7632	0.1683	0.2173	0.8979	1.0778	0.3743

There are no hat diagonals (h_{ii}) greater than the suggested cutoff of greater than 2p/n (0.95652). The highest is for observation 11 at 0.8558. |DFFITS| has a calculated cutoff of greater than 1.38313, notable observations for DFFITS are 7, 11, 12, 20, 23 all at above 1.38313. COVRAT has a calculated cutoff of greater than 2.434783, notable observations for COVRAT are 1, 3, 4, 5, 8, 9, 10, 13, 14, 15, 16, 18, 19, 21. However, due to the small sample size COVRAT is less useful than it otherwise would be. R-student indicates a questionable observation at 11 with -2.0066, but the cutoff being 2 this is extremely close and not concerning. However, observation 23 is 3.8875 which would be an extreme outlier. Cook's D indicates only observation 11 is large compared to the other values at 1.734. Observation 7 is slightly high at 0.763.

```
proc rsquare adjrsq mse cp;
model y=x1-x10;run;
```

The RSQUARE Procedure
Model: MODEL1
Dependent Variable: y

R-Square Selection Method

Number of Observations Read	23
Number of Observations Used	23

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
10	0.9379	0.8861	11.0000	0.31004	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10

The RSQUARE Procedure
Model: MODEL1
Dependent Variable: y

R-Square Selection Method

Number of Observations Read	22
Number of Observations Used	22

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
10	0.9536	0.9114	11.0000	0.24759	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10

Of the observations, 11 is the most likely candidate for removal. With observation 11,MSE is 0.310, CP is 11.00, R^2 is 0.938 and PRESS is 21.240. Without observation 11, MSE is 0.24759, CP is 11.00, R^2 is 0.9536, and PRESS is 11.468.

On looking at observation 11, it appears to be the wind speed making the observation appear to have leverage and influence. However, since X3 is average wind velocity, X8 is $X3^2$, and X10 is $X3^3$, intuitively these would hold more leverage since there are 3 variables reporting the same metric in different formats. Its almost certain that these 3 will be found to be multicolinear and therefore will not be in the final model together, so without a non-statistical reason to do so, observation 11 will not be removed from the dataset. Furthermore, no other observations need removal from this dataset.

```
/* part 2 */
proc reg; model y=x1-x10 / collin vif;run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

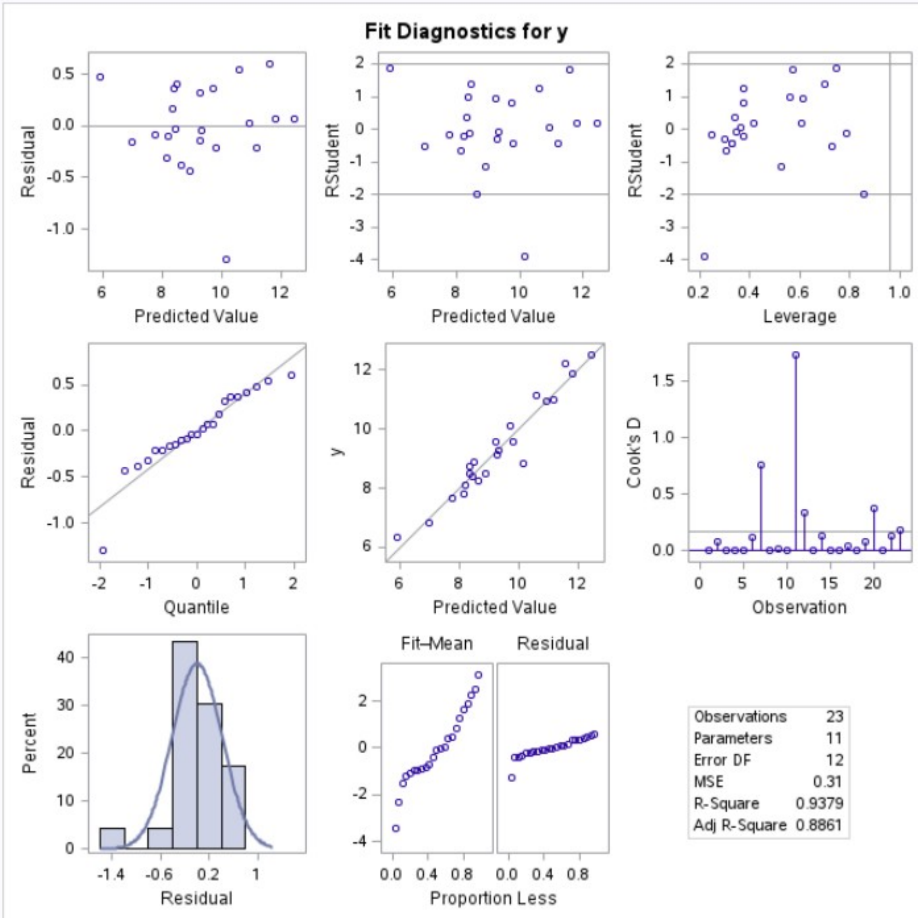
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	56.18481	5.61848	18.12	<.0001
Error	12	3.72046	0.31004		
Corrected Total	22	59.90526			

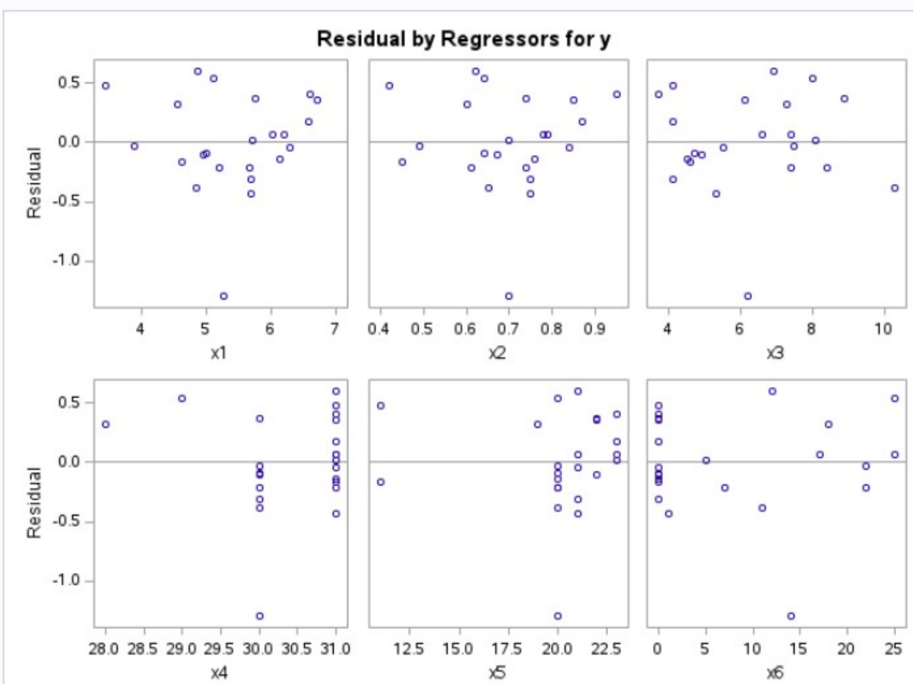
Root MSE	0.55681	R-Square	0.9379
Dependent Mean	9.31130	Adj R-Sq	0.8861
Coeff Var	5.97994		

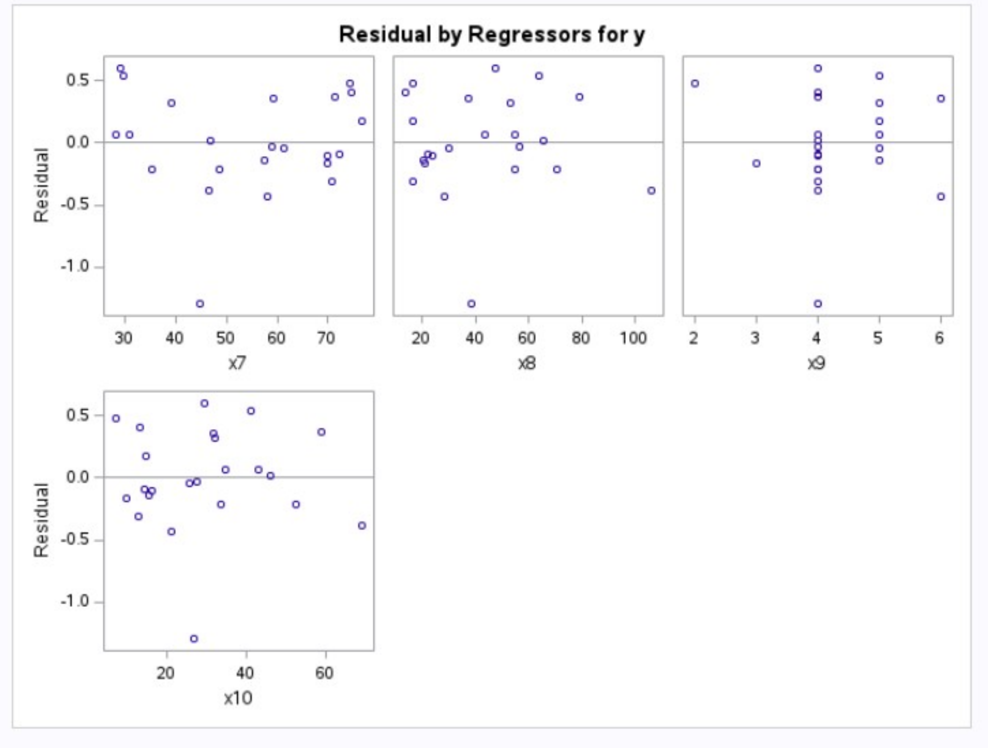
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.85758	7.34348	-0.12	0.9090	0
x1	1	0.88875	0.58383	1.52	0.1538	17.39576
x2	1	-2.85220	5.46655	-0.52	0.6113	36.58293
x3	1	1.06891	0.93474	1.14	0.2751	203.54313
x4	1	0.19613	0.20487	0.96	0.3573	1.84823
x5	1	0.18529	0.09514	1.95	0.0752	6.28608
x6	1	0.00047413	0.03168	0.01	0.9883	6.38823
x7	1	-0.07462	0.01678	-4.45	0.0008	5.56417
x8	1	-0.07922	0.05789	-1.37	0.1963	137.62823
x9	1	-0.35651	0.23593	-1.51	0.1566	3.02878
x10	1	-0.00409	0.09196	-0.04	0.9652	162.05539

Collinearity Diagnostics													
			Proportion of Variation										
Number	Eigenvalue	Condition Index	Intercept	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	9.87094	1.00000	0.00000251	0.00001303	0.00000888	0.00000362	0.00000347	0.00003601	0.00046060	0.00012930	0.00001496	0.00012424	0.00001271
2	0.76548	3.59099	0.00000480	0.00004887	0.00003691	0.00000789	0.00000791	0.00005099	0.06593	0.00305	0.00020799	0.00020600	0.00012325
3	0.26348	6.12080	0.00001454	0.00005375	0.00003547	0.00004822	0.00002110	0.00011078	0.11697	0.00013122	0.00157	0.00128	0.00159
4	0.06504	12.31944	0.00020282	0.00165	0.00200	0.00015877	0.00023617	0.00240	0.02291	0.08492	0.00082565	0.02753	0.00090748
5	0.01633	24.58326	0.00001216	0.00523	0.00778	0.00054303	0.00001719	0.00857	0.04639	0.00119	0.00171	0.51647	0.00275
6	0.01017	31.14699	0.00429	0.00513	0.00162	0.00191	0.00901	0.02806	0.23059	0.44279	0.00129	0.02995	0.00102
7	0.00624	39.77217	0.00014758	0.02176	0.00446	0.00175	0.00006961	0.37457	0.18896	0.14027	0.00340	0.02945	0.01782
8	0.00097906	100.40965	0.01146	0.30010	0.33726	0.03308	0.00370	0.12630	0.08019	0.21064	0.27684	0.00374	0.07969
9	0.00091939	103.61682	0.01406	0.46727	0.01205	0.00589	0.00149	0.00206	0.02629	0.01657	0.61399	0.03626	0.39081
10	0.00028320	186.69571	0.01485	0.00444	0.31011	0.60619	0.35944	0.31315	0.22004	0.00947	0.04452	0.35500	0.37816
11	0.00013749	267.94387	0.95494	0.19431	0.32464	0.35042	0.62601	0.14469	0.00127	0.09082	0.05562	0.00000228	0.12712

The REG Procedure
Model: MODEL1
Dependent Variable: y







The VIF output indicates possible multicolinearity between X3, X8, and X10, and X2. Of these, X2 is 35.58 which is much closer to the cutoff of 20 than X3 at 203.54, X8 at 137.628, and X10 at 162.055. Its possible that dropping one of the other 3 will bring X2 into normal range.

The condition index indicates possible multicolinearity on rows 8,9,10,and 11.

For row 8, this would be between X1 =.30 and X2=.337. However, this is a low proportion of variation and does not likely indicate colinearity.

For row 9, X1=.467 and X8=.614 may be colinear. X8 is a high proportion of variation. X10 is 0.391. While X1 and X10 are below 0.5, at least one is likely colinear with X8.

For row 10, X3=.60619, X4=0.35944, X5=0.31315, X9=0.35500, X10=0.37816. This would indicate that X4, X5, X9, and/or X10 is likely colinear with X3. X5 has a lower proportion of variation so is less likely than the other 3 candidates.

For row 11, intercept=0.95494, x4=0.62601 are colinear.

Combining the results of the VIF and colinearity diagnostics output, X3, X8, X10, and possibly X2 should not be in a model together. X4 appears to be colinear with the intercept and could create a model with a better fit, but worse prediction power. Ideally, leaving X4 out will likely lead to a better model.

proc reg; model y=x1 x2 x3 x4 x5 x6 x7 x9 / collin vif;run;

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	55.44997	6.93125	21.78	<.0001
Error	14	4.45529	0.31823		
Corrected Total	22	59.90526			

Root MSE	0.56412	R-Square	0.9256
Dependent Mean	9.31130	Adj R-Sq	0.8831
Coeff Var	6.05847		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.03214	6.61182	0.46	0.6536	0
x1	1	1.14776	0.56007	2.05	0.0597	15.59627
x2	1	-5.89956	3.88026	-1.52	0.1507	17.95729
x3	1	-0.09851	0.09754	-1.01	0.3297	2.15933
x4	1	0.18899	0.20667	0.91	0.3760	1.83243
x5	1	0.22878	0.07659	2.99	0.0098	3.96879
x6	1	0.00382	0.02671	0.14	0.8883	4.42531
x7	1	-0.08268	0.01565	-5.28	0.0001	4.71523
x9	1	-0.25950	0.20300	-1.28	0.2219	2.18457

Collinearity Diagnostics											
			Proportion of Variation								
Number	Eigenvalue	Condition Index	Intercept	x1	x2	x3	x4	x5	x6	x7	x9
1	8.19987	1.00000	0.00000467	0.00002141	0.00002667	0.00047832	0.00000515	0.00008344	0.00090068	0.00023022	0.00025334
2	0.66027	3.52405	0.00000141	0.00002552	0.00003775	0.00119	0.00000225	0.00001558	0.16328	0.00338	0.00003886
3	0.06113	11.58228	0.00024376	0.00217	0.00479	0.06713	0.00021868	0.00410	0.01209	0.08930	0.05185
4	0.04851	13.00114	0.00005953	0.00000771	0.00002395	0.48644	0.00008055	0.00065073	0.28375	0.06345	0.00029490
5	0.01542	23.05872	0.00001046	0.00717	0.01826	0.00773	0.00003573	0.02405	0.01273	0.00788	0.78148
6	0.00907	30.07395	0.00628	0.01141	0.00210	0.03500	0.01090	0.15353	0.14323	0.36349	0.00538
7	0.00465	41.98534	0.00498	0.04431	0.04395	0.26076	0.00572	0.58313	0.26267	0.27811	0.00262
8	0.00091888	94.46544	0.02141	0.74762	0.84574	0.00603	0.00032575	0.23138	0.05461	0.10051	0.00809
9	0.00016500	222.92449	0.96701	0.18726	0.08507	0.13524	0.98270	0.00307	0.06674	0.09366	0.15000

Removing X8 and X10 then rerunning the colinearity diagnostics brought the VIF below 20 for all regressors. Making this adjustment also made the colinearity between the intercept and X4 much clearer, the proportion of variation is now 0.96701 for the intercept and 0.98270 for X4 on row 9. This would indicate that X4 should not be used in the selected models, outside of the full model.

```
/* part 3 */
proc rsquare adjrsq mse cp;
model y=x1-x10;run;
```

The RSQUARE Procedure
Model: MODEL1
Dependent Variable: y

R-Square Selection Method

Number of Observations Read	23
Number of Observations Used	23

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
1	0.7044	0.6903	38.1127	0.84319	x7
1	0.3765	0.3468	101.4665	1.77853	x6
1	0.2834	0.2492	119.4679	2.04430	x5
1	0.2090	0.1713	133.8388	2.25647	x3
1	0.2084	0.1707	133.9454	2.25804	x10
1	0.1448	0.1041	146.2341	2.43947	x8
1	0.1415	0.1006	146.8799	2.44900	x1
1	0.1397	0.0987	147.2307	2.45418	x9
1	0.1020	0.0593	154.5072	2.56161	x2
1	0.0086	-.0386	172.5513	2.82801	x4
2	0.8642	0.8506	9.2386	0.40675	x1 x7
2	0.8468	0.8315	12.5990	0.45884	x5 x7
2	0.8447	0.8292	13.0006	0.46507	x2 x7
2	0.7615	0.7376	29.0856	0.71441	x4 x7
2	0.7399	0.7139	33.2482	0.77894	x7 x9
2	0.7281	0.7009	35.5459	0.81456	x6 x7
2	0.7127	0.6840	38.5059	0.86045	x7 x8
2	0.7100	0.6810	39.0380	0.86869	x3 x7
2	0.7047	0.6752	40.0553	0.88447	x7 x10
2	0.6846	0.6530	43.9469	0.94479	x1 x6
2	0.6130	0.5743	57.7778	1.15920	x2 x6
2	0.6092	0.5701	58.5055	1.17048	x5 x6

Candidate Models:

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
10	0.9379	0.8861	11.00	0.3100	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10
7	0.9238	0.8883	7.72	0.3042	x1 x2 x5 x6 x7 x8 x9
6	0.9222	0.8931	6.03	0.2912	x1 x2 x5 x7 x9 x10
5	0.9115	0.8854	6.11	0.3120	x1 x2 x5 x7 x9
4	0.8976	0.8749	6.78	0.3407	x1 x2 x5 x7

R-Square(Coefficient of Determination) – Proportion of variation explained by the model. 0.9379 would indicate that 93.79% of variation is explained by the model. This is helpful in determining how well the model can predict observations.

Adjusted R-Square – Similar definition to R-square but in this case, the value is adjusted for the number of regressors in the model (P). This is useful in comparing models with different numbers of regressors, since R-square increases with a larger model. This can lead to a larger model being selected when a smaller model is actually a better fit and so for comparisons, adjusted r-square is the better metric to use.

C(p)(Mallow's Cp) – Measure of bias of a model. If Cp=p (Mallow's Cp = number of predictors) then there model is unbiased. Selecting a model with a very large or very small C(p) would mean selecting a heavily biased model, and so staying near Cp=p is ideal when selecting candidate models.

MSE – Mean Square Error - This is the expected squared error between the model and the true model, which is the sum of the model's variance and the model's bias squared. Generally, the lower the MSE the better the fit of the model. This value is also related to C(p) in that C(p) is a measure of bias, and MSE increases with bias as a squared term.

PRESS – Predicted error sum of squares – this is the sum of squared press residuals, which are prediction errors weighted by 1 - the hat matrix diagonals. If h_{ii} is large, then PRESS residuals will be large, leading to a larger PRESS statistic. Larger PRESS statistics indicate that the model does not fit the data and should not be used to make predictions. This is a useful model selection tool, and generally due to the 1-h_{ii} diagonal, if the PRESS statistic goes high, it goes very high and is easier to identify poor fits than it would be using a linear term.

R^2_{pred} – Predicted Rsquare, indicates predictive power of the model. This is very similar to normal R^2 but with the sum of squares error term replaced with predicted error sum of squares, leading to a version of R^2 useful in determining if a model will be good for prediction, rather than just model fit.

proc reg; model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / cli p;run;

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	56.18481	5.61848	18.12	<.0001
Error	12	3.72046	0.31004		
Corrected Total	22	59.90526			

Root MSE	0.55681	R-Square	0.9379
Dependent Mean	9.31130	Adj R-Sq	0.8861
Coeff Var	5.97994		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.85758	7.34348	-0.12	0.9090
x1	1	0.88875	0.58383	1.52	0.1538
x2	1	-2.85220	5.46655	-0.52	0.6113
x3	1	1.06891	0.93474	1.14	0.2751
x4	1	0.19613	0.20487	0.96	0.3573
x5	1	0.18529	0.09514	1.95	0.0752
x6	1	0.00047413	0.03168	0.01	0.9883
x7	1	-0.07462	0.01678	-4.45	0.0008
x8	1	-0.07922	0.05789	-1.37	0.1963
x9	1	-0.35651	0.23593	-1.51	0.1566
x10	1	-0.00409	0.09196	-0.04	0.9652

Sum of Residuals	0
Sum of Squared Residuals	3.72046
Predicted Residual SS (PRESS)	21.23980

proc reg; model y=x1 x2 x5 x6 x7 x8 x9 / cli p;run;

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	55.34265	7.90609	25.99	<.0001
Error	15	4.56261	0.30417		
Corrected Total	22	59.90526			

Root MSE	0.55152	R-Square	0.9238
Dependent Mean	9.31130	Adj R-Sq	0.8883
Coeff Var	5.92312		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.51278	1.75164	4.86	0.0002
x1	1	1.34901	0.48898	2.76	0.0146
x2	1	-6.67357	3.59948	-1.85	0.0835
x5	1	0.23185	0.07351	3.15	0.0066
x6	1	-0.00099063	0.02569	-0.04	0.9698
x7	1	-0.08540	0.01477	-5.78	<.0001
x8	1	-0.00990	0.00634	-1.56	0.1393
x9	1	-0.34200	0.18558	-1.84	0.0852

Sum of Residuals	0
Sum of Squared Residuals	4.56261
Predicted Residual SS (PRESS)	13.04195

proc reg; model y=x1 x2 x5 x7 x9 x10 / cli p;run;

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	55.24599	9.20766	31.62	<.0001
Error	16	4.65928	0.29120		
Corrected Total	22	59.90526			

Root MSE	0.53963	R-Square	0.9222
Dependent Mean	9.31130	Adj R-Sq	0.8931
Coeff Var	5.79547		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.16307	1.23774	6.60	<.0001
x1	1	1.39218	0.46023	3.02	0.0080
x2	1	-6.35161	3.49541	-1.82	0.0880
x5	1	0.22333	0.06994	3.19	0.0057
x7	1	-0.08457	0.00886	-9.54	<.0001
x9	1	-0.34854	0.18223	-1.91	0.0739
x10	1	-0.01312	0.00882	-1.49	0.1562

Sum of Residuals	0
Sum of Squared Residuals	4.65928
Predicted Residual SS (PRESS)	11.81643

proc reg; model y=x1 x2 x5 x7 x9 / cli p;run;

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	54.60123	10.92025	35.00	<.0001
Error	17	5.30403	0.31200		
Corrected Total	22	59.90526			

Root MSE	0.55857	R-Square	0.9115
Dependent Mean	9.31130	Adj R-Sq	0.8854
Coeff Var	5.99885		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.79390	1.25517	6.21	<.0001
x1	1	1.35930	0.47584	2.86	0.0109
x2	1	-5.95057	3.60730	-1.65	0.1174
x5	1	0.19156	0.06893	2.78	0.0129
x7	1	-0.07854	0.00816	-9.63	<.0001
x9	1	-0.30292	0.18593	-1.63	0.1217

Sum of Residuals	0
Sum of Squared Residuals	5.30403
Predicted Residual SS (PRESS)	10.91862

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	23
Number of Observations Used	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	53.77309	13.44327	39.46	<.0001
Error	18	6.13217	0.34068		
Corrected Total	22	59.90526			

Root MSE	0.58367	R-Square	0.8976
Dependent Mean	9.31130	Adj R-Sq	0.8749
Coeff Var	6.26845		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.48915	1.29694	5.77	<.0001
x1	1	1.23527	0.49082	2.52	0.0215
x2	1	-6.06859	3.76866	-1.61	0.1247
x5	1	0.17112	0.07083	2.42	0.0265
x7	1	-0.07544	0.00829	-9.10	<.0001

Sum of Residuals	0
Sum of Squared Residuals	6.13217
Predicted Residual SS (PRESS)	12.29963

$R^2_{pred} = 1 - PRESS/SS_T$

Model	R-Square	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10	0.9379	0.8861	11.00	0.6454	0.3100	21.2398
x1 x2 x5 x6 x7 x8 x9	0.9238	0.8883	7.72	0.7823	0.3042	13.0420
x1 x2 x5 x7 x9 x10	0.9222	0.8931	6.03	0.8027	0.2912	11.8164
x1 x2 x5 x7 x9	0.9115	0.8854	6.11	0.8177	0.3120	10.9186
x1 x2 x5 x7	0.8976	0.8749	6.78	0.7947	0.3407	12.2996

Given the regression diagnostics, the best two models are y=x1 x2 x5 x7 x9 x10 and y=x1 x2 x5 x7 x9. These are very close, however y=x1 x2 x5 x7 x9 is the chosen model. Both R^2_{adj} and R^2_{pred} are close enough between the models that they're not likely truly different. Mallow's CP is slightly better on the larger model, and PRESS is slightly better on the smaller model. MSE is slightly better on the larger model. These are too close to choose using the diagnostics variables, but in the case that two models are very similar and one has fewer predictors, the smaller model is the best choice. In this case, that will be y=x1 x2 x5 x7 x9.

```
proc reg; model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=stepwise;run;
```

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	x7		1	0.7044	0.7044	38.1127	50.05	<.0001
2	x1		2	0.1598	0.8642	9.2386	23.53	<.0001
3	x5		3	0.0187	0.8829	7.6280	3.03	0.0978
4	x4		4	0.0225	0.9054	5.2756	4.29	0.0531
5		x1	3	0.0083	0.8972	4.8703	1.57	0.2262

Stepwise provided the model y=x1 x4 x5 x7

Method	Model	R-Square	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
Stepwise	x1 x4 x5 x7	0.9054	0.8844	5.27	0.8636	0.3148	8.1699
Selected	x1 x2 x5 x7 x9	0.9115	0.8854	6.11	0.8177	0.3120	10.9186

The generated stepwise model and the selected model are very similar. The differences are that the Mallow's Cp statistic indicates a slight bias on the stepwise model. PRESS is also lower on the stepwise model, which increases R^2_{pred} however, it is also known that X4 is colinear with the intercept which will also increase R^2 generally, and because R^2_{adj} is nearly identical between these models, the selected model is still the better choice and does not have a colinearity issue.

ytran=log(y)$^{1/3}$
x1tran=x1$^{1/3}$;
x2tran=x2$^{1/3}$;
x4tran=x4$^{1/3}$;
x5tran=x5$^{1/3}$;
x7tran=x7$^{1/3}$;
x9tran=x9$^{1/3}$;

Using these transforms gives the following version of the model:

Method	Model	R-Square	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
Stepwise	x1 x4 x5 x7	0.9313	0.9160	5.99	0.8997	0.0001	0.0026
Selected	x1 x2 x5 x7 x9	0.9387	0.9207	5.94	0.8542	0.0001	0.0038

For comparison, the transformed version was also run using the model generated by the stepwise selection method. The results are similar, but the c(p) statistic and R^2_{pred} are much closer after the transform. For the same reasons stated above, the selected model is the better choice in this case.