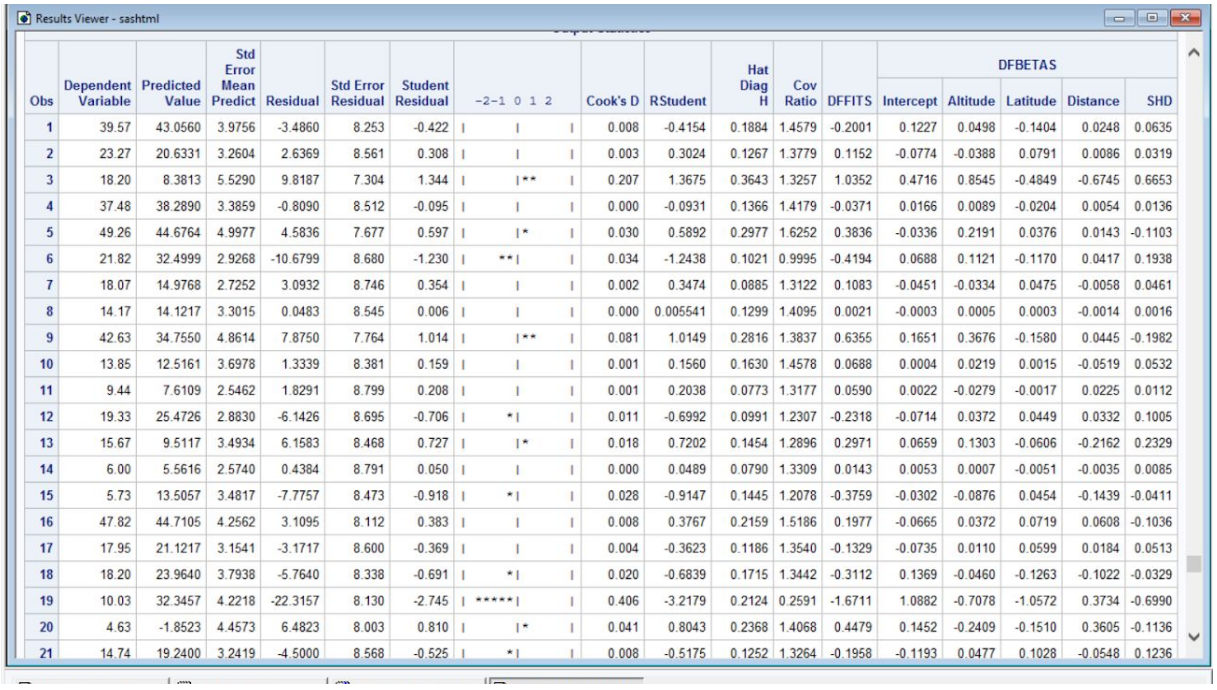
Case Study 3

Timothy Kim, Karl Keat, Leo Wang, Graham Gilliam

Analysis

The influential points in the data were analyzed using SAS. An initial model, Model 2, was used which had the following prediction equation:
$E(y) = -97.89 + .00221x_1 + 3.45x_2 - .0536x_3 - 15.86x_4$. From this model, a residual analysis was conducted to identify potential outliers and influential points. Part of the output is shown below in Table 1.

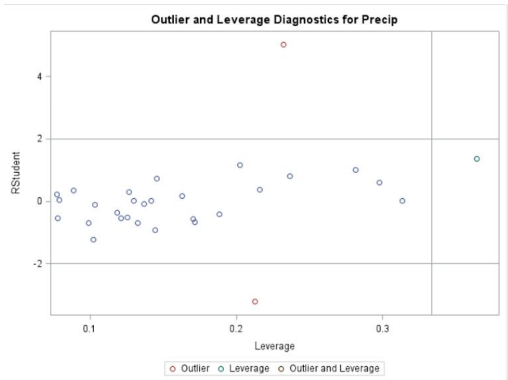
Table 1. Residual Analysis of initial Model

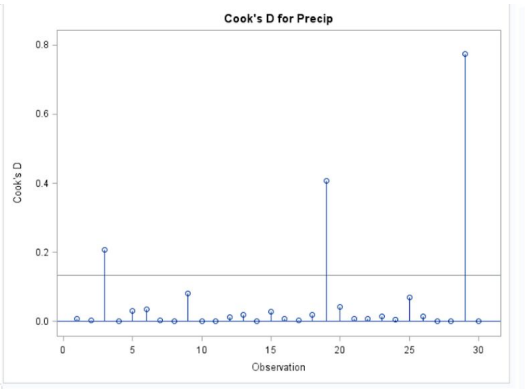

Results Viewer - sashtml

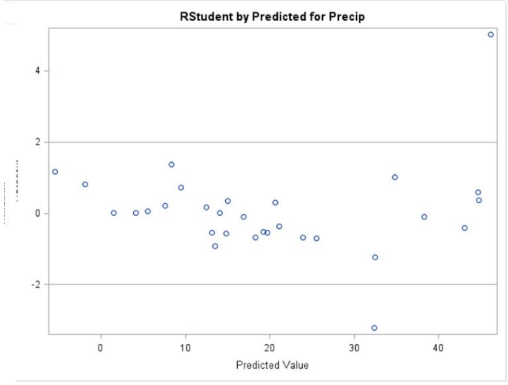
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS				
													Intercept	Altitude	Latitude	Distance	SHD
1	39.57	43.0560	3.9756	-3.4860	8.253	-0.422	| |	0.008	-0.4154	0.1884	1.4579	-0.2001	0.1227	0.0498	-0.1404	0.0248	0.0635
2	23.27	20.6331	3.2604	2.6369	8.561	0.308	| |	0.003	0.3024	0.1267	1.3779	0.1152	-0.0774	-0.0388	0.0791	0.0086	0.0319
3	18.20	8.3813	5.5290	9.8187	7.304	1.344	| |**	0.207	1.3675	0.3643	1.3257	1.0352	0.4716	0.8545	-0.4849	-0.6745	0.6653
4	37.48	38.2890	3.3859	-0.8090	8.512	-0.095	| |	0.000	-0.0931	0.1366	1.4179	-0.0371	0.0166	0.0089	-0.0204	0.0054	0.0136
5	49.26	44.6764	4.9977	4.5836	7.677	0.597	| |*	0.030	0.5892	0.2977	1.6252	0.3836	-0.0336	0.2191	0.0376	0.0143	-0.1103
6	21.82	32.4999	2.9268	-10.6799	8.680	-1.230	| **|	0.034	-1.2438	0.1021	0.9995	-0.4194	0.0688	0.1121	-0.1170	0.0417	0.1938
7	18.07	14.9768	2.7252	3.0932	8.746	0.354	| |	0.002	0.3474	0.0885	1.3122	0.1083	-0.0451	-0.0334	0.0475	-0.0058	0.0461
8	14.17	14.1217	3.3015	0.0483	8.545	0.006	| |	0.000	0.005541	0.1299	1.4095	0.0021	-0.0003	0.0005	0.0003	-0.0014	0.0016
9	42.63	34.7550	4.8614	7.8750	7.764	1.014	| |**	0.081	1.0149	0.2816	1.3837	0.6355	0.1651	0.3676	-0.1580	0.0445	-0.1982
10	13.85	12.5161	3.6978	1.3339	8.381	0.159	| |	0.001	0.1560	0.1630	1.4578	0.0688	0.0004	0.0219	0.0015	-0.0519	0.0532
11	9.44	7.6109	2.5462	1.8291	8.799	0.208	| |	0.001	0.2038	0.0773	1.3177	0.0590	0.0022	-0.0279	-0.0017	0.0225	0.0112
12	19.33	25.4726	2.8830	-6.1426	8.695	-0.706	| *|	0.011	-0.6992	0.0991	1.2307	-0.2318	-0.0714	0.0372	0.0449	0.0332	0.1005
13	15.67	9.5117	3.4934	6.1583	8.468	0.727	| |*	0.018	0.7202	0.1454	1.2896	0.2971	0.0659	0.1303	-0.0606	-0.2162	0.2329
14	6.00	5.5616	2.5740	0.4384	8.791	0.050	| |	0.000	0.0489	0.0790	1.3309	0.0143	0.0053	0.0007	-0.0051	-0.0035	0.0085
15	5.73	13.5057	3.4817	-7.7757	8.473	-0.918	| *|	0.028	-0.9147	0.1445	1.2078	-0.3759	-0.0302	-0.0876	0.0454	-0.1439	-0.0411
16	47.82	44.7105	4.2562	3.1095	8.112	0.383	| |	0.008	0.3767	0.2159	1.5186	0.1977	-0.0665	0.0372	0.0719	0.0608	-0.1036
17	17.95	21.1217	3.1541	-3.1717	8.600	-0.369	| |	0.004	-0.3623	0.1186	1.3540	-0.1329	-0.0735	0.0110	0.0599	0.0184	0.0513
18	18.20	23.9640	3.7938	-5.7640	8.338	-0.691	| *|	0.020	-0.6839	0.1715	1.3442	-0.3112	0.1369	-0.0460	-0.1263	-0.1022	-0.0329
19	10.03	32.3457	4.2218	-22.3157	8.130	-2.745	| *****|	0.406	-3.2179	0.2124	0.2591	-1.6711	1.0882	-0.7078	-1.0572	0.3734	-0.6990
20	4.63	-1.8523	4.4573	6.4823	8.003	0.810	| |*	0.041	0.8043	0.2368	1.4068	0.4479	0.1452	-0.2409	-0.1510	0.3605	-0.1136
21	14.74	19.2400	3.2419	-4.5000	8.568	-0.525	| *|	0.008	-0.5175	0.1252	1.3264	-0.1958	-0.1193	0.0477	0.1028	-0.0548	0.1236

Just from looking at Table 1, there appears to be at least one influential point, the 19th observation, which has a large RStudent value, -3.2179. To better visualize any influential points, a series of plots were constructed: a plot of Cook's D, a plot of RStudent by Predicted Value, and a plot of RStudent by Leverage. The three graphs are plotted in Table 2.

Table 2. Graphic Representation of Initial Residual Analysis





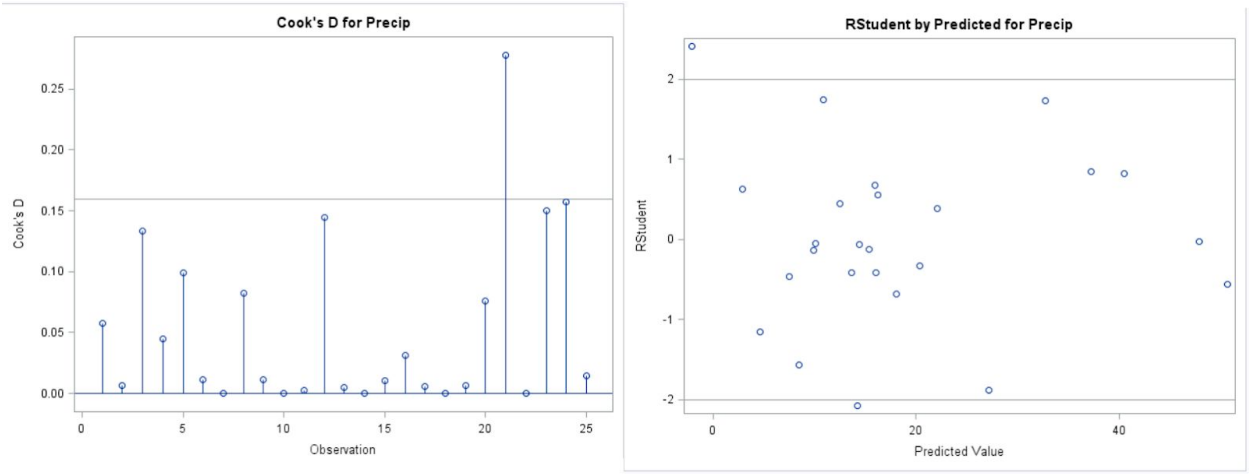


From the graph of Cook's D, there are three potential outliers: the previously identified Observation 19, as well as 3 and 29. From the last graph, we see that only point 3 had large leverage, but was actually not influential under a studentized residual. However, since there are only two influential points under RStudent, 29 and 19, those are the points that we will be considering to be most influential, as on their own Cook's D and RStudent have their own limitations caused by different ways of calculation. So, the next step was to remove points 19 and 29 from the model and to them reexamine the new model. Using the same graphical representations as the initial model, the existence of potential outliers and influential points was examined.

From this residual analysis, we saw that there are 3 potential influential points-15, 18, and 24- identified using Cook's D, with only two points being influential using a studentized residual -15 and 18. We do also see one additional point that has leverage, point 27. However, this point doesn't have significant influence, and actually has a very small Cook's D. The 24th observation is right at the cusp of being influential under a studentized residual, but isn't quite at the required cutoff, so we will leave it in the the regression and just remove the two clearly influential points.

When these next two points are removed, reducing the model to 26 observations, a final regression analysis was tested. This time, only the Cook's D and RStudent vs. Predicted Value was plotted, as we just wanted to observe if there were any remaining influential points.

Table 3. Graphical Representation of Third Residual Analysis



As we see in Table 3, there appears to be at least a one influential point. Ultimately, since there appear to be influential points appearing after each round of removing points, there is strong evidence to suggest that the set-up of the model is off- either an important interaction or a higher order term is likely missing. We wouldn't expect to keep identifying influential points if the model was adequate. To that end, an additional interaction term was included into the original model, and we decided to examine the adequacy of this new model as opposed to our original predicted model.

<u>Addition of interaction terms to original model:</u>

The model was modified to include interaction terms with the variable x1, yielding the following model expression:

$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_1x_2 + \beta_6x_1x_3 + \beta_7x_1x_4$

This generated the following prediction expression:

$E(y) = -148.57892 + 0.03952x_1 + 4.77960x_2 - 0.02675x_3 - 14.75271x_4 - 0.00092794x_1x_2 - 0.00000982x_1x_3 - 0.00303x_1x_4$

For the nested F test to compare whether introducing interaction terms improved the model, we used the formula

F = ((SSE_Reduced - SSE_Complete)/(k-g)) / (MSE_Complete)

where k is the number of β terms in the complete model and g is the number of β terms in the reduced model and are testing the hypothesis:

H_0: $\beta_5 = \beta_6 = \beta_7 = 0$

H_a: At least one of the β terms is not 0

For our models, this became:

F = ((2097.83621 - 1248.78860)/(7-4)) / (56.76312) = 4.98591

This F value yields a p of .008661375, which is below $\alpha = 0.01$. Therefore, we reject the null hypothesis that the interaction terms are not significant to the model.

The REG Procedure
Model: MODEL1
Dependent Variable: Precip

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5913.81738	1478.45434	17.62	<.0001
Error	25	2097.83621	83.91345		
Corrected Total	29	8011.65359			

Root MSE	9.16043	R-Square	0.7382
Dependent Mean	19.80733	Adj R-Sq	0.6963
Coeff Var	46.24766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-97.89872	24.13791	-4.06	0.0004
Altitude	1	0.00221	0.00113	1.95	0.0627
Latitude	1	3.45376	0.65609	5.26	<.0001
Distance	1	-0.05365	0.03879	-1.38	0.1789
SHD	1	-15.85771	4.37100	-3.63	0.0013

The REG Procedure
Model: MODEL1
Dependent Variable: Precip

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	6762.86498	966.12357	17.02	<.0001
Error	22	1248.78860	56.76312		
Corrected Total	29	8011.65359			

Root MSE	7.53413	R-Square	0.8441
Dependent Mean	19.80733	Adj R-Sq	0.7945
Coeff Var	38.03707		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-148.57892	25.64174	-5.79	<.0001
Altitude	1	0.03952	0.01188	3.33	0.0031
Latitude	1	4.77960	0.69469	6.88	<.0001
Distance	1	-0.02675	0.04001	-0.67	0.5107
SHD	1	-14.75271	4.80385	-3.07	0.0056
altLat	1	-0.00092794	0.00033189	-2.80	0.0105
altDist	1	-0.00000982	0.00002194	-0.45	0.6589
altSHD	1	-0.00303	0.00145	-2.08	0.0492

Residual analysis of model after including interaction terms:

After including the interaction terms in the model, we performed a second series of outlier and influential point tests to confirm the interaction terms improved the model overall.

Table 4: Residual Analysis of model including interaction terms

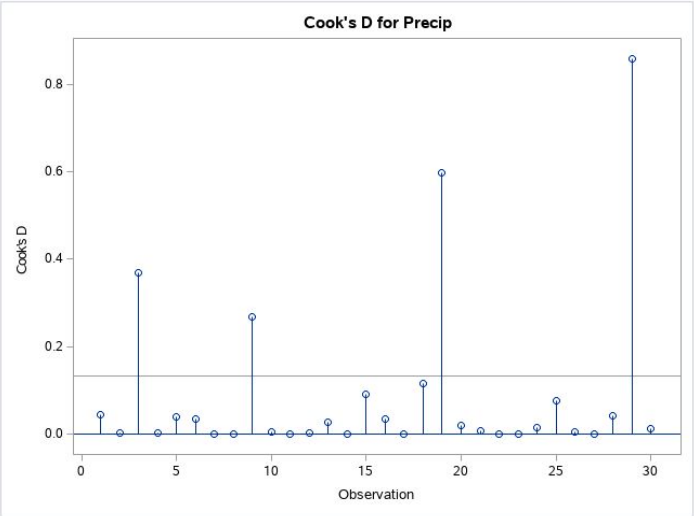
The REG Procedure
Model: MODEL1
Dependent Variable: Precip

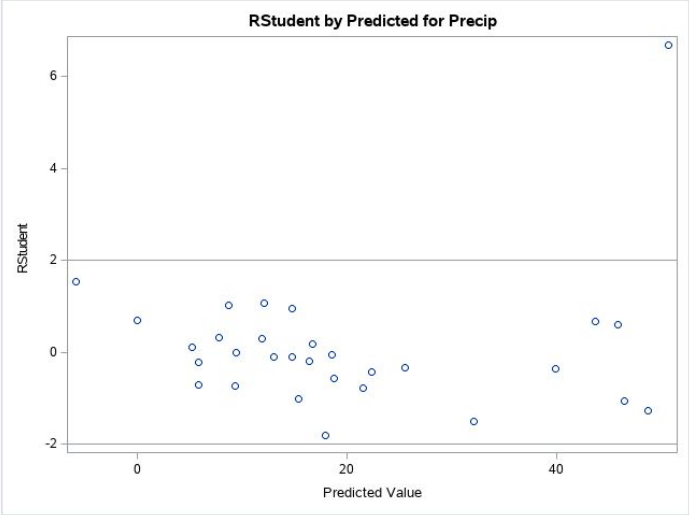
Output Statistics

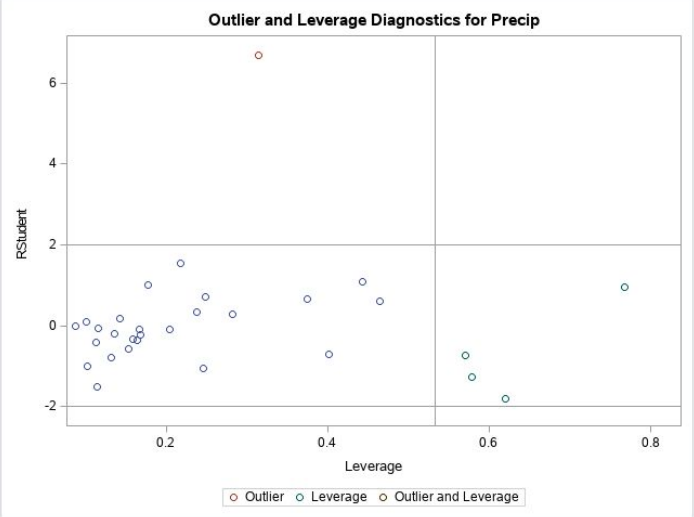
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS							
													Intercept	Altitude	Latitude	Distance	SHD	altLat	altDist	altSHD
1	39.57	46.4730	3.7320	-6.9030	6.545	-1.055	**	0.045	-1.0575	0.2454	1.2694	-0.6030	0.4145	-0.2427	-0.4550	-0.0462	0.2557	0.2455	0.0155	-0.1248
2	23.27	25.6120	3.0015	-2.3420	6.910	-0.339		0.003	-0.3320	0.1587	1.6538	-0.1442	0.1002	-0.0423	-0.1005	-0.0277	-0.0239	0.0363	0.0263	0.0289
3	18.20	14.7755	6.6021	3.4245	3.630	0.943	*	0.368	0.9410	0.7679	4.4921	1.7116	-0.1907	0.7229	0.1772	0.1066	-0.1671	-0.4682	-0.8360	0.6661
4	37.48	39.9287	3.0536	-2.4487	6.888	-0.356		0.003	-0.3484	0.1643	1.6578	-0.1544	0.0813	-0.0471	-0.0936	-0.0046	0.0700	0.0491	0.0019	-0.0359
5	49.26	45.8872	5.1315	3.3728	5.516	0.611	*	0.040	0.6025	0.4639	2.3596	0.5604	0.0741	-0.1271	-0.0780	-0.0901	0.0971	0.1522	0.0197	-0.2904
6	21.82	32.1847	2.5521	-10.3647	7.089	-1.462	**	0.035	-1.5034	0.1147	0.7234	-0.5413	0.1174	-0.0758	-0.1679	0.0027	0.2669	0.0961	-0.0244	-0.1448
7	18.07	18.5429	2.5576	-0.4729	7.087	-0.067		0.000	-0.0652	0.1152	1.6372	-0.0235	0.0083	-0.0037	-0.0084	0.0032	-0.0120	0.0041	-0.0029	0.0107
8	14.17	14.8201	3.4058	-0.6501	6.720	-0.097		0.000	-0.0945	0.2043	1.8173	-0.0479	-0.0046	0.0052	0.0040	0.0356	-0.0405	-0.0027	-0.0184	0.0239
9	42.63	48.7452	5.7301	-6.1152	4.892	-1.250	**	0.268	-1.2672	0.5784	1.9088	-1.4844	0.1937	-0.8711	-0.1916	0.0482	-0.0361	0.8378	-0.3855	0.6976
10	13.85	11.9303	3.9972	1.9197	6.386	0.301		0.004	0.2943	0.2815	1.9538	0.1842	0.0442	-0.0337	-0.0419	-0.1512	0.1587	0.0225	0.0801	-0.0887
11	9.44	9.4662	2.2270	-0.0262	7.197	-0.004		0.000	-0.003551	0.0874	1.5898	-0.0011	0.0000	0.0000	-0.0000	-0.0003	-0.0003	-0.0000	0.0001	0.0003
12	19.33	22.3898	2.5335	-3.0598	7.095	-0.431		0.003	-0.4231	0.1131	1.5284	-0.1511	-0.0545	0.0284	0.0406	0.0292	0.0502	-0.0184	-0.0239	-0.0325
13	15.67	8.7342	3.1728	6.9358	6.833	1.015	**	0.028	1.0157	0.1773	1.2017	0.4716	0.1607	-0.1139	-0.1587	-0.3165	0.3614	0.1276	0.0148	-0.1445
14	6.00	5.2808	2.3978	0.7192	7.142	0.101		0.000	0.0984	0.1013	1.6084	0.0330	0.0146	-0.0083	-0.0146	-0.0119	0.0224	0.0080	0.0032	-0.0124
15	5.73	9.3813	5.6910	-3.6513	4.937	-0.740	*	0.091	-0.7317	0.5706	2.7623	-0.8434	-0.0047	-0.2334	-0.0030	0.0453	0.0142	0.3591	-0.5711	-0.2870
16	47.82	43.7919	4.6121	4.0281	5.957	0.676	*	0.034	0.6676	0.3747	1.9616	0.5168	-0.0259	-0.1939	0.0262	0.1375	-0.1088	0.2486	-0.2289	-0.1140
17	17.95	16.7230	2.8458	1.2270	6.976	0.176		0.001	0.1720	0.1427	1.6734	0.0702	0.0430	-0.0229	-0.0375	-0.0157	-0.0167	0.0187	0.0097	0.0118
18	18.20	12.1587	5.0151	6.0413	5.622	1.074	**	0.115	1.0785	0.4431	1.6926	0.9620	0.0306	-0.2359	-0.0233	-0.0484	-0.0073	0.1100	0.4417	0.4417
19	10.03	17.9937	5.9302	-7.9637	4.647	-1.714	***	0.598	-1.7986	0.6196	1.2121	-2.2953	0.0688	1.3585	-0.0591	-0.1037	0.1052	-1.5096	1.0087	-1.1051
20	4.63	-0.0251	3.7524	4.6551	6.533	0.713	*	0.021	0.7043	0.2481	1.6008	0.4045	0.0473	0.0590	-0.0511	0.2890	-0.1144	-0.0699	-0.0189	0.0308
21	14.74	18.7427	2.9474	-4.0027	6.934	-0.577	*	0.008	-0.5683	0.1530	1.5161	-0.2416	-0.0855	0.0126	0.0720	-0.1054	0.1591	-0.0257	0.0938	-0.0610
22	15.02	16.3852	2.7766	-1.3652	7.004	-0.195		0.001	-0.1906	0.1358	1.6558	-0.0756	-0.0448	0.0208	0.0393	-0.0006	0.0300	-0.0197	0.0058	-0.0158
23	12.36	13.0525	3.0727	-0.6925	6.879	-0.101		0.000	-0.0984	0.1663	1.7339	-0.0439	-0.0302	0.0164	0.0270	0.0051	0.0125	-0.0134	-0.0058	-0.0083
24	8.26	15.4441	2.4046	-7.1841	7.140	-1.006	**	0.014	-1.0065	0.1019	1.1082	-0.3389	0.0518	-0.0097	-0.0540	0.0943	-0.2154	0.0166	-0.0527	0.1636
25	4.05	-5.8670	3.5140	9.9170	6.664	1.488	**	0.077	1.5330	0.2175	0.7936	0.8083	0.3758	-0.1702	-0.3849	0.4022	-0.0242	0.1637	-0.1095	-0.0966
26	9.94	7.7538	3.6749	2.1862	6.577	0.332		0.004	0.3256	0.2379	1.8287	0.1819	0.1467	-0.0819	-0.1356	-0.0409	-0.0248	0.0683	0.0355	0.0229
27	4.25	5.8265	3.0838	-1.5765	6.874	-0.229		0.001	-0.2243	0.1675	1.7098	-0.1006	-0.0229	-0.0242	0.0236	0.0083	-0.0198	0.0144	0.0275	-0.0197
28	1.66	5.8076	4.7731	-4.1476	5.829	-0.712	*	0.042	-0.7033	0.4014	2.0118	-0.5759	0.0767	-0.0283	-0.0691	-0.4917	0.2047	-0.0002	0.2474	0.0126
29	74.87	50.7322	4.2248	24.1378	6.238	3.869	******	0.858	6.6888	0.3144	0.0002	4.5300	-3.4145	1.9992	3.6871	0.4506	-1.8369	-1.9938	-0.1767	0.8710
30	15.95	21.5482	2.7371	-5.5982	7.019	-0.798	*	0.012	-0.7907	0.1320	1.3218	-0.3083	0.1663	-0.0821	-0.1676	-0.0221	-0.0994	0.0804	0.0016	0.1107

At least two outliers are visible from the table: Observation 19 has an RStudent value of -1.7986, and Observation 29 has an RStudent value of 6.6888. We created another set of plots (Cook's Distance, RStudent by Predicted Value, and RStudent by Leverage) to locate any influential points after the change to the model.

Table 5: Graphic Representation of Interaction Term Model Residual Analysis







Four potential outliers appear from the graph of Cook's Distance: Observations 3, 9, 15, and 19.

Conclusion:

We conclude that since the nested F-Test resulted in a p-value below 0.01, we reject the null hypothesis that interaction terms are not significant to the model. Our resulting model includes the interaction terms with respect to x1, and see that it produces a better model than the non-interaction terms model (having an Adj. R-Squared value of 0.7945 compared to 0.6963). We would not have been able to come up with a better fitting model had we not had an extensive knowledge of residual analysis, and the various methods that can be used to find certain points that may be influential which could then be removed. Through testing the fit of each model and graphical analysis, we can finally arrive at a decent fitting model that has a higher Adj. R-Squared value than our initial model.

Appendix

```
data model;
set CALIRAIN;
SHD = 0; * W is baseline ;
if Shadow = "L" then SHD = 1;
run;

proc reg data=model plots(only)=(cooksd rstudentbypredicted rstudentbyleverage);
model Precip = Altitude Latitude Distance SHD / r influence;
run;
```

```
*
  Cook's D (influential points): Obs 3, 19, 29
  RStudent (influential points): Obs 19, 29
  Studentized Residuals (outliers in y): Obs 19, 29
  Leverage (outliers in x): Obs 3
;

data model2;
set model;
if Precip=74.87 then DELETE;
if Precip=10.03 then DELETE;
run;
* Deletes the two influential points common to both Cook's D and RStudent ;

proc reg data=model2 plots(only)=(cooksd rstudentbypredicted rstudentbyleverage);
model Precip = Altitude Latitude Distance SHD / r influence;
run;

*
  Cook's D (influential points): Obs 15, 18, 24
  RStudent (influential points): Obs 15, 18
  Studentized Residuals (outliers in y): Obs 15, 18
  Leverage (outliers in x): Obs 24
;

data model3;
set model2;
if Precip=18.20 then DELETE;
if Precip=5.73 then DELETE;
run;

* Deletes the two influential points common to both Cook's D and RStudent ;

proc reg data=model3 plots(only)=(cooksd rstudentbypredicted rstudentbyleverage);
model Precip = Altitude Latitude Distance SHD / r influence;
run;

* Cook's D (influential points): Obs 21
  RStudent (influential points): Obs 20, 21
;
```

Testing if interaction terms improve the model after residual testing

```
data interactionTerms;
```

```
set model;
altLat = Altitude * Latitude;
altDist = Altitude * Distance;
altSHD = Altitude * SHD;
run;
* Model without interaction terms;
proc reg data = interactionTerms plots = none;
model Precip = Altitude Latitude Distance SHD;
run;
* Model with interaction terms;
proc reg data = interactionTerms plots = none;
model Precip = Altitude Latitude Distance SHD altLat altDist altSHD;
run;
data nestedFtest;
SSE_C = 1248.78860;
SSE_R = 2097.83621;
N = 30;
K = 7;
G = 4;
Fstat = ((SSE_R - SSE_C)/(K-G))/(SSE_C/(N-K-1));
pvalue = SDF('F',Fstat,K-G,N-K-1);
Fcritical = quantile('F',.95,K-G,N-K-1);
proc print data=nestedFtest;
run;
```

Second set of residual/influential point analysis after adding interaction terms;

```
proc reg data=interactionTerms plots(only)=(cooksd rstudentbypredicted rstudentbyleverage);
model Precip = Altitude Latitude Distance SHD altLat altDist altSHD / r influence;
run;
```