Analysis

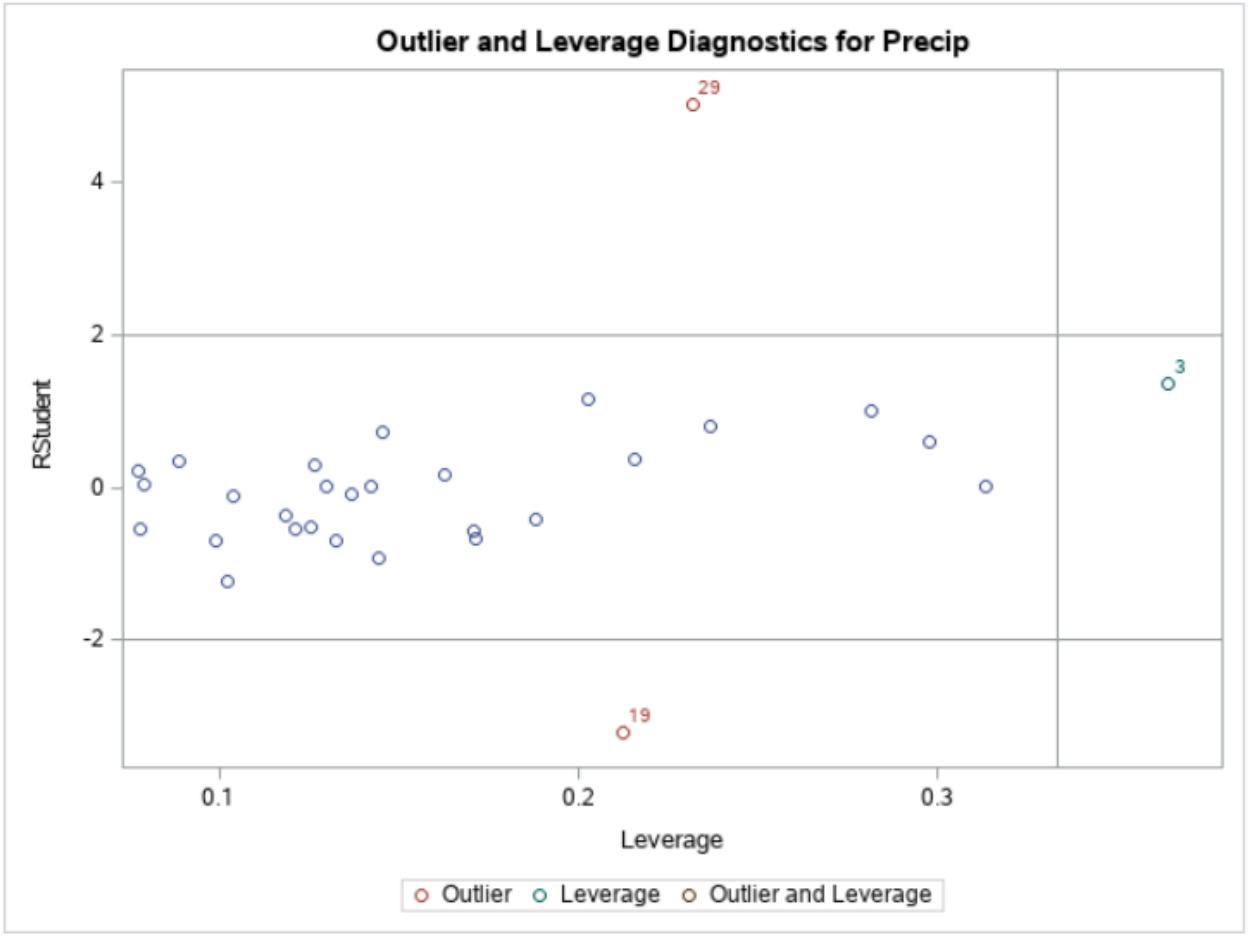

Figure 1. SAS Output of Leverage and Deleted Residuals for Model 2

Observation 3 is an outlier in the x-direction (leverage point).

$h_i > \frac{2(k+1)}{n} = \frac{2(4+1)}{30} = 0.33$. Observation 3 has a leverage greater than 0.33 so observation 3 is a leverage point.

Observations 19 and 29 are outliers in the y-direction since their deleted student residuals are greater than 2.

R-Student Critical Value: 2.06390

R-Student is found using the t-stat at *t*(*n*-*k*-2, alpha/2), where n is the number of observations in the data set, k is the number of independent variables in the model, and alpha is the desired Type 1 error probability (typically 0.05).

Cook's D Critical Value: 0.89425

Cook's D is found using the F-stat at $F(k+1, n-k-1, \text{alpha})$, where n is the number of observations in the data set, k is the number of independent variables in the model, and alpha is the 50^{th} percentile.

The critical values for R-Student and Cook's D were calculated using SAS and the codes used can be found in the Appendix.

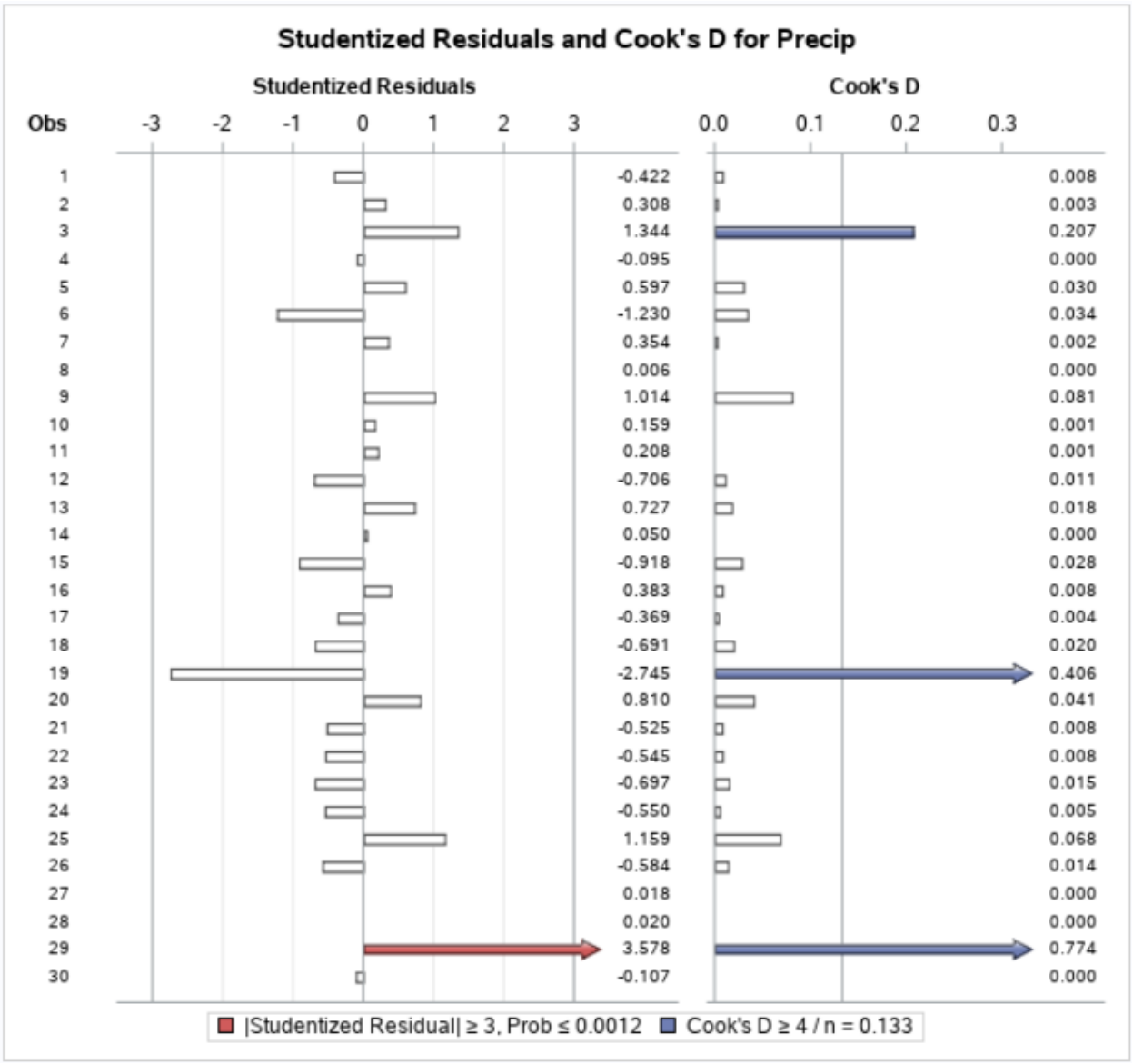

Figure 2. SAS Output for Residual Analysis of Original Model 2

Although no observations exceed the Cook's D critical value of 0.89425, observations 19 and 29 have an absolute value of deleted student residuals (r-student) greater than 2.06390 (-3.2179 and 5.0189 respectively). Therefore, we can conclude that observations 19 and 29 are influential points.

We would handle these observations by removing them from the dataset and rerunning the model. This will hopefully result in a model free of influential points. If not, we would have to consider removing the new influential points as well.

New model with interaction with x_1:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_4$$

First model (No observations removed):

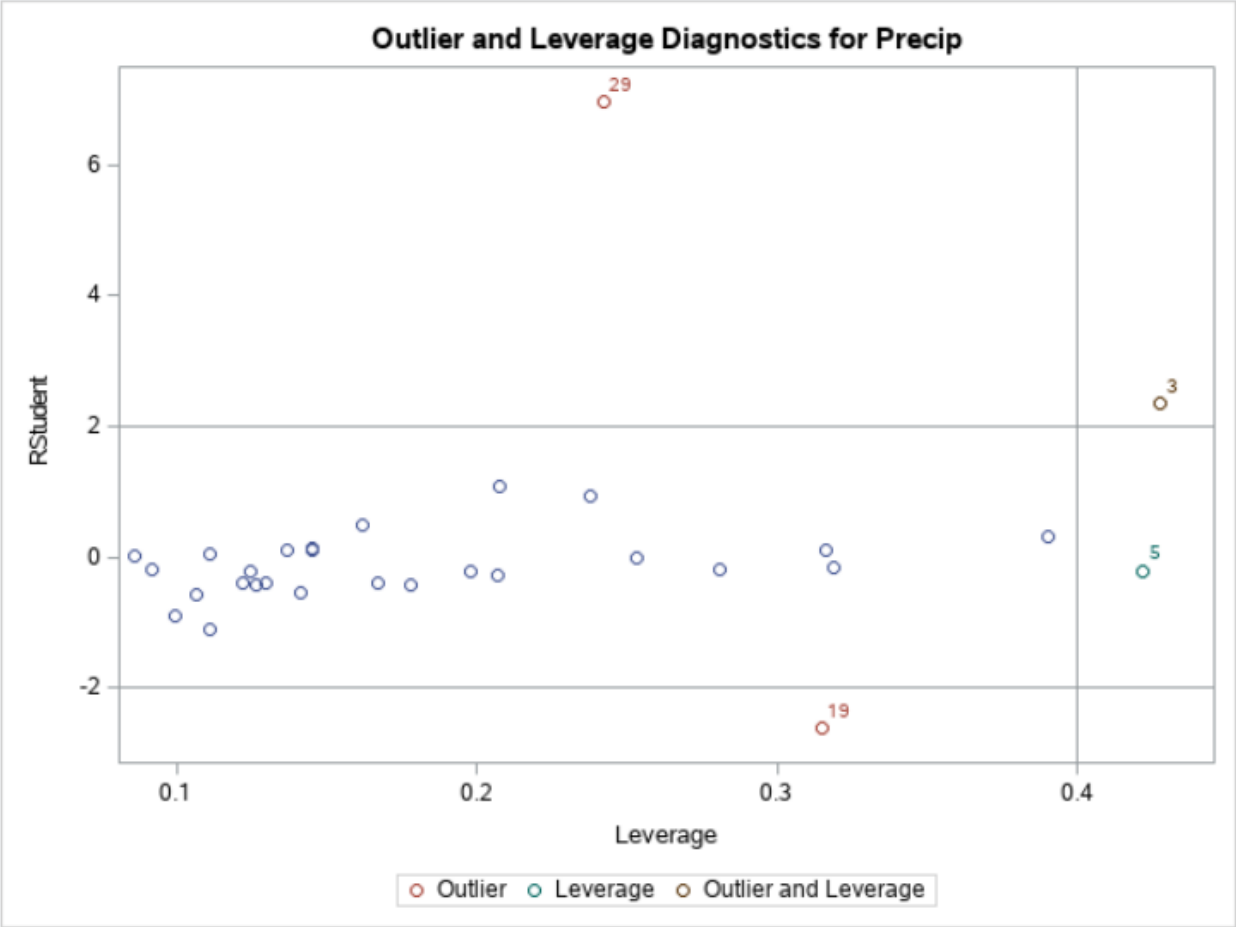

Figure 3. SAS Output of Leverage and Deleted Residuals for Model 2 with Interaction Term

Observations that are outliers in the x-direction (leverage point):

$$h_i > \frac{2(k+1)}{n} = \frac{2(5+1)}{30} = 0.4$$

Observations 3 and 5 have leverages greater than 0.4, thus they are outliers.

Observations that are outliers in the y-direction (deleted student residuals > 2):

Observations 19 and 29 are outliers as they have deleted student residuals greater than 2.

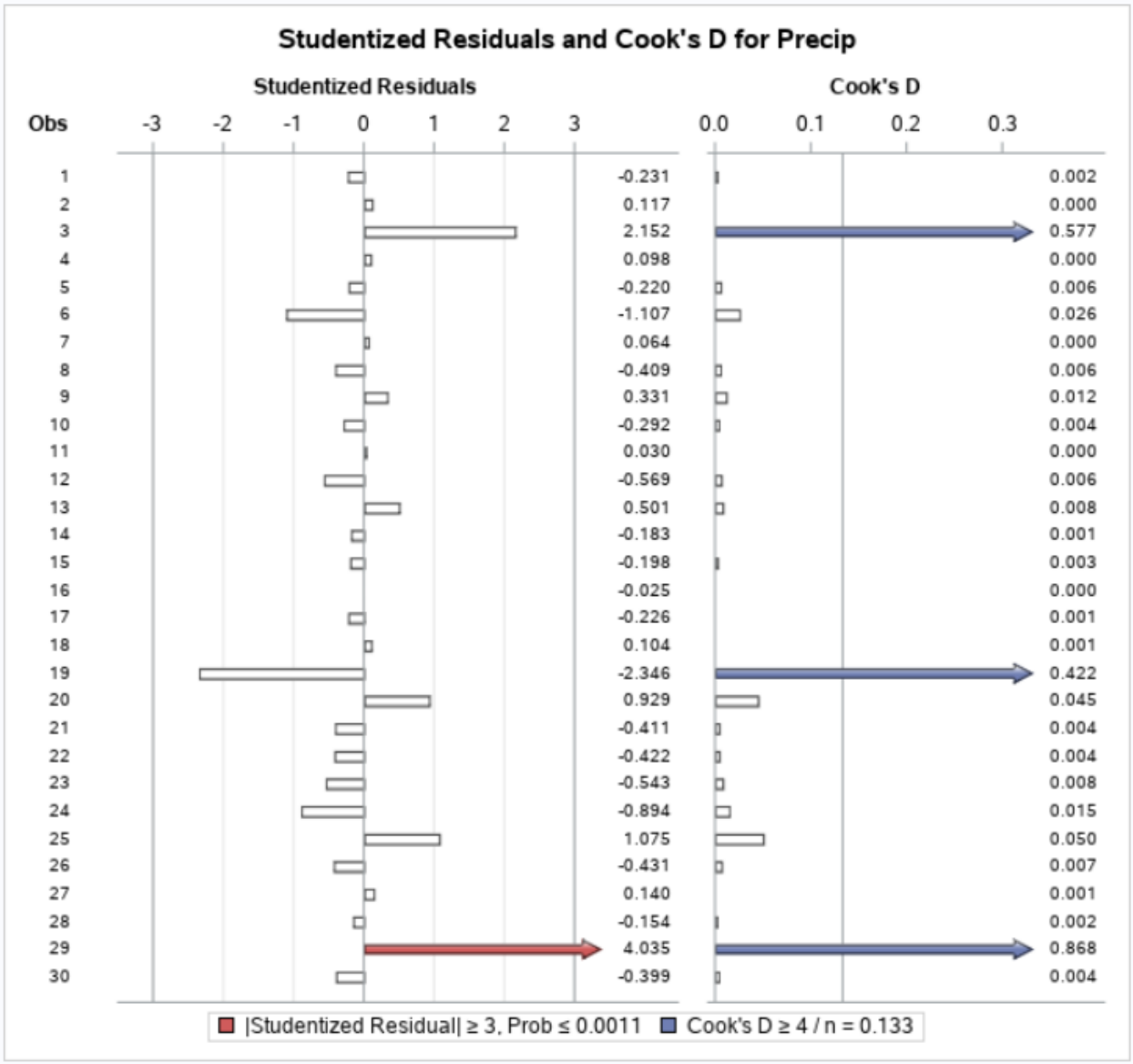

Figure 4. SAS Output for Residual Analysis of Model 2 with Interaction Term

R-Student Critical Value: 2.06866

Cook's D Critical Value: 0.91687

Observations 3, 19, 29 are influential points as they have R-Student values exceeding the critical value. No observation is considered as an influential point based on Cook's D critical value, but observation 29 comes close to being one as it has 0.868 as its Cook's D value.

We therefore removed observations 3, 19, 29 from the dataset and reran the model.

Second Model (Observations 3, 19, 29 Removed)

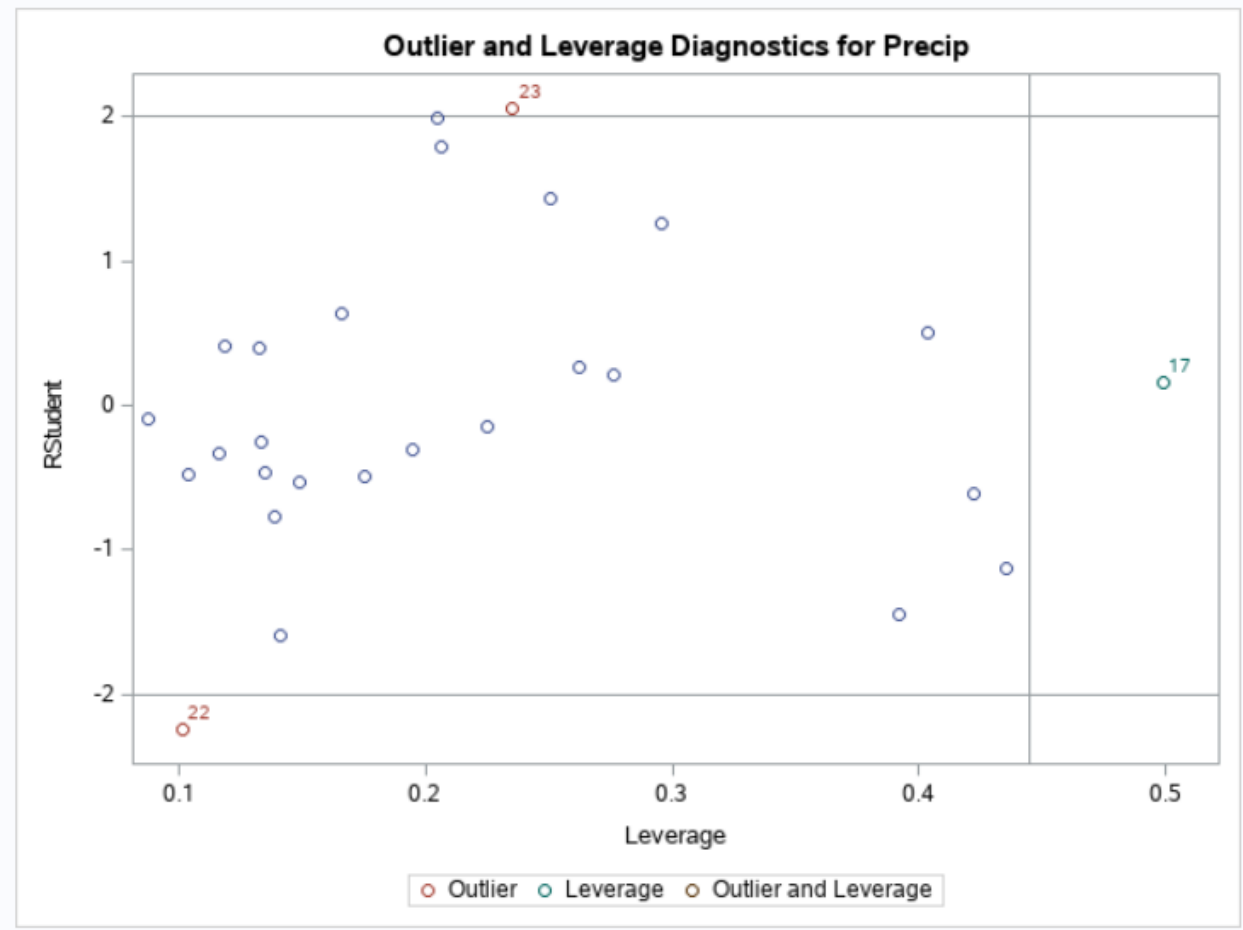

Figure 5. SAS Output of Leverage and Deleted Residuals for Second Model

Observations that are outliers in the x-direction (leverage point):

$$h_i > \frac{2(k+1)}{n} = \frac{2(5+1)}{27} = 0.44$$

Observation 17 has a leverage greater than 0.44, thus it is an outlier.

Observations that are outliers in the y-direction (deleted student residuals > 2):

Observations 22 and 23 are outliers as they have deleted student residuals greater than 2.

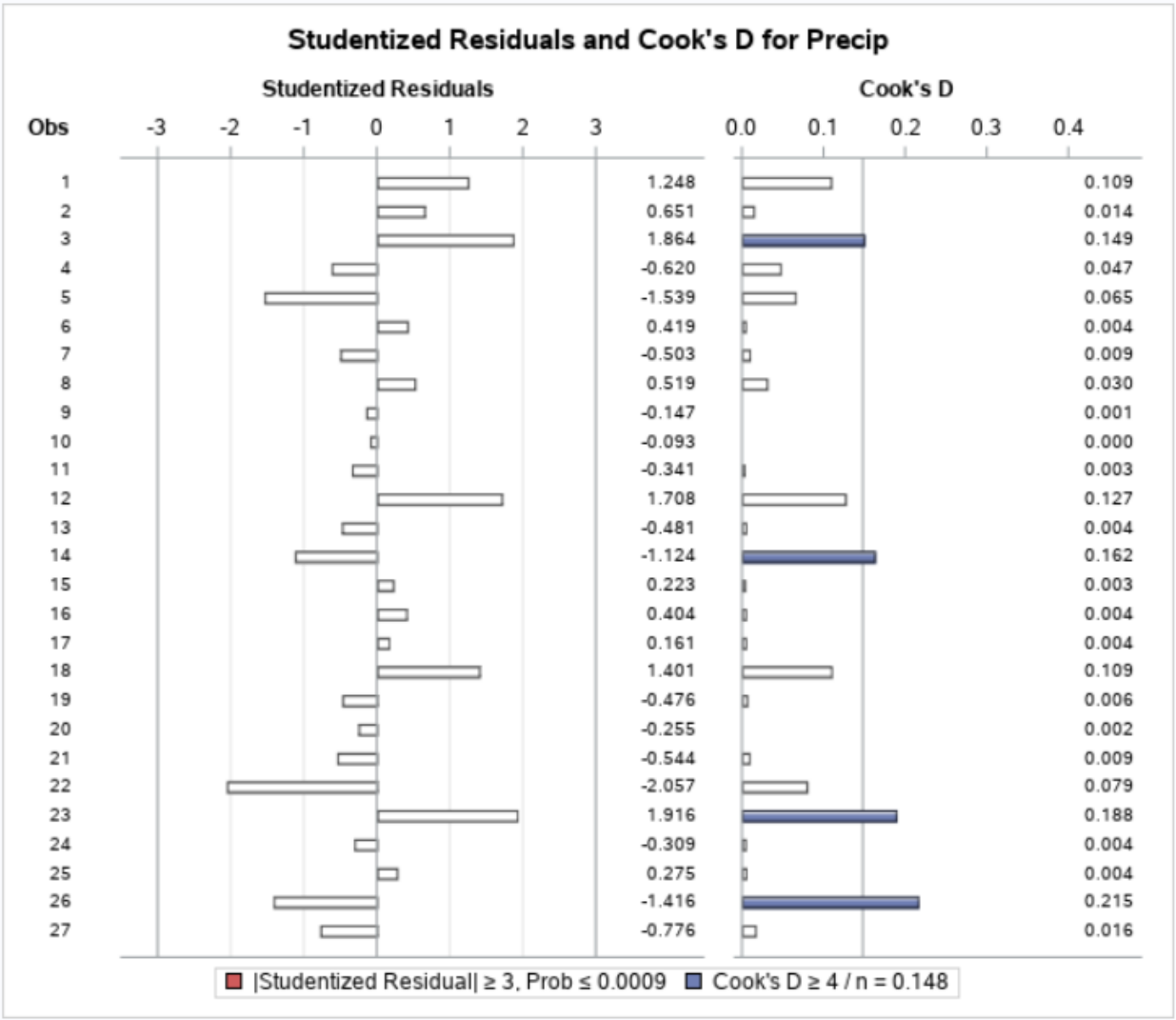

Figure 6. SAS Output for Residual Analysis of Second Model

R-Student Critical Value: 2.08596
Cook's D Critical Value: 0.92060

None of the observations is considered as an influential point as they do not have R-Student values exceeding the critical value. Additionally, no observation is considered as an influential point based on Cook's D critical value.

Since there are no influential points remaining in the model, we will perform a residual analysis to check the assumptions, and then make the appropriate adjustments if needed.

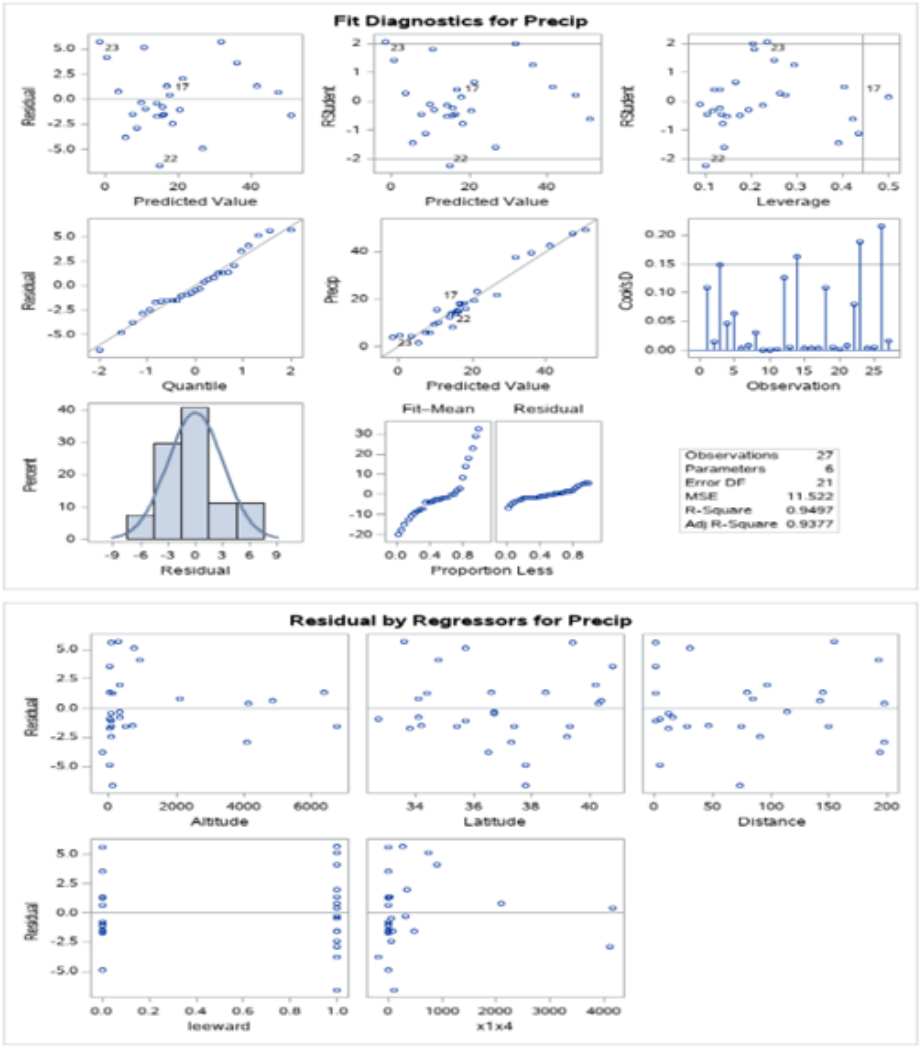

Figure 7. SAS Output for Residuals

The residual assumptions are largely satisfied. There is some evidence of fanning but not enough in our opinion to compromise the model.

We will use the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_4$ without observations 3, 19, 29 for estimating the mean level of precipitation.

Conclusion

After analyzing and adjusting the model three times over, we have come to the conclusion that our second model (without observations 3, 19, and 29) is appropriate for estimating the mean level of precipitation based on four independent variables and an interaction variable: the altitude of the station recording the precipitation level, its latitude, its distance from the Pacific coast, a qualitative variable indicating the direction of the slope upon which the station stands, and an interaction variable between the direction of the slope and the altitude. We were able to remove any influential points from the data in order to restrict outliers from overly influencing the results of our model. Because of this, we believe the resulting model is effective at estimating the precipitation at each station.

The model does not have any influential points, but two of the residual plots exhibit slight fanning. We attempted to apply a transformation to the dependent variable to correct for this, but these transformations lead to new influential points. As a result, we chose to leave the model as is without performing any transformations. Additionally, we attempted similar transformations on earlier forms of the model as well in an attempt to limit the effects of the outliers without needing to drop them. We were unsuccessful in finding a transformation that removed the effects of the influential points. Therefore, we dropped the influential points in the original model rather than using the transformed models. This model could be improved by finding an effective transformation that sets the variance constant for all variables without creating new influential points.

While we have found the interaction of the qualitative slope direction variable and the altitude variable to be beneficial to the overall accuracy of the model, we did not check the hypothetical interaction variable with the slope direction variable and the other two quantitative variables. We acknowledge that the model could be further improved with the addition of these interaction variables, but further analysis would have to be performed in order to be sure.

Appendix
SAS Codes Used

```
data cs4;
infile '/folders/myfolders/CALIRAIN.txt' dlm='09'x firstobs=2;
input station name  $ Precip    Altitude        Latitude        Distance        Shadow $;
run;

data sad;
set cs4;
leeward = 0;
if shadow = 'L' then leeward = 1;
x1x4 = altitude*leeward;
precip3=precip**3;
precip2=precip**2;
precipln=log(precip);
precipsqrt=precip**.5;
precipe=exp(precip);
run;

*Residual analysis with original model 2
proc reg data=sad plots(label)=(cooksd rstudentbypredicted
rstudentbyleverage);
model precip = altitude latitude distance leeward / r influence;
run;

*Calculation for RStudent and Cook's D Critical Values for the original model 2
data teststat;
t=quantile('T',0.975,30 - 4 - 2);
f=quantile('F',0.5,4+1,30-4-1);
run;
proc print data=teststat;
run;
*t=2.06390;
*f=0.89425;

*Residual analysis for new second model with interaction
proc reg data=sad plots(label)=(cooksd rstudentbypredicted
rstudentbyleverage);
model precip = altitude latitude distance leeward x1x4 / r influence;
run;

*New second model calculation for RStudent and Cook's D critical values
data teststat;
t=quantile('T',0.975,30 - 5 - 2);
f=quantile('F',0.5,5+1,30-5-1);
run;
```

```
proc print data=teststat;
run;
*t=2.06866;
*f=0.91687;

*New second model with observations 3,19,29 removed;
data cs44;
input station name  $ Precip Altitude Latitude Distance Shadow $;
cards;
1 Eureka 39.57 43 40.8 1 W
2 RedBluff 23.27 341 40.2 97 L
4 FortBragg 37.48 74 39.4 1 W
5 SodaSprings 49.26 6752 39.3 150 W
6 SanFrancisco 21.82 52 37.8 5 W
7 Sacramento 18.07 25 38.5 80 L
8 SanJose 14.17 95 37.4 28 L
9 GiantForest 42.63 6360 36.6 145 W
10 Salinas 13.85 74 36.7 12 L
11 Fresno 9.44 331 36.7 114 L
12 PtPiedras 19.33 57 35.7 1 W
13 PasaRobles 15.67 740 35.7 31 L
14 Bakersfield 6.00 489 35.4 75 L
15 Bishop 5.73 4108 37.3 198 L
16 Mineral 47.82 4850 40.4 142 W
17 SantaBarbara 17.95 120 34.4 1 W
18 Susanville 18.20 4152 40.3 198 L
20 Needles 4.63 913 34.8 192 L
21 Burbank 14.74 699 34.2 47 W
22 LosAngeles 15.02 312 34.1 16 W
23 LongBeach 12.36 50 33.8 12 W
24 LosBanos 8.26 125 37.8 74 L
25 Blythe 4.05 268 33.6 155 L
26 SanDiego 9.94 19 32.7 5 W
27 Daggett 4.25 2105 34.1 85 L
28 DeathValley 1.66 -178 36.5 194 L
30 Colusa 15.95 60 39.2 91 L
;
run;

data sad1;
set cs44;
leeward = 0;
if shadow = 'L' then leeward = 1;
x1x4 = altitude*leeward;
run;
```

```
*Residual analysis for new second model without the 3 outliers
proc reg data=sad1 plots(label)=(cooksd rstudentbypredicted
rstudentbyleverage);
model precip = altitude latitude distance leeward x1x4 / r influence;
run;

*Calculation for RStudent and Cook's D Critical Values for the second model without 3 outliers
data teststat;
t=quantile('T',0.975,27 - 5 - 2);
f=quantile('F',0.5,5+1,27-5-1);
run;
proc print data=teststat;
run;
*t=2.08596;
*f=0.92060;

*Variable transformation with all observations included
proc reg data=sad plots(label)=(cooksd rstudentbypredicted
rstudentbyleverage);
model precipln = altitude latitude distance leeward x1x4  / r influence;
run;
*precipln -> transformed y variable. It was swapped with precip2/precip3/precipsqrt/precipe to
test other variable transformations.
```