problem 1:
formally show that the average of a hat diagonal is p/n

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p}{n}$$

Solution:-
The element of h_{ij} of matrix H may be interpreted as the amount of leverage exerted by the i^{th} observation y_i on the j^{th} fitted value $\hat{y}_j$.
The hat matrix diagonal is standardized measure of the i^{th} observation from the center of the x-space, Thus large hat diagonal reveal observations that are potentially influential because they are remote in x space from the rest of the sample space.

- a general guideline is to flag cases where $h_{ii} > 2p/n$, where,
p → is number of coln of X, equal to (k+1) in MLR

$$\hat{y}_i = (Hy)_i$$

A linear combination of all responses often it is difficult to find a case with leaverage by examine each predictor separately.

We know,

$$\text{var}(\hat{y}_i) = h_{ii}\,\sigma^2$$

$$\text{var}(e_i) = (1 - H_{ii})\,\sigma^2$$

In simple linear regression —

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

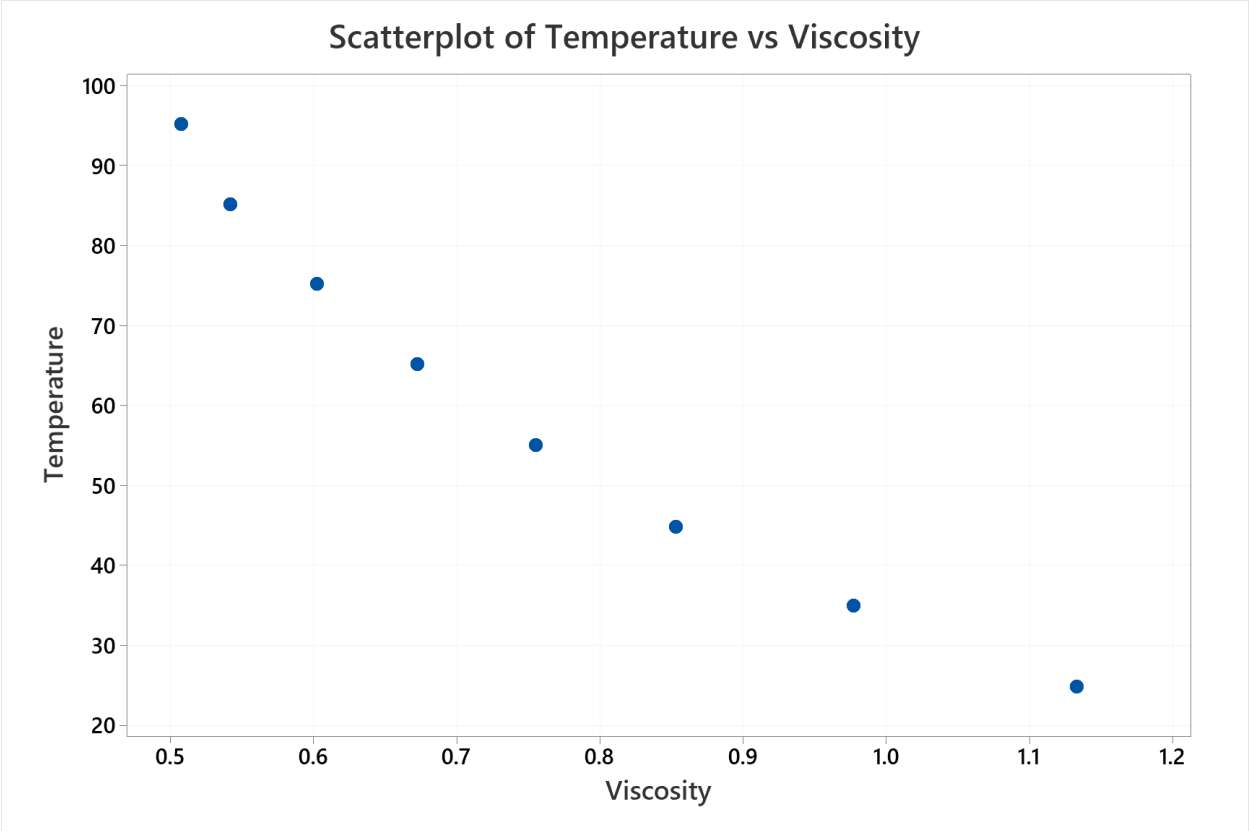
$$\bar{h} = \frac{(k+1)}{n} = \frac{p}{n}$$

hence $\boxed{\bar{h} \Rightarrow \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{p}{n}}$

Problem 2:

Solution:

a. Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?



Yes, it is a straight line model. If the scatter plot is a straight line model then the model is not adequate.

b. Fit the straight - line model. Compute the summary statistics and the residual plots. What are your conclusions regarding model adequacy?

Regression Equation

Viscosity = 1.2815 - 0.008758 Temperature

Coefficients

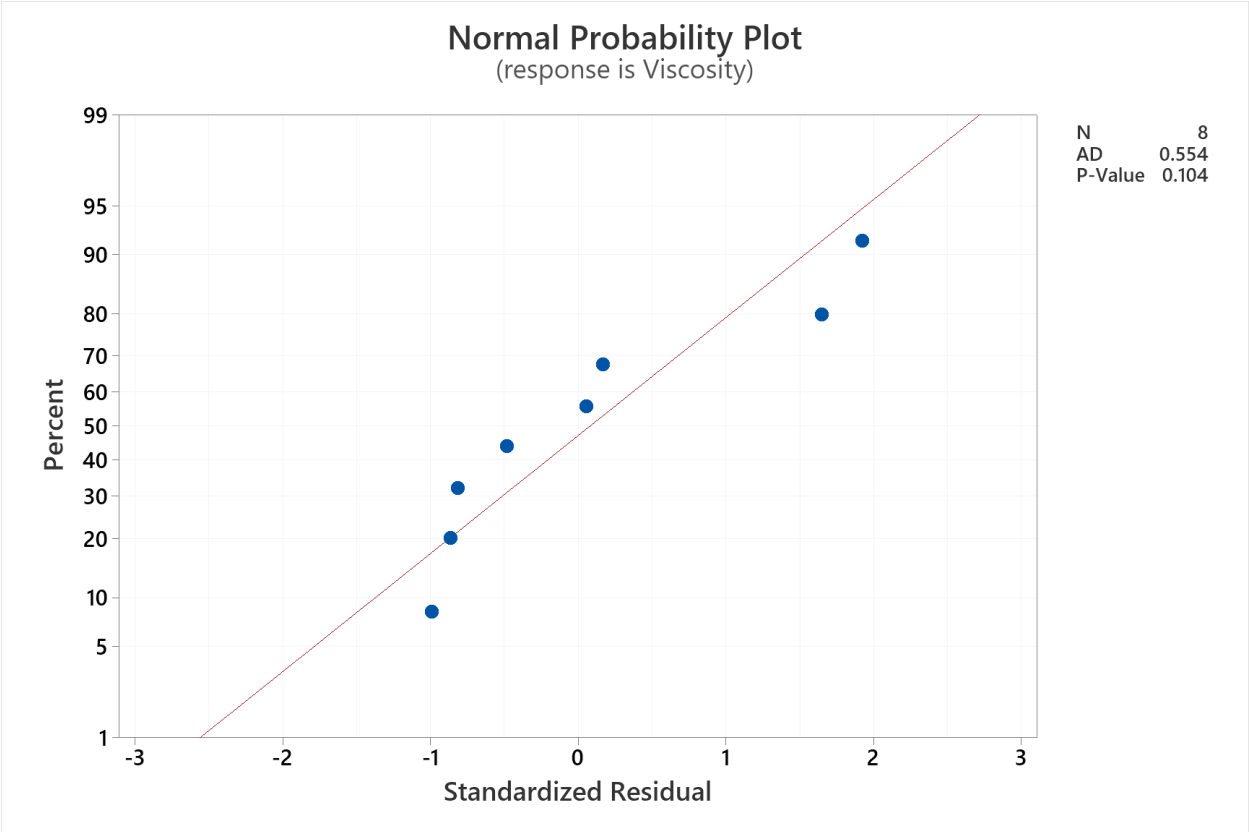
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	1.2815	0.0469	(1.1668, 1.3962)	27.34	0.000	
Temperature	-0.008758	0.000728	(-0.010540, -0.006976)	-12.02	0.000	1.00

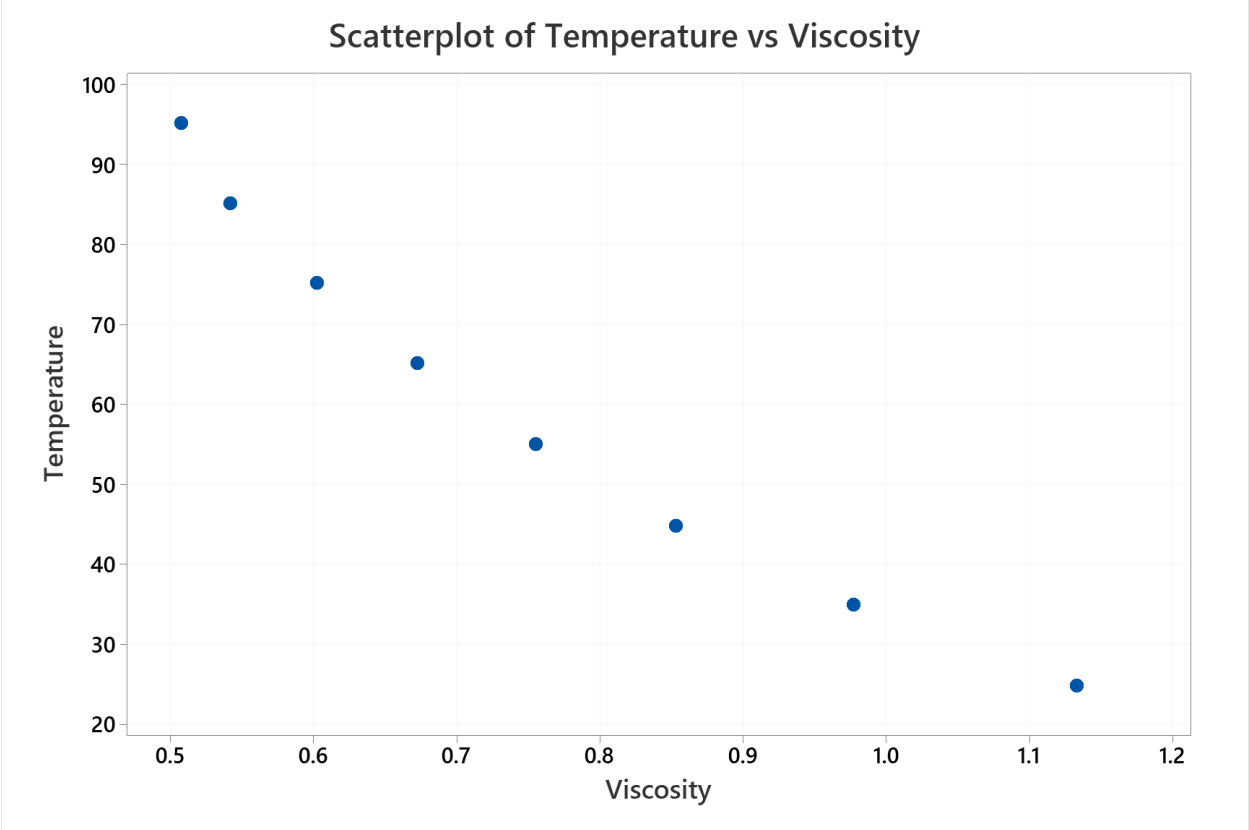
Model Summary

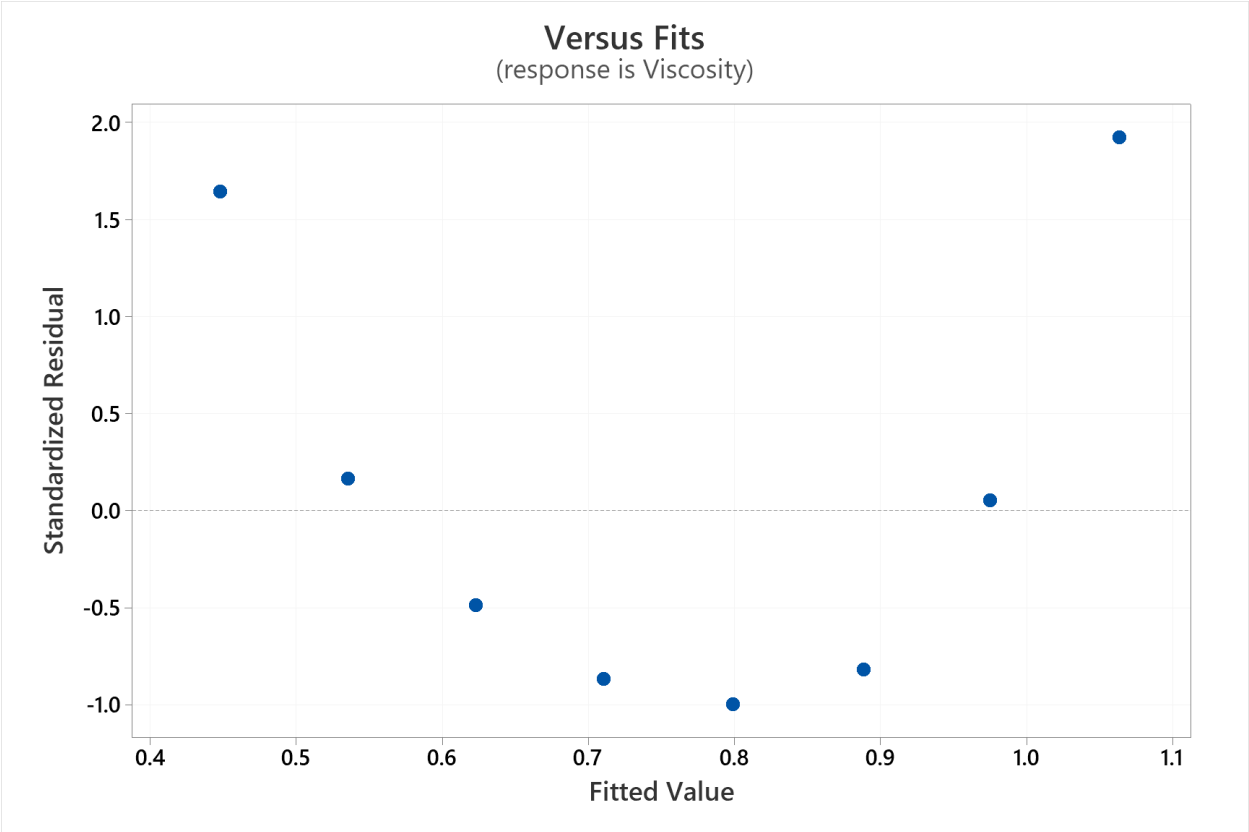
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0474336	96.02%	95.35%	0.0317038	90.64%	-16.37	-22.13

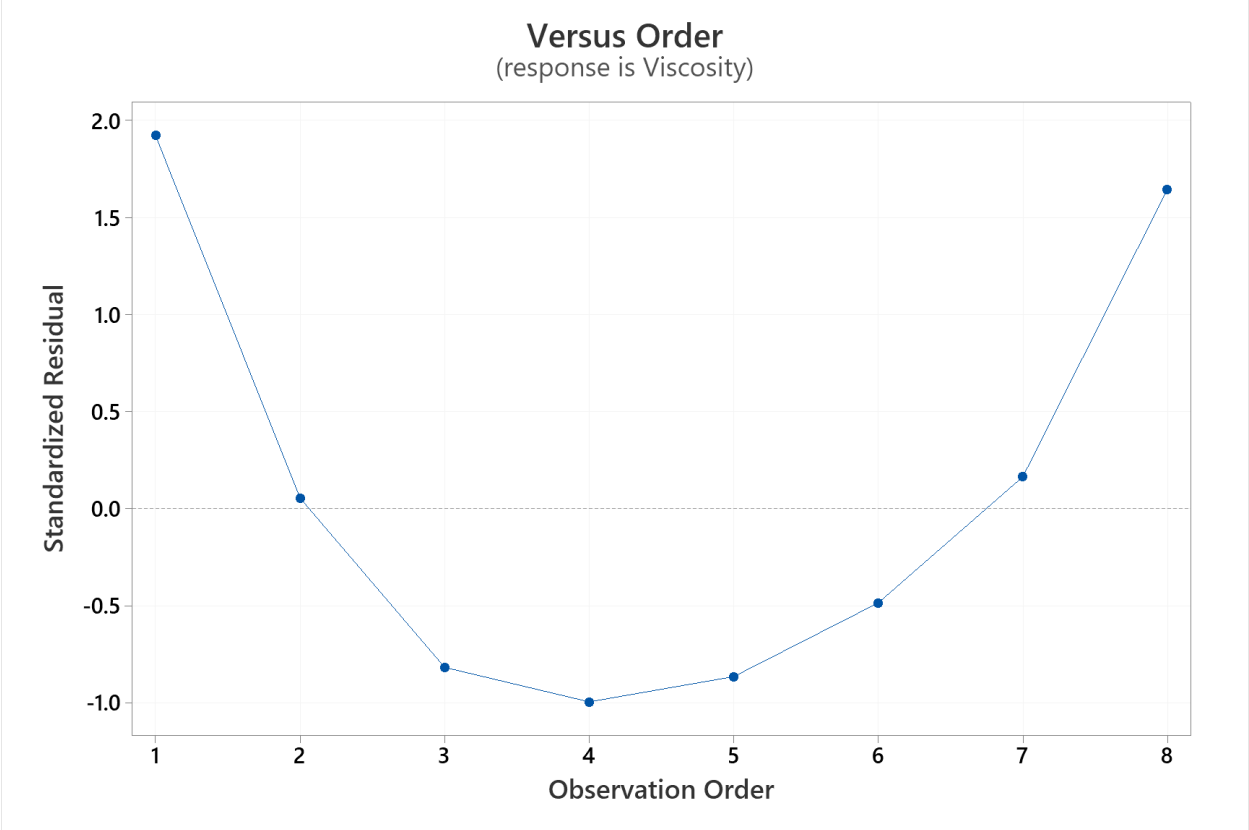
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.32529	96.02%	0.32529	0.325292	144.58	0.000
Temperature	1	0.32529	96.02%	0.32529	0.325292	144.58	0.000
Error	6	0.01350	3.98%	0.01350	0.002250		
Total	7	0.33879	100.00%				


Normal Probability Plot
(response is Viscosity)
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-3
-2
-1
0
1
2
3
Standardized Residual
N 8
AD 0.554
P-Value 0.104


Scatterplot of Temperature vs Viscosity
Temperature
100
90
80
70
60
50
40
30
20
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Viscosity


Versus Fits
(response is Viscosity)
Standardized Residual
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Fitted Value


Versus Order
(response is Viscosity)
Standardized Residual
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
6
7
8
Observation Order

R-square is 96.02%

To determine whether a linear model is appropriate, we examine the residual plot. If we see a curved relationship in the residual plot, the linear model is not appropriate.

The scatter plot is a straight-line model and there is a curve in residual data hence the assumptions for model adequacy are violated.

c. Basic principles of physical chemistry suggest that the viscosity is an exponential function of the temperature. Repeat part b using the appropriate transformation based on this information.

Regression Equation

Viscosity = 2.6651 - 0.47622 logeTemp

Coefficients

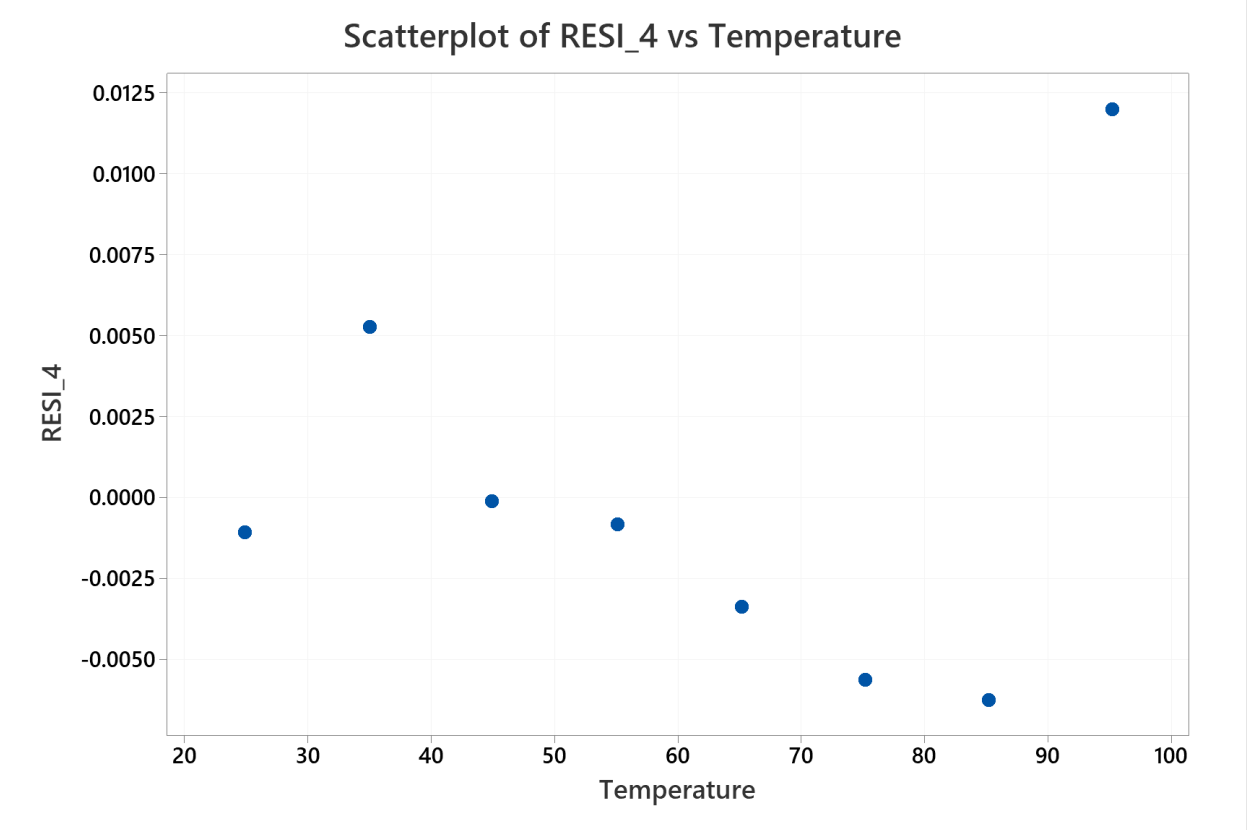
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.6651	0.0215	(2.6124, 2.7178)	123.72	0.000	
logeTemp	-0.47622	0.00534	(-0.48929, -0.46315)	-89.17	0.000	1.00

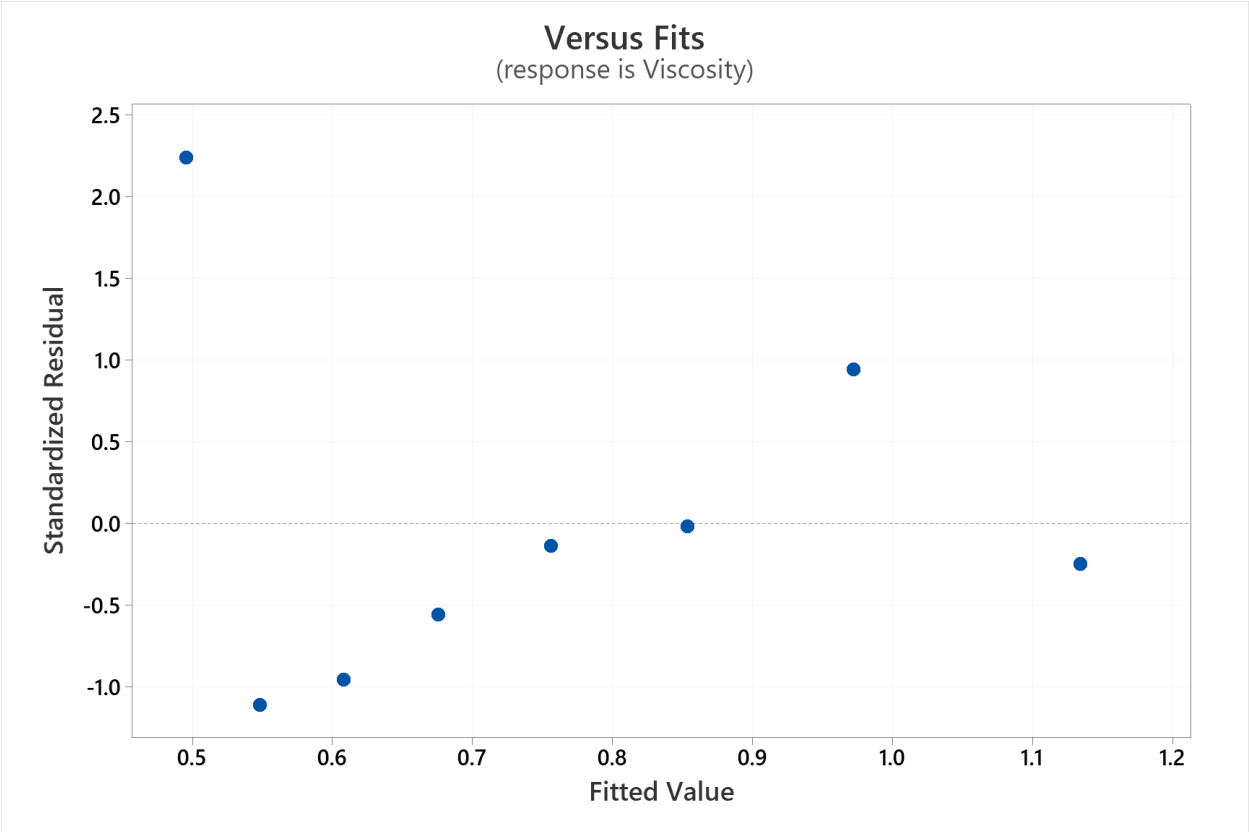
Model Summary

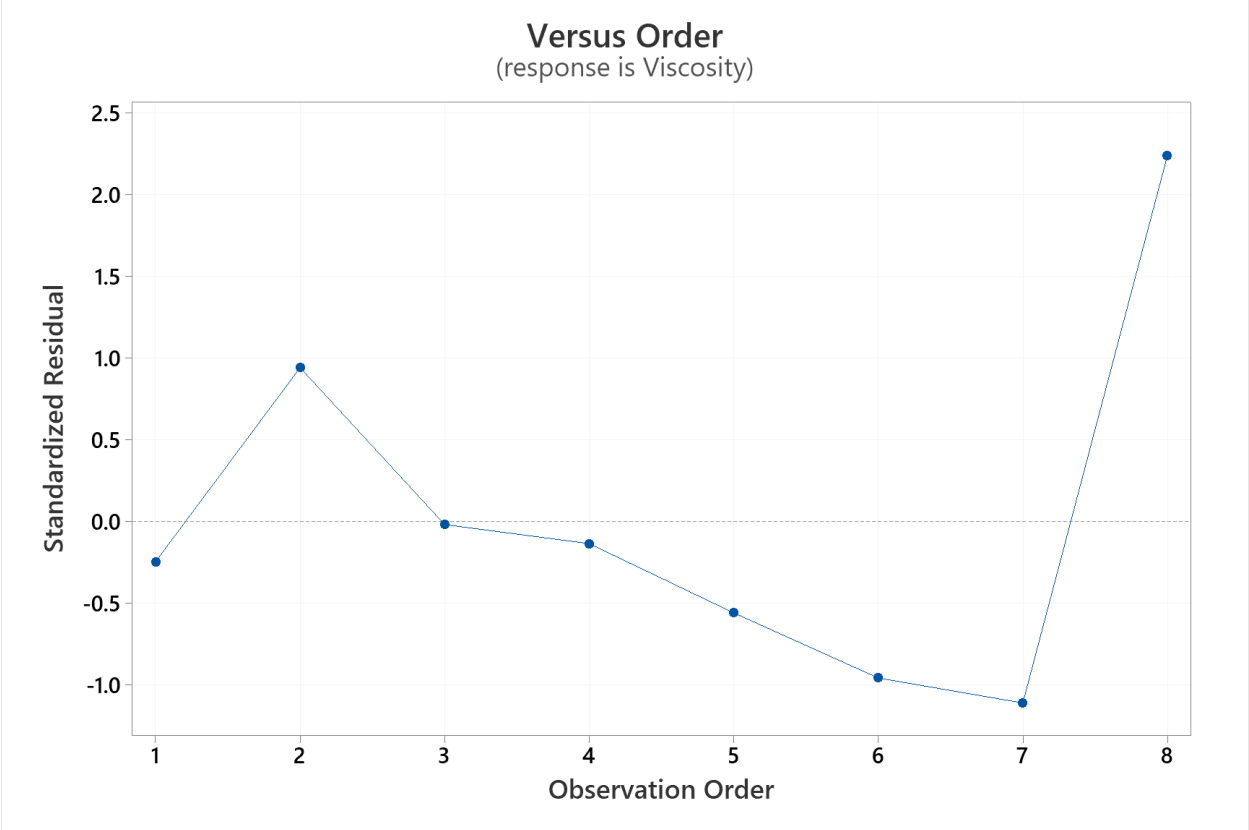
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0065249	99.92%	99.91%	0.0005062	99.85%	-48.11	-53.87

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.338536	99.92%	0.338536	0.338536	7951.59	0.000
logeTemp	1	0.338536	99.92%	0.338536	0.338536	7951.59	0.000
Error	6	0.000255	0.08%	0.000255	0.000043		
Total	7	0.338792	100.00%				




Versus Fits
(response is Viscosity)
Standardized Residual
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Fitted Value


Versus Order
(response is Viscosity)
Standardized Residual
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
6
7
8
Observation Order

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
	Temperature	Viscosity	RESI1	FITS	RESI	FITS_1	RESI_1	HI	FITS_2	RESI_2	FITS_3	RESI_3	log10Temp	RESI_4
1	24.9	1.1330	-0.0641049	18.6761	6.22393	1.06344	0.0695591	0.416941	1.06344	0.0695591	18.6761	6.22393	1.39620	-0.0010760
2	35.0	0.9772	-0.0956711	35.7570	-0.75702	0.97499	0.0022131	0.273400	0.97499	0.0022131	35.7570	-0.75702	1.54407	0.0052677
3	44.9	0.8532	-0.0342627	49.3516	-4.45161	0.88828	-0.0350844	0.179387	0.88828	-0.0350844	49.3516	-4.45161	1.65225	-0.0001106
4	55.1	0.7550	-0.0660624	60.1176	-5.01765	0.79895	-0.0439546	0.130865	0.79895	-0.0439546	60.1176	-5.01765	1.74115	-0.0008225
5	65.2	0.6723	-0.0207350	69.1844	-3.98437	0.71050	-0.0382006	0.131163	0.71050	-0.0382006	69.1844	-3.98437	1.81425	-0.0033699
6	75.2	0.6021	-0.0507510	76.8807	-1.68066	0.62292	-0.0208224	0.178851	0.62292	-0.0208224	76.8807	-1.68066	1.87622	-0.0056171
7	85.2	0.5420	-0.0124967	83.4697	1.73035	0.53534	0.0066558	0.273696	0.53534	0.0066558	83.4697	1.73035	1.93044	-0.0062609
8	95.2	0.5074	-0.0228784	87.2630	7.93702	0.44777	0.0596340	0.415698	0.44777	0.0596340	87.2630	7.93702	1.97864	0.0119894
9														

To calculate the exponential function of temperature we have calculated the log to the base e of temperature and fitted the model with viscosity. After fitting the regression model, We are getting above results and graphs.

As we can see in the graphs,

R-square is 99% which is slightly increased than the previous regression model.

To determine whether a linear model is appropriate, we examine the residual plot. If we see a curved relationship in the residual plot, the linear model is not appropriate.

The versus fitted graph is not a curved shape and the scatter plot does not follow the straight line model. Also, From the above statistic, we can say that the P-value is small.

From the above observations, The new regression model is appropriate and does not violate the assumptions.

Problem 3:

Perform a thorough influence analysis of the solar thermal energy test data given in Table B.2. Discuss your results.

Solution:

Regression Equation

y = 183 + 0.0875 x1 + 3.85 x2 + 4.27 x3 - 19.16 x4 + 2.66 x5

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
12.5916	75.45%	70.11%	6121.27	58.78%	241.83	246.06

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	11205.3	75.45%	11205.3	2241.1	14.13	0.000
x1	1	4759.7	32.05%	588.2	588.2	3.71	0.067
x2	1	982.0	6.61%	615.9	615.9	3.88	0.061
x3	1	1826.9	12.30%	552.7	552.7	3.49	0.075
x4	1	3496.8	23.54%	3246.8	3246.8	20.48	0.000
x5	1	139.8	0.94%	139.8	139.8	0.88	0.357
Error	23	3646.6	24.55%	3646.6	158.5		
Total	28	14851.9	100.00%				

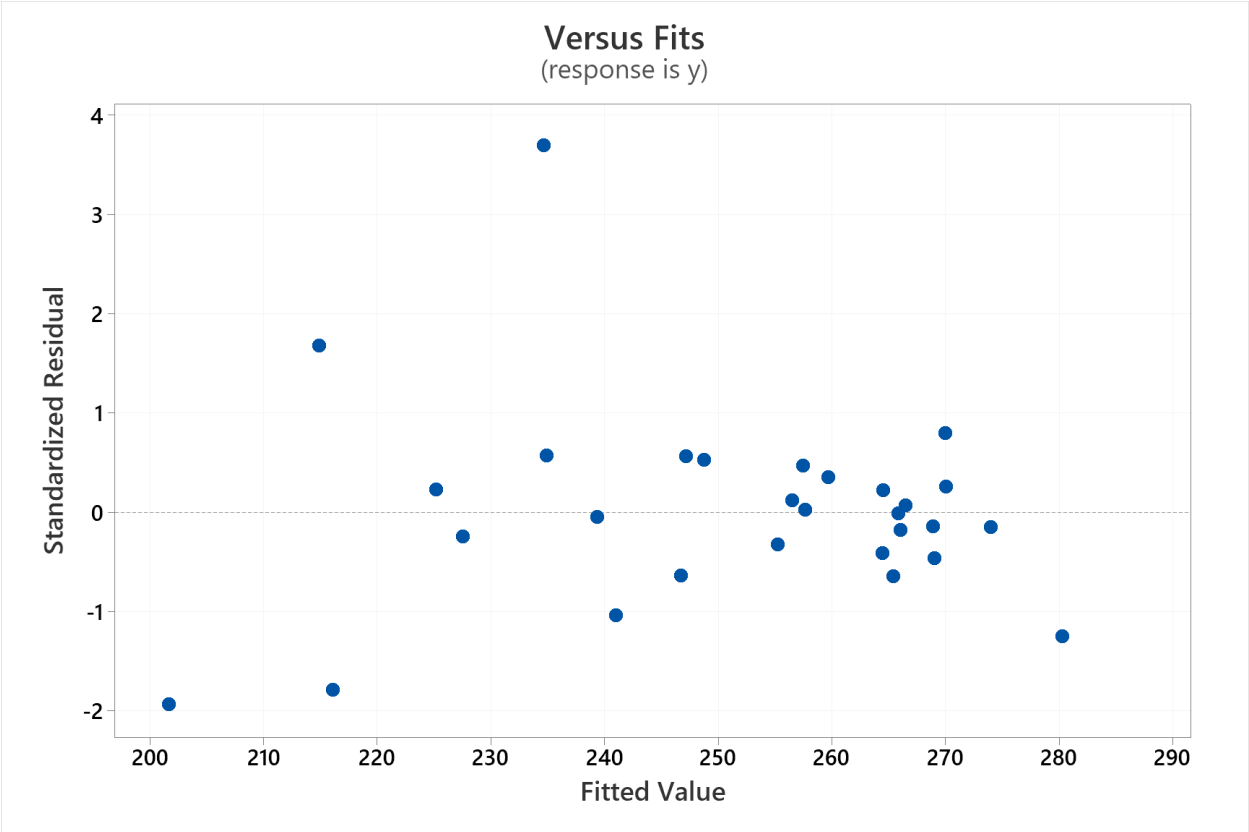
Fits and Diagnostics for Unusual Observations

Obs	y	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
1	271.80	270.04	10.58	(248.16, 291.92)	1.76	0.26	0.25	0.705461
17	277.20	234.62	5.05	(224.17, 245.08)	42.58	3.69	5.66	0.161091

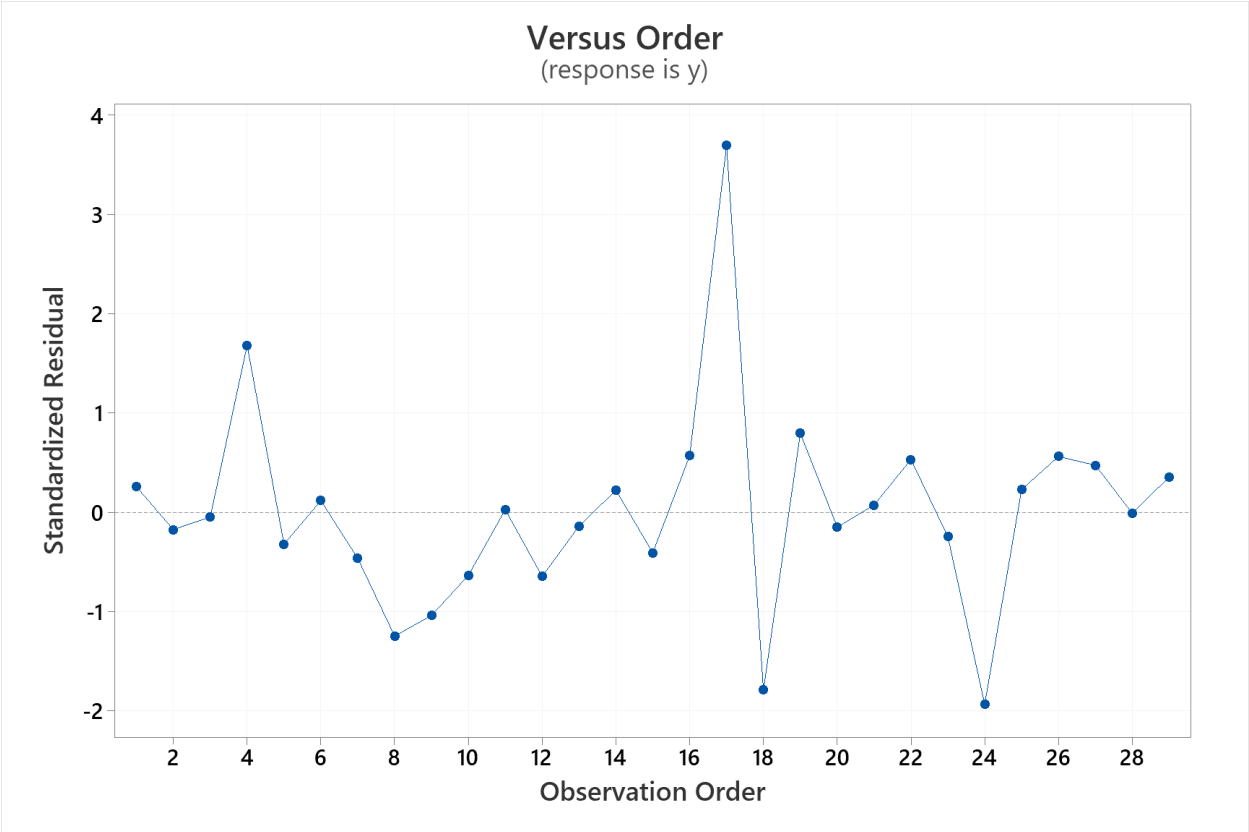
Obs	Cook's D	DFITS	
1	0.03	0.39050	X
17	0.44	2.47851	R

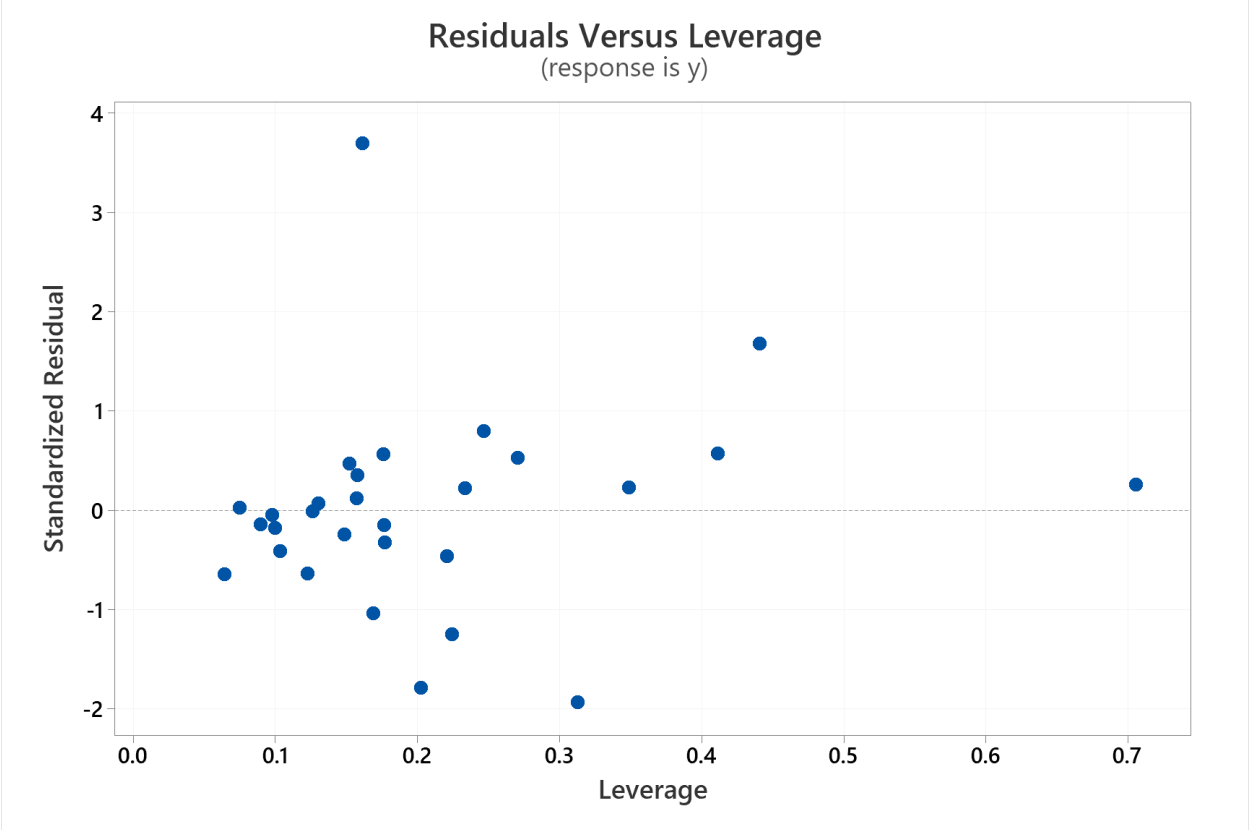
R Large residual
X Unusual X

y	x1	x2	x3	x4	x5	FITS	RESI	Leverage	FITS_1	RESI_1	Leverage_	COOK
271.8	783.35	33.53	40.55	16.66	13.20	270.039	1.7605	0.705461	270.039	1.7605	0.705461	0.026494
264.0	748.45	36.50	36.19	16.46	14.11	266.055	-2.0551	0.099829	266.055	-2.0551	0.099829	0.000547
238.8	684.45	34.66	37.31	17.66	15.68	239.333	-0.5327	0.097692	239.333	-0.5327	0.097692	0.000036
230.7	827.80	33.13	32.52	17.50	10.53	214.875	15.8245	0.440963	214.875	15.8245	0.440963	0.371421
251.6	860.45	35.75	33.71	16.40	11.00	255.242	-3.6418	0.176817	255.242	-3.6418	0.176817	0.003638
257.9	875.15	34.46	34.14	16.28	11.31	256.518	1.3819	0.157045	256.518	1.3819	0.157045	0.000444
263.9	909.45	34.60	34.85	16.06	11.96	269.036	-5.1363	0.220710	269.036	-5.1363	0.220710	0.010079
266.5	905.55	35.38	35.89	15.93	12.58	280.285	-13.7851	0.224346	280.285	-13.7851	0.224346	0.074487
229.1	756.00	35.85	33.53	16.60	10.66	240.985	-11.8852	0.169111	240.985	-11.8852	0.169111	0.036373
239.3	769.35	35.68	33.79	16.41	10.85	246.755	-7.4552	0.122370	246.755	-7.4552	0.122370	0.009282
258.0	793.50	35.35	34.72	16.17	11.41	257.658	0.3417	0.074643	257.658	0.3417	0.074643	0.000011
257.6	801.65	35.04	35.22	15.92	11.91	265.434	-7.8339	0.063917	265.434	-7.8339	0.063917	0.004706
267.3	819.65	34.07	36.50	16.04	12.85	268.940	-1.6404	0.089362	268.940	-1.6404	0.089362	0.000305
267.0	808.55	32.20	37.60	16.19	13.58	264.531	2.4690	0.233600	264.531	2.4690	0.233600	0.002549
259.6	774.95	34.32	37.89	16.62	14.21	264.436	-4.8362	0.103433	264.436	-4.8362	0.103433	0.003164
240.4	711.85	31.08	37.71	17.37	15.56	234.881	5.5188	0.411107	234.881	5.5188	0.411107	0.037954
277.2	694.85	35.73	37.00	18.12	15.83	234.625	42.5752	0.161091	234.625	42.5752	0.161091	0.436154
196.0	638.10	34.11	36.76	18.53	16.41	216.079	-20.0789	0.202199	216.079	-20.0789	0.202199	0.134633
278.7	774.55	34.79	34.62	15.54	13.10	269.984	8.7162	0.246732	269.984	8.7162	0.246732	0.034727
272.3	757.90	35.77	35.40	15.70	13.63	273.980	-1.6804	0.176443	273.980	-1.6804	0.176443	0.000772
267.4	753.35	36.44	35.96	16.45	14.51	266.525	0.8747	0.130428	266.525	0.8747	0.130428	0.000139
254.5	704.70	37.82	36.26	17.62	15.38	248.763	5.7372	0.270723	248.763	5.7372	0.270723	0.017612
224.7	666.80	35.07	36.34	18.12	16.10	227.526	-2.8262	0.148327	227.526	-2.8262	0.148327	0.001717
181.5	568.55	35.26	35.90	19.05	16.73	201.639	-20.1387	0.312686	201.639	-20.1387	0.312686	0.282191
227.5	653.10	35.56	31.84	16.51	10.58	225.158	2.3418	0.348858	225.158	2.3418	0.348858	0.004743
253.6	704.05	35.73	33.16	16.02	11.28	247.162	6.4381	0.176124	247.162	6.4381	0.176124	0.011306
263.0	709.60	36.46	33.83	15.89	11.91	257.490	5.5099	0.152345	257.490	5.5099	0.152345	0.006766
265.8	726.90	36.26	34.89	15.83	12.65	265.880	-0.0800	0.126199	265.880	-0.0800	0.126199	0.000001
263.8	697.15	37.20	36.27	16.71	14.06	259.684	4.1164	0.157438	259.684	4.1164	0.157438	0.003950



We can see that in residual vs fitted graph, there is some pattern that indicates a non linear association in the data.


Versus Order
(response is y)
Standardized Residual
Observation Order
4
3
2
1
0
-1
-2
2
4
6
8
10
12
14
16
18
20
22
24
26
28


Residuals Versus Leverage
(response is y)
Standardized Residual
Leverage
4
3
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

In the above graph we can see that, observations 1,4 and 24 have high leverage points and most influential points according to cook's distance

Problem 4:Project topic write-up (a short paragraph is sufficient)

For the Linear Regression Project, I will work on the Climate Change data available at Kaggle and draw insights from the statistical analysis performed. This project is collaborated with one of my classmates, Vaishnavi Solunke. There are many studies showing that the average temperature of the earth has risen over the last century. The consequences of a sustained rise in global temperatures will be catastrophic. Rising sea levels and increasing frequency of extreme weather events affect billions of people. Many corporate organizations are yet to see the impact of climate change at organizational level. That the impact of Climate change on Organization's Human Resource can affect their revenues and other metrics.

In this issue, we will examine the relationship between the average temperature of the earth and several other factors.This problem is an attempt to study the relationship between average global temperature and other factors.

Dataset link: https://www.kaggle.com/econdata/climate-change