Home Work 5-1

1. A chemical engineer studied the effect of the amount of surfactant (x_1) and time (x_2) on clathrate formation (y). Clathrates are used as cool storage media. File dat_Table_B8.xlsx summarizes the experimental results.

a. Fit a multiple linear regression model relating clathrate formation to these regressors.

Below are the output from R.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.0870	1.6692	6.64	0.0000
x1	350.1192	39.6829	8.82	0.0000
x2	0.1089	0.0100	10.91	0.0000

Residual standard error: 4.782 on 33 degrees of freedom

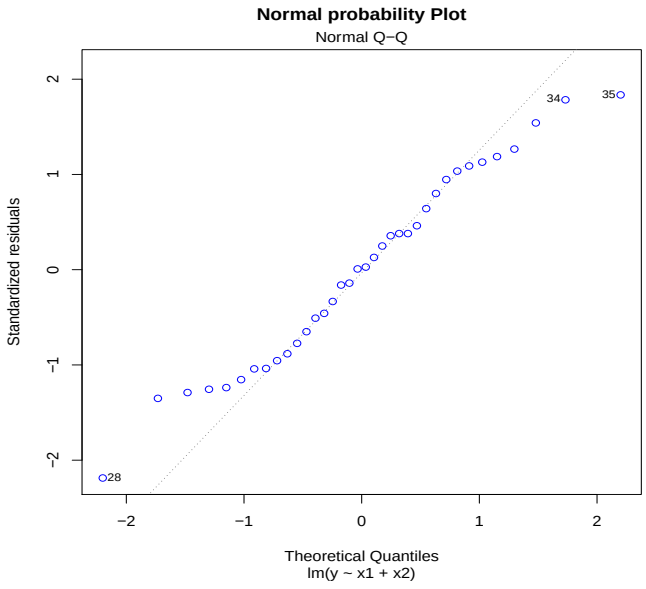
Multiple R-squared: 0.8415, Adjusted R-squared: 0.8319

F-statistic: 87.6 on 2 and 33 DF, p-value: 6.316e-14

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	1283.90	1283.90	56.14	0.0000
x2	1	2723.17	2723.17	119.07	0.0000
Residuals	33	754.74	22.87		

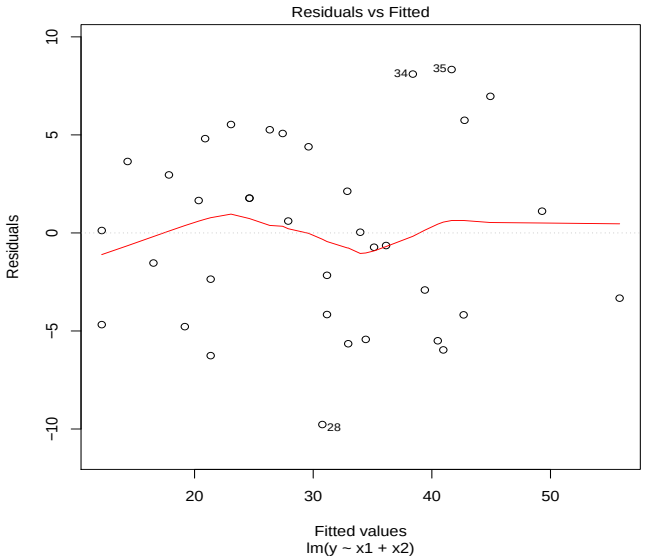
b. Construct a normality plot of the residuals from the full model. Does there seem to be any problem with the normality assumption?

This plot indicates that our data is partially normal with huge deviations at the end. Data seems to have a heavy-tailed distribution as Above the line is lower percentile and below the line is higher percentile. Nonlinearity might exist



c. Construct and interpret a plot of the residuals versus the predicted response.

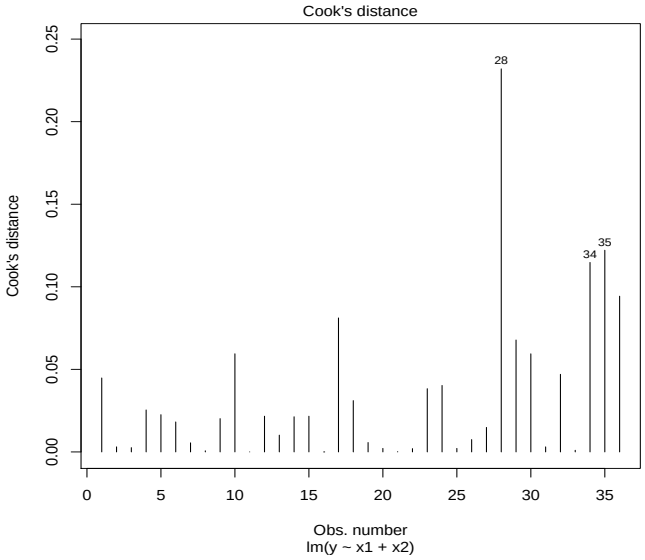
By visual inspection, 3 points seem to have high values but as it is not standardized it does not confirm that those are outliers.



d. Perform a thorough influence analysis (any influential points) of the clathrate formation model.

Plotting of Cook's D values indicates that 3 points are in question. Among them point 28 seems to be outlier.

Note: There are various rules about how to detect outliers based on Cook's D. According to the original paper the cutoff point should be F(α, p, n-p) and choice of α can be 0.5 or lower.



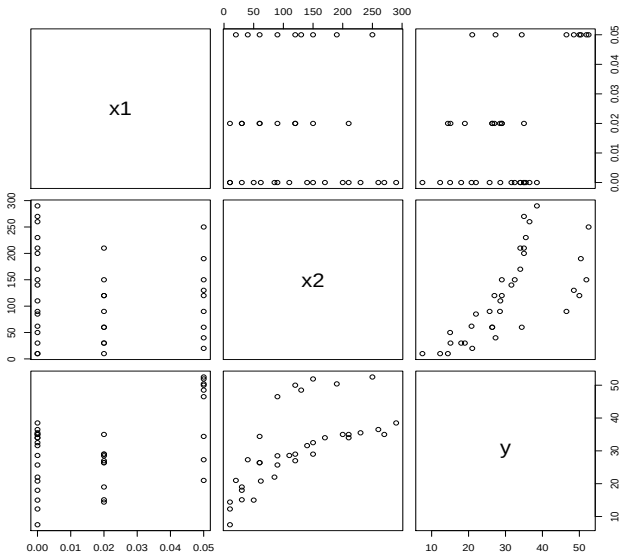
	Y	Residual	Stand_Res	Student_Res	R_Student	Lev_hii	CookD	Dffit
1	7.5000	-4.6763	-0.9778	-1.0371	-1.0383	0.1111	0.0448	-0.3670
2	15.0000	-1.5337	-0.3207	-0.3338	-0.3292	0.0767	0.0031	-0.0949
3	22.0000	1.6536	0.3458	0.3563	0.3515	0.0581	0.0026	0.0873
4	28.6000	5.5302	1.1564	1.1873	1.1949	0.0514	0.0254	0.2780
5	31.6000	5.2622	1.1003	1.1292	1.1341	0.0504	0.0226	0.2614
6	34.0000	4.3942	0.9188	0.9464	0.9448	0.0574	0.0182	0.2331
7	35.0000	2.1261	0.4446	0.4615	0.4560	0.0721	0.0055	0.1271
8	35.5000	-0.6419	-0.1342	-0.1411	-0.1390	0.0947	0.0007	-0.0450
9	36.5000	-2.9099	-0.6085	-0.6506	-0.6448	0.1252	0.0202	-0.2439
10	38.5000	-4.1780	-0.8736	-0.9552	-0.9539	0.1635	0.0594	-0.4217
11	12.3000	0.1237	0.0259	0.0274	0.0270	0.1111	0.0000	0.0095
12	18.0000	3.6450	0.7622	0.7999	0.7955	0.0922	0.0216	0.2534
13	20.8000	2.9591	0.6187	0.6413	0.6355	0.0691	0.0102	0.1732
14	25.7000	4.8089	1.0056	1.0351	1.0363	0.0563	0.0213	0.2532
15	32.5000	5.0729	1.0607	1.0894	1.0926	0.0519	0.0216	0.2555
16	34.0000	0.0368	0.0077	0.0080	0.0079	0.0788	0.0000	0.0023
17	35.0000	-5.4993	-1.1499	-1.2379	-1.2483	0.1371	0.0811	-0.4976
18	14.4000	-4.7787	-0.9992	-1.0413	-1.0427	0.0792	0.0311	-0.3058
19	19.0000	-2.3574	-0.4929	-0.5090	-0.5032	0.0621	0.0057	-0.1294
20	26.4000	1.7746	0.3711	0.3793	0.3743	0.0429	0.0021	0.0792
21	28.5000	0.6065	0.1268	0.1289	0.1269	0.0315	0.0002	0.0229
22	29.0000	-2.1615	-0.4520	-0.4584	-0.4529	0.0280	0.0020	-0.0769
23	35.0000	-5.9656	-1.2474	-1.2898	-1.3034	0.0646	0.0383	-0.3426
24	15.1000	-6.2574	-1.3084	-1.3510	-1.3688	0.0621	0.0403	-0.3521
25	26.4000	1.7746	0.3711	0.3793	0.3743	0.0429	0.0021	0.0792
26	27.0000	-4.1615	-0.8702	-0.8826	-0.8796	0.0280	0.0075	-0.1494
27	29.0000	-5.4295	-1.1353	-1.1542	-1.1602	0.0324	0.0149	-0.2123
28	21.0000	-9.7716	-2.0433	-2.1869	-2.3289	0.1270	0.2320	-0.8884
29	27.3000	-5.6503	-1.1815	-1.2554	-1.2669	0.1143	0.0678	-0.4551
30	48.5000	5.7456	1.2014	1.2664	1.2785	0.1000	0.0594	0.4263
31	50.4000	1.1095	0.2320	0.2487	0.2451	0.1298	0.0031	0.0947
32	52.5000	-3.3265	-0.6956	-0.7733	-0.7685	0.1908	0.0470	-0.3732
33	34.4000	-0.7290	-0.1524	-0.1611	-0.1587	0.1050	0.0010	-0.0544
34	46.5000	8.1030	1.6943	1.7837	1.8478	0.0977	0.1148	0.6079
35	50.0000	8.3349	1.7428	1.8352	1.9071	0.0981	0.1222	0.6291
36	51.9000	6.9669	1.4568	1.5411	1.5753	0.1065	0.0943	0.5437

By looking at the values of different kinds of residuals and previous plots, it can be concluded that 3 of the points require attention which are 28, 34 and 35. One of the ways to identify outliers is to compare $|t_i|$ with $t_{\alpha/2,n-p-1}$. But it is important to know that these considerations are simultaneous, so we need to look at it as multiple comparison (it is covered later). Here if we set $\alpha = 0.05$, as a result cutoff $= t_{0.05/2,36-2-1} = 2.03$. By that rule the point 28 should be considered outlier while other two may not be considered as outliers. Also Cook's D supports the decision.

Note: In this analysis, statistical test should be considered along with the graph and other values.

f. Identify any appropriate transformation for these data. Fit this model and compare.

In this multiple plot, it seems that there is a quadratic relationship between x_2 and y (i.e $y^2 \approx x_2$). Hence to try the model $y = \beta_0 + \beta_1.\sqrt{x_2} + e$



The summary of this new model is as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0508	1.9584	0.54	0.5952
x1	333.6476	32.6876	10.21	0.0000
sqrt(x2)	2.2928	0.1667	13.75	0.0000

Residual standard error: 3.957 on 33 degrees of freedom

Multiple R-squared: 0.8915, Adjusted R-squared: 0.8849

F-statistic: 135.6 on 2 and 33 DF, p-value: $<$ 2.2e-16

This shows improvement is R^2 which can be considered a better fit.

Note: Just better R^2 does not confirm that other statistics are better as well.

2. The kinematic viscosity (y) of a certain solvent system depends on the ratio of the two solvents (x_1) and the temperature (x_2). File data_Table_B10.xlsx summarizes a set of experimental results.

a. Fit a multiple linear regression model relating the viscosity to the two regressors.

Summary of the model is as follows:

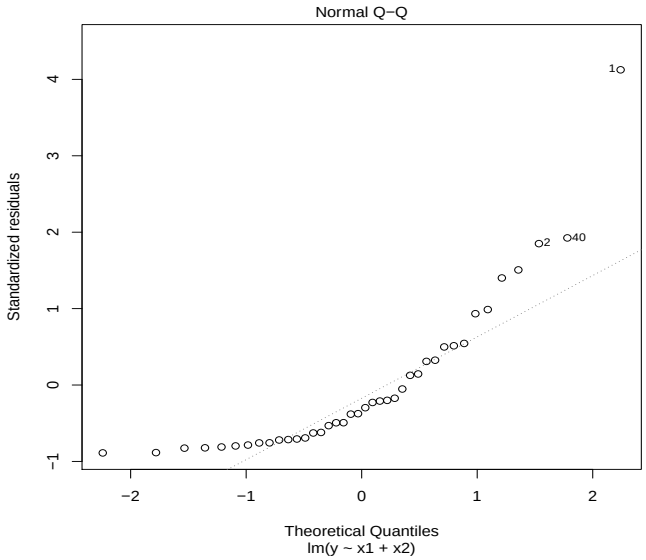
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6794	0.1435	4.73	0.0000
x1	1.4073	0.1969	7.15	0.0000
x2	-0.0156	0.0014	-10.95	0.0000

Residual standard error: 0.2593 on 37 degrees of freedom
Multiple R-squared: 0.822, Adjusted R-squared: 0.8124
F-statistic: 85.46 on 2 and 37 DF, p-value: 1.351e-14

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	3.43	3.43	51.07	0.0000
x2	1	8.06	8.06	119.85	0.0000
Residuals	37	2.49	0.07		

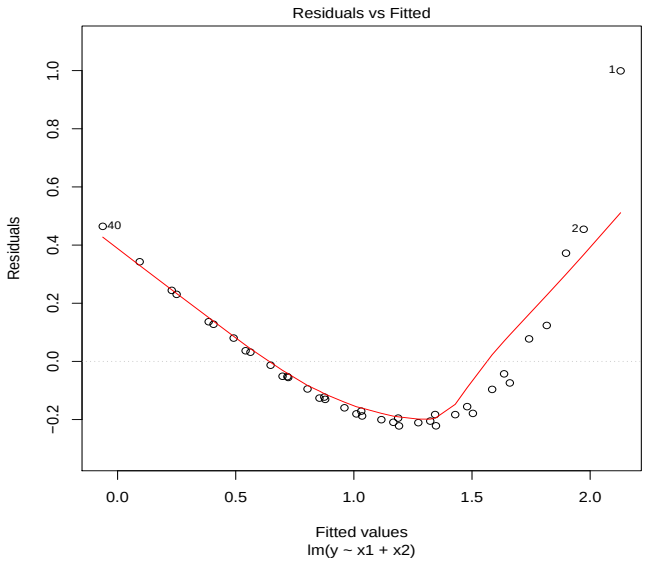
b. Construct a normality plot of the residuals from the full model. Does there seem to be any problem with the normality assumption?

This plot indicates that assumption of normality for this model is violated.



c. Construct and interpret a plot of the residuals versus the predicted response.

The graph clearly suggests the violation of linearity. But it can also happen due to any missing important independent variable as well.



d. Compute the PRESS statistic for the model and interpret it.

Computed PRESS statistic is 3.11. It implies that $R_p^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{3.11}{3.43+8.06+2.49} = 0.7775(i.e\,77.75\%)$

Using PRESS statistic, we can say that 77.75% of variation in viscosity can be explained by this model.

e. Perform a thorough influence analysis (any influential points) for the model.

	Y_Value	Lev_hii	CookD	Dffit	(Intercept)	x1	x2	I_CookD	I_Dffit	I_Dfbeta1	I_Dfbeta2
1	3.13	0.13	0.83	2.12	-0.03	0.18	-0.00	Inspect	Inspect	0	0
2	2.43	0.10	0.13	0.65	-0.02	0.08	-0.00	0	Inspect	0	0
3	1.94	0.09	0.01	0.15	-0.01	0.02	-0.00	0	0	0	0
4	1.59	0.07	0.00	-0.08	0.00	-0.01	0.00	0	0	0	0
5	1.32	0.07	0.01	-0.19	0.01	-0.03	0.00	0	0	0	0
6	1.13	0.07	0.02	-0.24	0.02	-0.04	-0.00	0	0	0	0
7	0.97	0.07	0.02	-0.25	0.02	-0.04	-0.00	0	0	0	0
8	0.85	0.09	0.02	-0.23	0.02	-0.03	-0.00	0	0	0	0
9	0.75	0.10	0.01	-0.18	0.02	-0.02	-0.00	0	0	0	0
10	0.67	0.13	0.00	-0.09	0.01	-0.01	-0.00	0	0	0	0
11	2.27	0.09	0.08	0.49	0.01	0.02	-0.00	0	0	0	0
12	1.82	0.07	0.00	0.08	0.00	0.00	-0.00	0	0	0	0
13	1.49	0.05	0.00	-0.09	-0.00	-0.01	0.00	0	0	0	0
14	1.25	0.04	0.01	-0.14	-0.00	-0.01	0.00	0	0	0	0
15	1.06	0.03	0.01	-0.15	0.00	-0.01	0.00	0	0	0	0
16	0.92	0.03	0.01	-0.14	0.00	-0.01	-0.00	0	0	0	0
17	0.80	0.04	0.01	-0.12	0.00	-0.01	-0.00	0	0	0	0
18	0.71	0.05	0.00	-0.09	0.00	-0.01	-0.00	0	0	0	0
19	0.63	0.07	0.00	-0.01	0.00	-0.00	-0.00	0	0	0	0
20	0.57	0.09	0.00	0.10	-0.01	0.01	0.00	0	0	0	0
21	1.59	0.09	0.00	-0.05	-0.00	0.00	0.00	0	0	0	0
22	1.32	0.07	0.01	-0.16	-0.02	0.01	0.00	0	0	0	0
23	1.12	0.05	0.01	-0.18	-0.02	0.01	0.00	0	0	0	0
24	0.96	0.04	0.01	-0.16	-0.02	0.01	0.00	0	0	0	0
25	0.83	0.03	0.01	-0.12	-0.01	0.01	0.00	0	0	0	0
26	0.73	0.03	0.00	-0.09	-0.01	0.01	-0.00	0	0	0	0
27	0.65	0.04	0.00	-0.04	-0.00	0.00	-0.00	0	0	0	0
28	0.58	0.05	0.00	0.03	0.00	-0.00	0.00	0	0	0	0
29	0.52	0.07	0.01	0.14	0.00	-0.01	0.00	0	0	0	0
30	0.47	0.09	0.03	0.31	0.00	-0.01	0.00	0	0	0	0
31	1.16	0.13	0.03	-0.30	-0.04	0.04	0.00	0	0	0	0
32	0.99	0.11	0.03	-0.28	-0.04	0.04	0.00	0	0	0	0
33	0.86	0.09	0.02	-0.22	-0.03	0.03	0.00	0	0	0	0
34	0.75	0.08	0.01	-0.14	-0.02	0.02	0.00	0	0	0	0
35	0.67	0.07	0.00	-0.06	-0.01	0.01	0.00	0	0	0	0
36	0.59	0.07	0.00	0.04	0.00	-0.01	0.00	0	0	0	0
37	0.53	0.08	0.01	0.15	0.02	-0.02	0.00	0	0	0	0
38	0.48	0.09	0.03	0.30	0.03	-0.04	0.00	0	0	0	0
39	0.44	0.11	0.08	0.50	0.04	-0.06	0.00	0	0	0	0
40	0.40	0.13	0.19	0.79	0.05	-0.09	0.00	0	Inspect	0	0

By looking at the influence analysis, it appears that point 1 seems to be an outlier. On the other hand point 2 and 40 are relatively high leverage points which does effect the dffit values. Hence those two are not considered as outliers.

f. Perform a thorough residual analysis of these data.

	Y_value	Residual	Stand_Res	Student_Res	R_Student	Lev_hii	CookD	Dffit
1	3.1280	0.9991	3.8524	4.1251	5.5366	0.1278	0.8312	2.1195
2	2.4270	0.4544	1.7520	1.8505	1.9161	0.1036	0.1319	0.6513
3	1.9400	0.1237	0.4768	0.4986	0.4934	0.0854	0.0077	0.1508
4	1.5860	-0.0741	-0.2856	-0.2966	-0.2930	0.0733	0.0023	-0.0824
5	1.3250	-0.1788	-0.6893	-0.7137	-0.7089	0.0672	0.0122	-0.1903
6	1.1260	-0.2215	-0.8540	-0.8843	-0.8816	0.0672	0.0188	-0.2366
7	0.9694	-0.2218	-0.8552	-0.8884	-0.8858	0.0733	0.0208	-0.2491
8	0.8473	-0.1876	-0.7234	-0.7564	-0.7520	0.0854	0.0178	-0.2298
9	0.7481	-0.1305	-0.5033	-0.5315	-0.5263	0.1036	0.0109	-0.1789
10	0.6671	-0.0552	-0.2130	-0.2280	-0.2251	0.1278	0.0025	-0.0862
11	2.2700	0.3722	1.4350	1.5065	1.5338	0.0926	0.0772	0.4899
12	1.8190	0.0774	0.2986	0.3094	0.3056	0.0683	0.0023	0.0828
13	1.4890	-0.0963	-0.3712	-0.3809	-0.3764	0.0502	0.0026	-0.0865
14	1.2460	-0.1830	-0.7055	-0.7194	-0.7146	0.0380	0.0068	-0.1421
15	1.0620	-0.2107	-0.8124	-0.8257	-0.8221	0.0320	0.0075	-0.1494
16	0.9160	-0.2004	-0.7727	-0.7854	-0.7812	0.0320	0.0068	-0.1420
17	0.8005	-0.1596	-0.6155	-0.6275	-0.6223	0.0380	0.0052	-0.1238
18	0.7091	-0.0947	-0.3652	-0.3748	-0.3704	0.0502	0.0025	-0.0851
19	0.6345	-0.0130	-0.0503	-0.0521	-0.0514	0.0683	0.0001	-0.0139
20	0.5715	0.0803	0.3095	0.3249	0.3209	0.0926	0.0036	0.1025
21	1.5930	-0.0428	-0.1650	-0.1730	-0.1707	0.0903	0.0010	-0.0538
22	1.3240	-0.1555	-0.5996	-0.6205	-0.6152	0.0660	0.0091	-0.1636
23	1.1180	-0.2052	-0.7913	-0.8110	-0.8071	0.0478	0.0110	-0.1809
24	0.9576	-0.2093	-0.8072	-0.8220	-0.8183	0.0357	0.0083	-0.1575
25	0.8302	-0.1804	-0.6958	-0.7063	-0.7015	0.0297	0.0051	-0.1226
26	0.7282	-0.1262	-0.4864	-0.4938	-0.4887	0.0297	0.0025	-0.0854
27	0.6470	-0.0511	-0.1969	-0.2005	-0.1979	0.0357	0.0005	-0.0381
28	0.5784	0.0366	0.1412	0.1447	0.1428	0.0478	0.0004	0.0320
29	0.5219	0.1364	0.5260	0.5443	0.5390	0.0660	0.0070	0.1433
30	0.4735	0.2443	0.9420	0.9877	0.9873	0.0903	0.0323	0.3110
31	1.1610	-0.1828	-0.7048	-0.7577	-0.7532	0.1348	0.0298	-0.2973
32	0.9925	-0.1950	-0.7519	-0.7972	-0.7932	0.1105	0.0263	-0.2796
33	0.8601	-0.1711	-0.6597	-0.6925	-0.6875	0.0924	0.0163	-0.2193
34	0.7523	-0.1226	-0.4728	-0.4930	-0.4879	0.0802	0.0071	-0.1441
35	0.6663	-0.0523	-0.2017	-0.2097	-0.2069	0.0742	0.0012	-0.0586
36	0.5940	0.0317	0.1221	0.1269	0.1252	0.0742	0.0004	0.0354
37	0.5338	0.1278	0.4926	0.5137	0.5085	0.0802	0.0077	0.1502
38	0.4804	0.2306	0.8894	0.9335	0.9319	0.0924	0.0296	0.2973
39	0.4361	0.3426	1.3212	1.4009	1.4200	0.1105	0.0813	0.5006
40	0.4016	0.4644	1.7908	1.9252	2.0020	0.1348	0.1925	0.7901

Similar to influence analysis, it indicates that point one is an outlier.