16.3.3 FINDING INFLUENTIAL POINTS IN A REGRESSION

Per 16.3.2, a point not belonging to the original regression may be outlying; notwithstanding, it cannot influence coefficients already regressed. But if a point in the original data regression is outlying, it +may indeed influence the coefficients, or it may not: perhaps it outlies in an uninfluential direction. Consider the data cluster bounded by the dashed circle in Figure 16.3-3, where the dashed line is the regression line for those data.

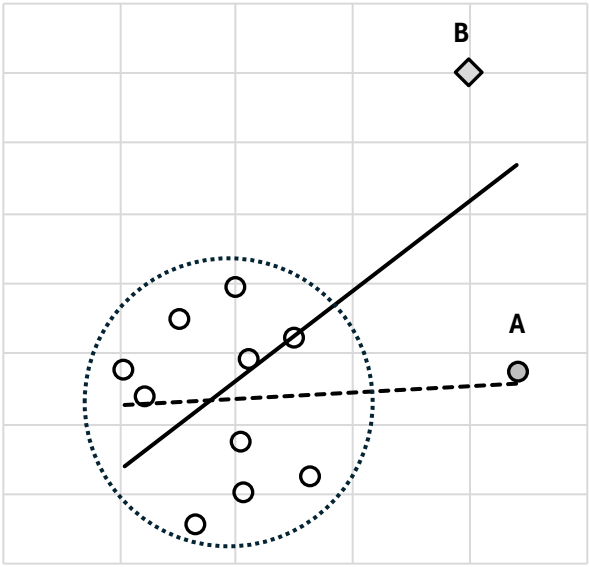

FIGURE 16.3-3 Outlying and influential points. Point A (gray circle) is outlying but not influential as it sits nearly on the regression line (dashed) defined by the remaining data (unfilled circles). By contrast, Point B (gray diamond) is both outlying and influential as it markedly changes the slope of the regression line when included (dotted regression line shifts to the solid line).

Point A is outlying, and its leverage would flag it as so. Nonetheless, if it were included in the regression, it would not have much influence because it very nearly sits on the (dashed) line already regressed. Thus, Point A is outlying but it is not *influential*. However, if Point B were included in the regression, the regression line would shift markedly toward it (solid line) because its leverage is in an influential direction. In short, Point A is outlying but not influential; Point B is both outlying and influential. Several statistics flag influential points.

16.3.4 Cook's Distance, DFFITS, and DFBETAS

As demonstrated, it is possible that a point is outlying but not in a way that significantly affects the estimated regressors. To see if this is so, one may remove the k^{th} point from the data set to see how it affects the regression. Such measures make use of the standardized

residual with the k^{th} observation omitted, $MSR_{(k)}$, where the parentheses indicate the deleted observation, per Equation 16.3-5.

$$MSR_{(k)} = \frac{\Sigma\left[y - \hat{y}_{(k)}\right]}{n - m - 1} = MSR(1 - h_{kk}) \tag{16.3-5}$$

Here, $\hat{y}_{(k)}$ is the predicted response when the k^{th} observation is omitted from the regressed model. Fortunately, the second equality shows that there is no need to perform n regressions (one for each deletion), as each $MSR_{(k)}$ derives from MSR and the k^{th} leverage value. Equation 16.3-6 gives the standard deviation for the regression with the k^{th} residual deleted, $s_{(k)}$, where $\epsilon_k = y_k - \hat{y}_k$, $SSR = \Sigma\epsilon_k^2$, and $DFR_{(k)} = DFR - 1 = n - m - 1$. Because an additional degree of freedom is removed with the omission of the k^{th} point, $s_{(k)}$ tends toward a $t(DFR - 1)$ distribution, or if squared, an $F(1, DFR - 1)$ distribution.

$$s_{(k)} = \epsilon_k \sqrt{\frac{DFR_{(k)}}{SSR(1 - h_{kk}) - \epsilon_k^2}} \sim t \qquad s_{(k)}^2 \sim F \tag{16.3-6a, b}$$

Three related measures of influence are in common use: *Cook's Distance* ($\Delta_{(k)}\bar{y}$), *DFFBETAS* ($\Delta_{j(k)}a$), and *DFFITS* ($\Delta_{(k)}\hat{y}$), per Equations 16.3-7, 8, and 9, respectively, where s_j is the standard error of the j^{th} effect per Equation 10.8-2.

$$Cook's\ D\text{:} \quad \Delta_{(k)}\bar{y} = \frac{\epsilon_k^2}{m \cdot MSR} \cdot \frac{h_{kk}}{(1 - h_{kk})^2} = \frac{\left(\mathbf{a} - \mathbf{a}_j\right)\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{a} - \mathbf{a}_j\right)}{m \cdot MSR} \sim \frac{1}{m} F \tag{16.3-7}$$

$$DFBETAS\text{:} \quad \Delta_{j(k)}a = \frac{a_j - a_{j(k)}}{s_{x,j}} = \frac{x_{k,j}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}{s_{x,j}}\left(\frac{\epsilon_k}{1 - h_{kk}}\right) \tag{16.3-8}$$

$$DFFITS\text{:} \quad \Delta_{(k)}\hat{y} = s_{(k)}\sqrt{\frac{h_{kk}}{1 - h_{kk}}} \sim t\sqrt{\frac{h_{kk}}{1 - h_{kk}}} \tag{16.3-9}$$

DFBETAS is so named because coefficients were originally denominated by β_j [7]; however, for consistency this text maintains a_j to indicate the j^{th} regressor. The second equality of Equation 16.3-7 shows that Cook's distance [8] is the joint confidence region for the regression model [9], while DFFITS concerns the influence of the k^{th} point, and DFBETAS concerns the influence of the k^{th} deletion on the j^{th} coefficient. Note that $\mathbf{\Delta A}$ is an $n \times m$ matrix, while $\mathbf{\Delta}\bar{\mathbf{y}}$ and $\mathbf{\Delta}\hat{\mathbf{y}}$ are $n \times 1$ vectors. As general rules of thumb [7 – 10], Equation 16.3-10a flags potential outliers, and Equations 16.3-10b, c, and d flag potentially influential outliers.

$$h_{kk} > 2\frac{m}{n} \qquad \Delta_{(k)}\bar{y} > \frac{4}{n} \qquad \left|\Delta_{j(k)}a\right| > \frac{2}{\sqrt{n}} \qquad \Delta_{(k)}\hat{y} > 2\sqrt{\frac{m}{n}} \tag{16.3-10a, b, c, d}$$

Note that $\mathbf{\Delta A/a}$ may be expressed as either the Δ% or fractional difference of the coefficients with and without deletion, which may be more informative and intuitive than $\mathbf{\Delta A}$ alone; accordingly, $\mathbf{\Delta A/a}$ may be tested against $\left|\Delta_{j(k)}a/a_j\right| > \left|2/a_j\sqrt{n}\right|$. For larger data sets ($n \gtrsim 20$), one may also make inference from the percentile distributions by using a second-order regression in log statistics against percentile and then inverting.

Example 16.3-1 Identifying Influential Points

Problem Statement: The data of Table 16.3-1 captures 11 flame length observations of a boiler as it ramps up in firing rate from 0 to 100% of full load. 1. Use Equations 16.3-1, 7, 8, 9, and 10 to flag outlying and influential points for further consideration. 2. Which point(s) are outlying? 3. Which are influential?

Table 16.3-1 Observations of Flame Length for a Boiler at Various Loads and Oxygen Levels in the Flue Gas

Pt	Load [%]	O_2 [%]	Flame L [ft]	Pt	Load [%]	O_2 [%]	Flame L [ft]	Pt	Load [%]	O_2 [%]	Flame L [ft]
1	8.6	11.0	1	5	43.5	4.9	15	9	86.3	4.0	13
2	10.8	8.9	5	6	54.9	3.7	15	10	96.7	3.8	15
3	21.6	5.9	5	7	64.9	3.2	15	11	100.0	4.0	16
4	32.5	5.3	10	8	75.5	3.5	10				

Solution: *See Figure 16.3-4*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		X Matrix			y	ŷ					DFBETAS, $[a - a_{(k)}]/s_k$			Coeff. with deletion		
2	Pt	ξ_0	Load	O2	Flame		ε =	h_{kk}	Cook's	DFFITS	$\Delta_{(0)}a$	$\Delta_{(1)}a$	$\Delta_{(2)}a$	$a_{(0)}$	$a_{(1)}$	$a_{(2)}$
3			[%], ξ_1	[%], ξ_2	L [ft]		y - ŷ		D, $\Delta\bar{y}$	$\Delta\hat{y}$						
4	1	1	8.6	11.0	1	1.3	-0.3	0.675	0.022	-0.240	0.134	-0.074	-0.196	14.8	0.048	-1.218
5	2	1	10.8	8.9	5	4.2	0.8	0.307	0.022	0.242	-0.021	-0.035	0.101	15.6	0.047	-1.383
6	3	1	21.6	5.9	5	8.6	-3.6	0.262	0.302	-1.079	-0.733	0.756	0.513	19.1	0.014	-1.610
7	4	1	32.5	5.3	10	9.9	0.1	0.204	0.000	0.017	0.014	-0.014	-0.011	15.4	0.046	-1.321
8	5	1	43.5	4.9	15	10.9	4.1	0.147	0.158	0.796	0.501	-0.415	-0.384	13.0	0.063	-1.114
9	6	1	54.9	3.7	15	13.1	1.9	0.196	0.055	0.396	0.323	-0.231	-0.296	13.9	0.055	-1.163
10	7	1	64.9	3.2	15	14.2	0.8	0.194	0.010	0.161	0.120	-0.070	-0.118	14.9	0.048	-1.261
11	8	1	75.5	3.5	10	14.3	-4.3	0.144	0.170	-0.848	-0.183	-0.068	0.207	16.4	0.048	-1.441
12	9	1	86.3	4.0	13	14.1	-1.1	0.201	0.018	-0.218	0.081	-0.148	-0.066	15.1	0.052	-1.290
13	10	1	96.7	3.8	15	14.8	0.2	0.301	0.001	0.050	-0.026	0.040	0.021	15.6	0.044	-1.338
14	11	1	100.0	4.0	16	14.7	1.3	0.370	0.074	0.453	-0.275	0.390	0.232	16.8	0.029	-1.455
15		X^TX			X^Ty			Critical Values						$a_{(k)}$		
16		11	595.3	58.2	120			0.5	0.4	1.0	0.6	0.6	0.6	15.5	0.045	-1.327
17		595.3	43331	2491	7873			Analysis of Variance (ANOVA)								
18		58.2	2491	369.7	523				SS	DF	MS	F	p	ȳ	10.909	
19		$(X^TX)^{-1}$			a	S.E.		M	211.08	2	105.5	15.12	0.002	m	3	
20		3.520	-0.027	-0.373	15.471	4.957		R	55.83	8	6.98	s=	2.642	n	11	
21		-0.027	0.000	0.003	0.045	0.041		T	266.91	10	26.69	R^2 =	0.791			
22		-0.373	0.003	0.044	-1.327	0.553										

Figure 16.3-4. Spreadsheet for Outliers. Cells B4:D14 comprise **X**, Cells E4:E14 comprise **y**. Cells F4:F14 comprise $\hat{\mathbf{y}}$, derived in the usual way, with the error reported in Cells G4:G14. Cell J20 contains the DFR and Cell K20 contains the MSR. These values lead to the statistics in Cells H4:M14 per Equations 16.3-1, 7, 8, and 9. Using Equation 16.3-8 to solve for $a_{j(k)}$ leads to the values of Cells N4:P14, which are the a coefficients with the k^{th} observation deleted. Equation 16.3-10 leads to the critical values held in Cells H16:M16. For comparison with $a^{\mathrm{T}}_{j(k)}$, Cells N16:P16 show $\mathbf{a}^{\mathrm{T}}$. As examples, for m=3, H4 = MMULT (MMULT (B4:D4,$B$20:$D$22), TRANSPOSE (B4:D4)), I4=G4^2/$K$20*(H4/(1-H4)^2)*1/m, J4=G4*SQRT((H4/(1-H4)*($J$20-1) / ($K$20*(1-H4)*$J$20-G4^2))), K4:M4=MMULT (B4:D4,$B$20:$D$22) / TRANSPOSE(F$20:F$22) * G4/(1-H4), and N4:P4=TRANSPOSE ($E$20:$E$22) - (K4:M4) * TRANSPOSE($F$20:$F$22).

1. Statistics that exceed the critical values of Cells H16:M16 are boxed and shaded.
2. For Pt 1, Cell H4 exceeds the threshold of Cell H16 and is therefore outlying. However, Cells I4:M4 compared to respective Cells I16:M16 show that Pt 1 is not influential.
3. Pt 3 is influential as indicated by the values of Cells I4:M4, all of which exceed the respective thresholds in Cells I16:M16. Also compare Cells N6:O6 with N16:O16, which show large coefficient differences between $a_{0(3)} = 19.1$ and $a_{1(3)} = 0.014$ versus $a_0 = 15.5$ and $a_1 = 0.045$, indicating strong influence of Point 3 on the flame length and load coefficient.

Discovery of influential points should encourage the investigator to further examine the data, not merely exclude them without further analysis. Some possibilities for outlying or influential points are serial correlations, errors in transcription, wayward measurement equipment, etc. Although randomization can mute serial correlation, it is often infeasible to randomize the firing rate of a boiler even for investigatory purposes, since so many downstream processes depend on the boiler's steam production and rate. In the present case, the boiler was fired from a cold start and gradually ramped up in firing rate from partial to full load. Yet only after a boiler is equilibrated (warm) do observations become reliable. Note the inverse relation of oxygen and firing rate for less than 33% load (typical).

If randomization is infeasible, a better run order for flame length and combustion-related emissions in fired equipment may be from high- to low-fire. This will at least ameliorate the influences of a cold boiler on the observations. In the present case, Points 1–4 are suspect. Flame length at partial load may not be a concern. Regardless, these are the data.

Importantly, outliers may in fact be valid data points representing unrealized but potentially profitable effects. In the present case, they have alerted the investigator to the perils of data collected near start-up. In all cases, the best policy is to flag influential outliers and investigate them carefully. Great discoveries may await.

REFERENCE

6. See https://real-statistics.com/multiple-regression/outliers-and-influencers/, last accessed 2 June 2025.
7. Belsley, D.A. et al, Chapter 2, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley & Sons, 1980. pp. 11-16.
8. Cook, R.D., *Detection of Influential Observations in Linear Regression*. Technical Report# 256, Department of Applied Statistics, University of Minnesota, St. Paul. December 15, 1975. https://conservancy.umn.edu/server/api/core/bitstreams/ee41cf4b-7ba7-4695-916b-7da1d6e43b25/content, last accessed 6 June 2025.
9. Neter, J. et al, Appendix C.13. Building the Regression Model II: Diagnostics in *Applied Linear Statistical Models*, 4th ed. McGraw-Hill, 1974. pp. 378-383.
10. Montgomery, D. et al, Appendix C. Computation of Influence Diagnostics in *Introduction to Linear Regression Analysis*, 5th ed. John Wiley & Sons, 2012. pp. 600-601.