

Outliers and leverage: finding the observations that **don't fit**

An observation can be unusual on one variable, or only in combination: big but understaffed, cheap but high-margin. XL Edge measures how far each row sits from the rest, all variables at once, and pairs that with the regression diagnostics that tell you whether an unusual row actually changes your answer.

Functions	<code>fx.RiskOutliersλ</code> , with <code>fx.RiskRegressResidsλ</code> and <code>fx.RiskEllipseλ</code>
In XL Edge	XL Edge tab > Productivity Tools > Stat Tools > Insert Statistical Diagnostics > Outliers and Leverage
Example	<i>XL Edge Example - Outliers and Leverage.xlsx</i> (same folder). Every number in this guide is read from it.
Method	Mahalanobis distance with the sample covariance; hat-matrix leverage; the Belsley-Kuh-Welsch $2(k + 1)/n$ rule.

- | | |
|------------------------------------|---|
| 1 The idea: unusual in combination | 6 Does it matter? Residuals and influence |
| 2 Using the ribbon | 7 With and without |
| 3 The example: forty stores | 8 Practical advice |
| 4 Reading the block | 9 Argument reference |
| 5 Which cut-off? | |

1 The idea: unusual in combination

The usual outlier check looks at one column at a time: anything more than two or three standard deviations from the average is suspicious. That misses the rows that are ordinary on every variable and odd only in how the variables go together. When floor area and staff rise together, a large store with a small team stands out, even though neither its size nor its headcount is extreme on its own.

The **Mahalanobis distance** measures exactly that. It is the distance from the average row, in standard deviations, after allowing for the spread of each variable and the correlation between them. Picture the data as an oval cloud: points along the cloud's long axis can sit far from the center and still be typical,

while a point just off the side of a narrow cloud is very unusual. On a chart, rows with the same distance lie on the same ellipse around the center.

Leverage is the regression view of the same thing: how much pull a row *could* have on a fitted line because of where its explanatory values sit. A row with high leverage is not necessarily a problem. It is a problem when it is also off the line, and the regression diagnostics in section 6 tell those cases apart.

2 Using the ribbon

XL Edge tab > Productivity Tools > Stat Tools (arrow) > Insert Statistical Diagnostics
> Outliers and Leverage

- 1 **Select the data:** a table with one variable per column (headers above are fine), or Ctrl-click the spilled results of several simulated variables.
- 2 **Pick an empty cell.** XL Edge states the size first, (n + 2) rows by 5 columns, one row per observation, and asks before writing over anything. Above 5,000 rows it checks you really want one row each.

The result is one live formula. For the regression view, use *Insert Statistical Diagnostics > Regression Residuals* on the same sheet: it asks for Y (the variable you are explaining), then the X columns.

3 The example: forty stores

Forty stores with floor area (thousand sq ft), staff (full-time equivalents) and monthly sales (\$ thousand), simulated so that bigger stores have more staff and sell more. Four stores were planted to show the four kinds of unusual row:

Store	What is unusual	Kind
S07	Very large and fully staffed; sales on the trend.	Extreme, but consistent ("good leverage")
S15	Fairly large but understaffed; sales on the trend.	Unusual only in combination
S23	A typical store that sells far less than its size and staff predict.	Off the line, but no pull ("outlier")
S31	Large and well staffed, and sells far less than predicted.	Off the line, with pull ("influential")

4 Reading the block

```
floor area and staff, the two explanatory variables  
=fx.RiskOutliersλ(Data!B5:C44)
```

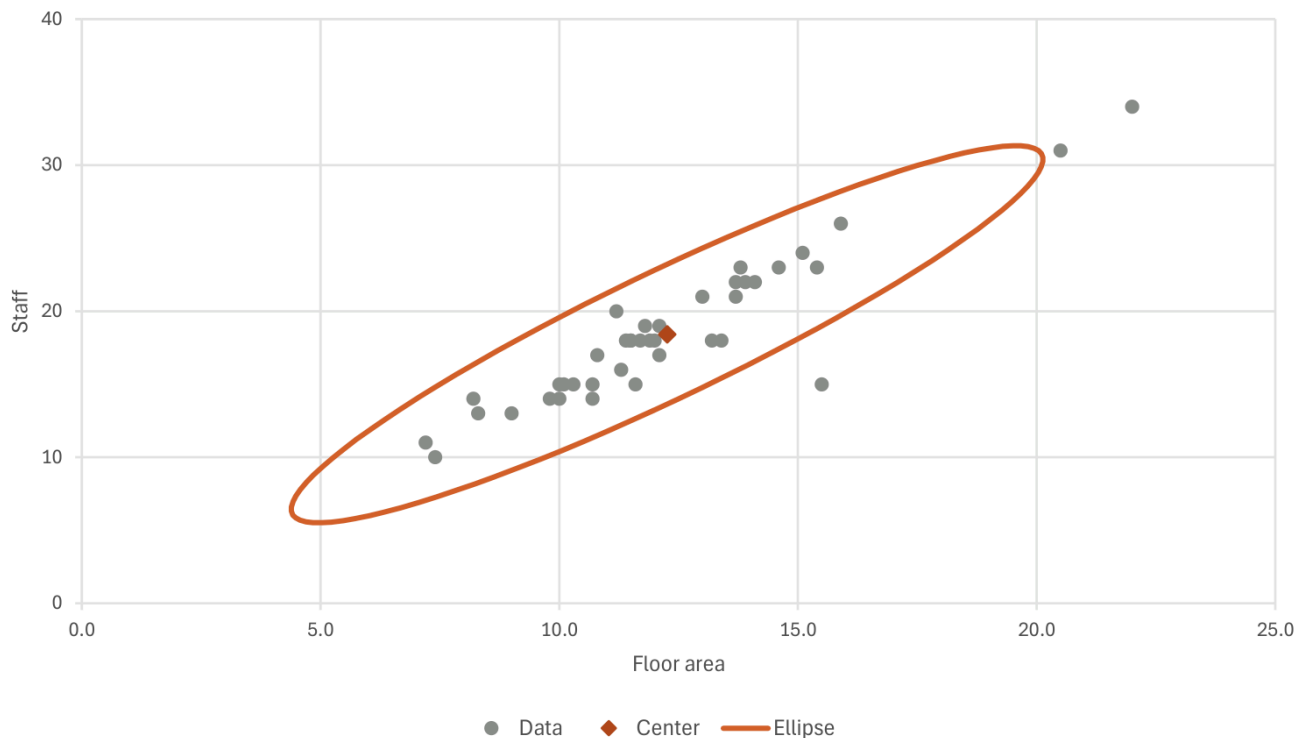
One row per store: **Observation** (the row number), **Distance** (Mahalanobis, in standard deviations), **Distance flag**, **Leverage** and **Leverage flag**. Beside the block, the example adds each variable's own z-score (sheet 1 *Outliers*), so the two views can be compared:

Sheet 1 Outliers

Store	z, floor area	z, staff	Distance	Leverage
S07	3.23	3.15	3.26	0.297
S15	1.08	-0.69	4.68	0.587
S23	-0.09	-0.09	0.09	0.025
S31	2.73	2.55	2.73	0.217

S15 is the lesson. Its floor area is 1.08 standard deviations above average and its staff -0.69: unremarkable on each, so a column-by-column check passes it. Yet its distance, 4.68, is the largest of all forty stores, because stores of its size normally have far more staff. S07 is extreme on both variables, but in the usual proportion, so its distance (3.26) is smaller than S15's. S23 sits right at the center (0.09): nothing about its size or staff is unusual. Its problem is its sales, which this function does not look at. Section 6 finds it.

95% prediction ellipse: Staff vs Floor area



The same picture, drawn with XL Edge's **Confidence Ellipse chart** (sheet 1 Outliers). The ellipse holds 95% of stores like these. S15 sits just below it at (15.5, 15), inside the range of both axes but outside the cloud; S07 and S31 sit beyond its top-right end. The ellipse's radius, 2.61, is a Mahalanobis distance: a store is outside exactly when its Distance exceeds it.

LEVERAGE AND DISTANCE ARE THE SAME INFORMATION

With the explanatory variables alone, leverage is a fixed transformation of the distance: $\text{leverage} = 1/n + \text{distance}^2 / (n - 1)$. The check cell on sheet 1 Outliers confirms it (largest difference 1.1e-16). So the two flags always rank the rows in the same order; they differ only in where they draw the line, which is the subject of the next section.

5 Which cut-off?

No cut-off is "the" right one. The block uses two, and the example adds a third:

Rule	Cut-off here	Stores flagged	What it means
Distance flag: mean distance + 1 standard deviation	1.97	3	Relative: the rows furthest out <i>in this data set</i> . It flags something in almost any data, even clean data.
Leverage flag: $2(k + 1) / n$	0.15	3	Belsley, Kuh and Welsch's rule: twice the average leverage. Equivalent here to a distance above 2.21.
Chi-square: distance ² above the 97.5th percentile of $\chi^2(k)$	2.72	3	For roughly normal data, only 2.5% of rows should cross it by chance (one store in forty).

the chi-square flag beside the block (k = 2 variables)
`=IF(B8:B47^2 >CHISQ.INV(0.975, 2), "High", "")`

Here all three agree on the same 3 stores: S07, S15 and S31. In real data they often disagree at the margin. Treat any flag as "look at this row", not "delete this row".

MASKING: OUTLIERS HIDE EACH OTHER

The distance is measured from the average, using a covariance that the outliers themselves help estimate. A cluster of several bad rows drags the average toward itself and inflates the spread, so each looks less unusual than it is. If you remove one flagged row and recompute, others can appear. Check again after any change.

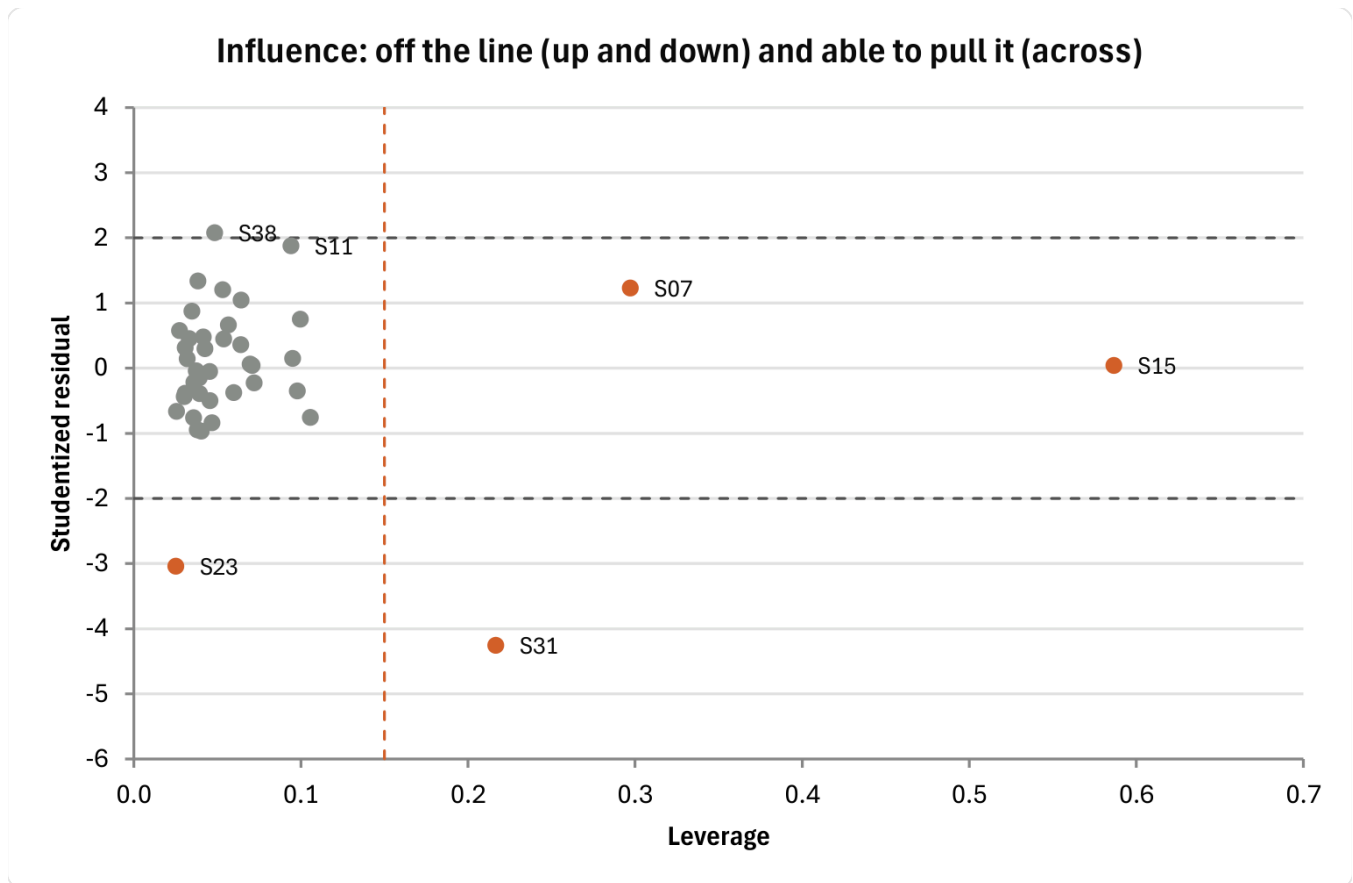
6 Does it matter? Residuals and influence

Unusual explanatory values matter only if you fit something to them. Regressing sales on floor area and staff, the regression diagnostics ask the two remaining questions: is the store *off the line* (its residual), and *did* it move the line (Cook's distance)?

Y = sales; X = floor area and staff; names from the header row
`=fx.RiskRegressResids(Data!D5:D44, Data!B5:C44, , HSTACK(Data!D4, Data!B4:C4))`

Sheet 2 *Regression*. Flags: Outlier |studentized| > 2; Leverage > $2m/n = 0.15$; Influential: Cook's distance > $4/n = 0.10$

Store	Residual (\$k)	Studentized	Leverage	Cook's distance	Flag
S07	50	1.23	0.297	0.210	Influential, Leverage
S15	1	0.04	0.587	0.001	Leverage
S23	-133	-3.04	0.025	0.065	Outlier
S31	-152	-4.26	0.217	1.142	Outlier, Influential, Leverage
S11	85	1.88	0.094	0.114	Influential
S38	95	2.08	0.048	0.067	Outlier



An influence plot (sheet 2 *Regression*, an ordinary Excel chart of the block's Leverage and Studentized columns). Up and down: how far off the line. Across: how much pull. Trouble lives in the corners away from the origin.

Reading it, store by store.

- **S15** has the most leverage of any store (0.587) but sits on the line (studentized 0.04), so its Cook's distance is 0.001: it could pull the line and doesn't. Keep it; it actually sharpens the estimates.
- **S07** also has high leverage and a modest residual. Its Cook's distance, 0.210, crosses the screening rule of $4/n$, which is why it is flagged Influential, but it is far below the stricter rule of 1. A big store on the trend is information, not a problem.
- **S23** is a clear outlier (studentized -3.04 : it sells about \$133k a month less than predicted) but sits at the center, so it barely moves the line (Cook's 0.065). Investigate the store; the regression is fine either way.
- **S31** is the dangerous one: far off the line (-4.26) *and* high leverage (0.217), so its Cook's distance, 1.142, is above 1. It is pulling the fitted line toward itself.
- **S11 and S38** are ordinary stores that cross a rule by chance. With forty stores, about two should exceed $|\text{studentized}| > 2$ even when nothing is wrong. That is why flags are prompts to look, not verdicts.

7 With and without

The direct test of influence is to refit without the row and see what changes. Sheet 3 *With and without* does it with `LINEST` and `FILTER`:

LINEST returns staff, floor area, intercept (right to left)
`=LINEST(FILTER(Data!D5:D44, Data!A5:A44 <> "S31"), FILTER(Data!B5:C44, Data!A5:A44 <> "S31"))`

Fitted on	Floor area	Staff	Intercept
All 40 stores	36.33	7.74	35.6
Without S07	34.99	7.33	57.7
Without S15	36.02	7.92	35.9
Without S23	36.28	7.71	40.2
Without S31	40.78	7.79	-15.0

Dropping S31 moves the floor-area coefficient from 36.33 to 40.78 and the intercept from 35.6 to -15.0. No other store moves the floor-area coefficient by more than 1.34. That is what Cook's distance predicted. Note also what it does *not* say: the data were simulated with a true floor-area effect of 38, and the refit without S31 overshoots it. Deleting a row is a sensitivity test, not a correction.

8 Practical advice

- **Investigate before you exclude.** The common causes are a data error (a misplaced decimal, a wrong unit), a row from a different population (a flagship store, a one-off contract), or a real event the model should learn from. Fix the first, model or separate the second, keep the third.
- **Report both ways.** If a conclusion changes when one row goes, say so; that fragility is a finding.
- **Units don't matter.** The distance allows for each variable's spread, so rescaling a column (dollars to thousands) leaves every distance unchanged.
- **More variables, more leverage.** The average leverage is $(k + 1)/n$, so with many columns and few rows every row looks influential. Keep n well above k (the function needs $n > k + 1$).
- **Simulated inputs too.** On Monte Carlo trials (Ctrl-click the spills), the distance finds the trials where inputs combine unusually: useful for checking that a copula produces the joint extremes you intended.

9 Argument reference

fx.RiskOutliersλ(Data)

Argument	Default	Meaning
Data	required	$n \times k$, one variable per column, $n > k + 1$. Numbers only.
Output, $(n + 2) \times 5$: a title; Observation, Distance, Distance flag, Leverage, Leverage flag. Distance = $\sqrt{((x - \bar{x})^T S^{-1} (x - \bar{x}))}$ with the sample covariance S ; flag "High" above the mean distance + 1 SD. Leverage = $1/n + \text{distance}^2/(n - 1)$, the hat-matrix diagonal of a regression on these columns with an intercept; flag "High" above $2(k + 1)/n$.		

fx.RiskRegressResidsλ(Y, X, [Intercept], [Names])

Argument	Default	Meaning
Y	required	$n \times 1$, the variable explained.
X	required	$n \times k$, the explanatory variables.
Intercept	TRUE	FALSE fits through the origin.
Names	-	Y's name, then one per column of X.

Output, $(n + 2) \times 9$: Observation, Actual, Fitted, Residual, Standardized, Studentized (externally; follows t with $n - m - 1$ df), Leverage, Cook's distance, Flag (Outlier |studentized| > 2 ; Influential, Cook $> 4/n$; Leverage $> 2m/n$, m = number of coefficients).

For the ellipse, see the *Confidence Ellipse* guide.

References: Mahalanobis (1936), *Proceedings of the National Institute of Sciences of India* 2(1); Belsley, Kuh and Welsch (1980), *Regression Diagnostics*, Wiley; Cook (1977), *Technometrics* 19(1). The example data are simulated for illustration. XL Edge and the fx.Risk... functions: © 2026 W Phillips, edgewisedata.com, MIT License.