Outliers

Author: John M. Cimbala, Penn State University
Latest revision: 12 September 2011

Introduction

- ***Outliers*** are defined as *data points that are statistically inconsistent with the rest of the data.*
- We must be careful because some "questionable" data points end up being outliers, but others do not.
- ***Questionable data points should never be discarded without proper statistical justification***.
- Here, we discuss statistical methods that help us to know whether to *keep* or *discard* suspected outliers.
- We discuss two types of outliers: (1) outliers in a sample of a *single variable* (x), and (2) outliers in a set of *data pairs* (y vs. x).

Outliers in a sample of a single variable

- Consider a sample of n measurements of a *single* variable x, i.e., x_1, x_2, ... , x_n. It is helpful to arrange the x values in increasing order, so that the suspected outliers are easily spotted (typically either the first or last data point(s) are suspected, since they are the lowest and highest values of x in the sample, respectively).
- The ***modified Thompson tau technique*** is a *statistical method for deciding whether to keep or discard suspected outliers in a sample of a single variable.* Here is the procedure:
  - The sample mean $\bar{x}$ and the sample standard deviation S are calculated in the usual fashion.
  - For each data point, the ***absolute value of the deviation*** is calculated as $\delta_i = |d_i| = |x_i - \bar{x}|$.
  - The data point *most suspected* as a possible outlier is *the data point with the maximum value of* δ_i.
  - The value of ***the modified Thompson*** τ (Greek letter tau) is calculated from the critical value of the student's t PDF, and is therefore a function of the number of data points n in the sample.
  - τ is obtained from the expression $\tau = \dfrac{t_{\alpha/2} \cdot (n-1)}{\sqrt{n}\sqrt{n-2+t_{\alpha/2}^{\,2}}}$, where
    - n is the number of data points
    - $t_{\alpha/2}$ is the critical student's t value, based on $\alpha = 0.05$ and df = n-2 (note that here df = n-2 instead of n-1). In Excel, we calculate $t_{\alpha/2}$ as TINV(α, df), i.e., here $t_{\alpha/2}$ = TINV(α, n-2)
  - A table of the modified Thompson τ is provided below:

Values of the Modified Thompson τ

n	τ	n	τ	n	τ
3	1.1511	21	1.8891	40	1.9240
4	1.4250	22	1.8926	42	1.9257
5	1.5712	23	1.8957	44	1.9273
6	1.6563	24	1.8985	46	1.9288
7	1.7110	25	1.9011	48	1.9301
8	1.7491	26	1.9035	50	1.9314
9	1.7770	27	1.9057	55	1.9340
10	1.7984	28	1.9078	60	1.9362
11	1.8153	29	1.9096	65	1.9381
12	1.8290	30	1.9114	70	1.9397
13	1.8403	31	1.9130	80	1.9423
14	1.8498	32	1.9146	90	1.9443
15	1.8579	33	1.9160	100	1.9459
16	1.8649	34	1.9174	200	1.9530
17	1.8710	35	1.9186	500	1.9572
18	1.8764	36	1.9198	1000	1.9586
19	1.8811	37	1.9209	5000	1.9597
20	1.8853	38	1.9220	($\rightarrow \infty$)	1.9600

  - We determine whether to reject or keep this suspected outlier, using the following simple rules:
    - **If $\delta_i > \tau S$, *reject* the data point**. It *is* an outlier.
    - **If $\delta_i \le \tau S$, *keep* the data point**. It is *not* an outlier.
  - With the modified Thompson τ technique, ***we consider only one suspected outlier at a time*** – namely, **the data point with the largest value of δ_i**. If that data point is rejected as an outlier, we remove it and start over. In other words, we calculate a *new* sample mean and a *new* sample standard deviation, and search for more outliers. This process is repeated ***one at a time*** until no more outliers are found.

- **Example**:
  Given: Ten values of variable *x* are measured. The data have been arranged in increasing order for convenience: 48.9, 49.2, 49.2, 49.3, 49.3, 49.8, 49.9, 50.1, 50.2, and 50.5.
  To do: Apply the modified Thompson tau test to see if any data points can be rejected as outliers.
  Solution:
  - The number of data points is counted: $n = 10$.
  - You suspect the first (smallest *x*) and last (largest *x*) data points to be possible outliers.
  - The sample mean and sample standard deviation are calculated: $\bar{x} = 49.64$, and $S = 0.52957$.
  - For the smallest (first) *x* value, the absolute value of the deviation is $\delta_1 = |x_1 - \bar{x}| = |48.9 - 49.64| = 0.74$. [We keep one extra digit here to avoid round-off error in the remaining calculations.]
  - Similarly, for the largest (n^{th}) *x* value, the absolute value of the deviation is $\delta_{10} = |x_{10} - \bar{x}| = |50.5 - 49.64| = 0.86$.
  - Since the absolute value of the deviation of the largest data point is bigger than that of the smallest data point, the modified Thompson tau technique is applied *only* to data point number 10. (Point number 10 is the most suspect data point.)
  - For $n = 10$, the value of the modified Thompson τ is 1.7984 (see the above table). We calculate $\tau S = (1.7984)\cdot(0.52957) = 0.95238$.
  - Since $\delta_{10} = 0.86 < \tau S = 0.95238$, ***there are no outlier points***. All ten data points must be kept.

  Discussion: Since the tenth data point had the largest value of δ, there is no need to test any other data point – we are sure that there are no outliers in this sample.
- If an outlier point is identified, it can be removed without guilt. (The engineer is not "fudging" the data.)
- To identify any *further* suspected outliers, however, the mean and standard deviation must be re-calculated, and *n* must be decreased by 1, before re-applying the modified Thompson tau technique.
- This process should be continued again and again as necessary until no further outlier points are found.

Outliers in a set of data pairs

- Now consider a ***set of n data pairs*** (*y* vs. *x*), for which we have used regression analysis to find the best-fit line or curve through the data. The best-fit line or curve is denoted by *Y* as a function of *x*.
- In regression analysis, we defined the ***error*** or ***residual*** e_i for each data pair (x_i, y_i), as ***the difference between the predicted or fitted value and the actual value***: e_i = error or residual at data pair *i*, or $e_i = Y_i - y_i$.
- As mentioned in our discussion of regression analysis, a useful measure of overall error is called the ***standard error of estimate*** or simply the ***standard error***, $S_{y,x}$. In general, for regression analysis with *m* independent variables (and so also for a polynomial fit of order *m*), $S_{y,x}$ is defined by $S_{y,x} = \sqrt{\dfrac{\sum_{i=1}^{i=n}(y_i - Y_i)^2}{\text{df}}}$, where df = ***degrees of freedom***, $\text{df} = n - (m+1)$.
- When we perform a regression analysis in Excel, $S_{y,x}$ is calculated for us as part of Excel's **Summary Output**. $S_{y,x}$ has the same units and dimensions as variable *y*.
- To determine if there are any outliers, we calculate the ***standardized residual*** $e_i / S_{y,x}$ for each data pair.
- Since $S_{y,x}$ is a kind of standard deviation of the original data compared to the predicted least-squares fit values, we would expect that approximately 95% of the standardized residuals should lie between –2 and 2, in other words, within ±2 standard deviations from the best-fit curve, assuming random errors.
- So, to determine if a particular (*x*,*y*) data point is an outlier, we check if $|e_i / S_{y,x}| > 2$. If so, we suspect that data point to be an outlier. But wait! We are not finished yet.
- Unfortunately, this is not the *only* test for outliers. There is another criterion for calling the data point an outlier – namely, *the standardized residual must be inconsistent with its neighbors*.
- Bottom line: There are ***two criteria*** for a data pair to be called an outlier:
  - $|e_i / S_{y,x}| > 2$
  - The standardized residual is inconsistent with its neighbors when plotted as a function of *x*.
- If *either* of these criteria fail, we *cannot* claim this data point to be an outlier.
- The second criterion (inconsistency) is rather subjective, and is best illustrated by examples.

- **Example**:

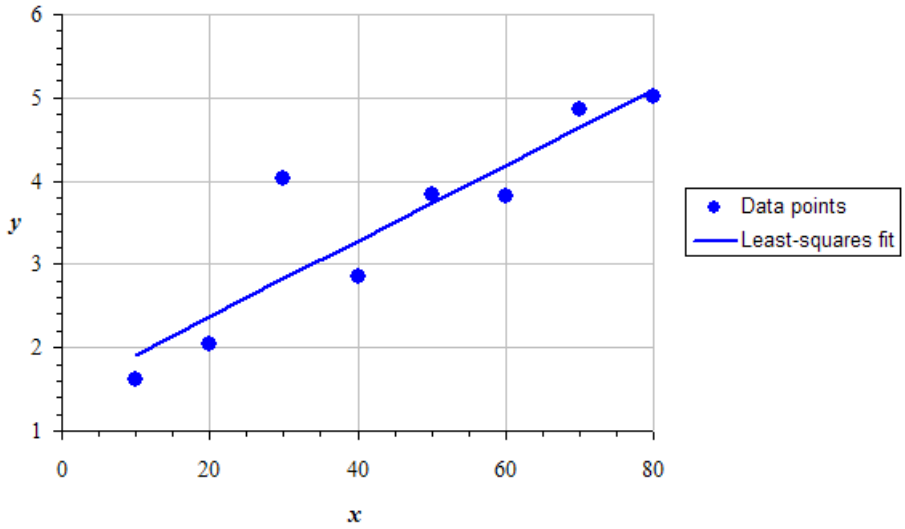
Given: Eight data pairs (x,y) are measured in a calibration experiment, as shown below:

i	*x*	*y*
1	10.00	1.62
2	20.03	2.04
3	30.01	4.03
4	40.02	2.85
5	50.02	3.84
6	60.01	3.81
7	70.00	4.86
8	80.01	5.02

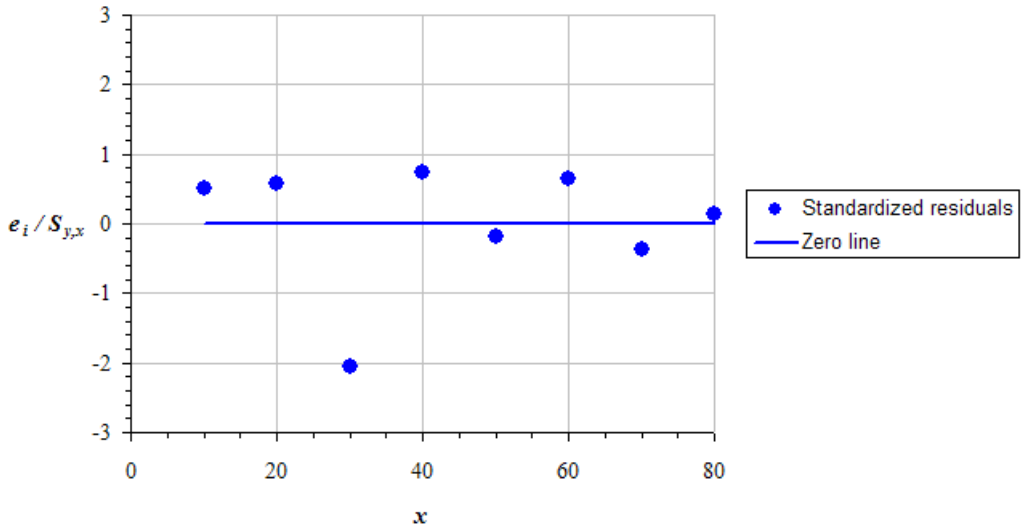
To do: Fit a straight line through these data, and determine if any of the data points should be thrown out as legitimate outliers.

Solution:

- We plot the data pairs and perform a linear (1st-order) regression analysis. Standard error = $S_{y,x}$ = 0.5833.



- By eye, we suspect that the third data point (at x = 30.01) might be an outlier. To test our suspicion, we calculate and plot the standardized residuals $e_i / S_{y,x}$. For comparison, we also plot the zero line.



- For the third data point, $e_i / S_{y,x}$ = –2.064. Since |–2.064| > 2, we meet the *first* criterion for an outlier.
- What about the *second* criterion? Here, the third data point is indeed *inconsistent* with its neighbors (it is not smooth and *does not follow any kind of pattern*, but is simply way lower than any of the others).
- Since *both* criteria are met, we say that **the third data point is definitely an outlier**, and we remove it.

Discussion: After removing the outlier, we should re-do the regression analysis, obtaining a better fit.

- **Example**:

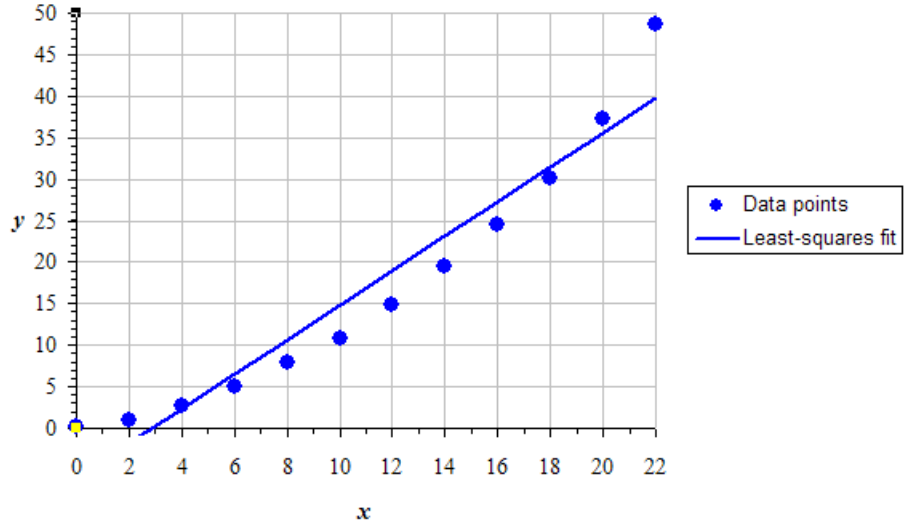
Given: Twelve data pairs (x,y) are measured in a calibration experiment, as shown below:

i	*x*	*y*
1	0.00	0.23
2	2.00	1.05
3	4.00	2.74
4	6.00	5.03
5	8.00	7.87
6	10.00	10.86
7	12.00	14.89
8	14.00	19.44
9	16.00	24.56
10	18.00	30.12
11	20.00	37.28
12	22.00	48.57

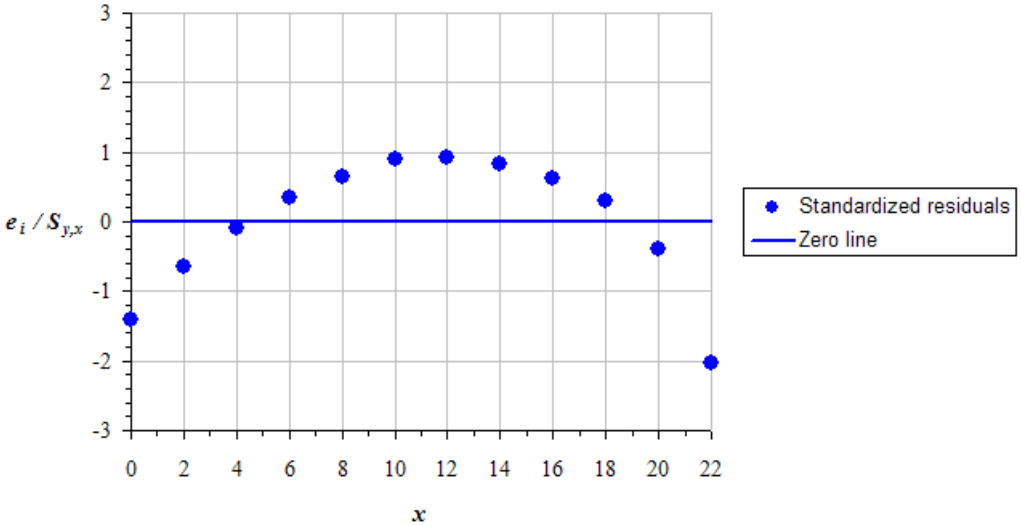
To do: Fit a straight line through these data, and determine if any of the data points should be thrown out as legitimate outliers.

Solution:

- We plot the data pairs and perform a linear (1st-order) regression analysis. Standard error = $S_{y,x}$ = 4.385.



- By eye, we suspect that perhaps the last data point (x = 22.00) might be an outlier. To test our suspicion, we calculate and plot the standardized residuals $e_i / S_{y,x}$. For comparison, we also plot the zero line.

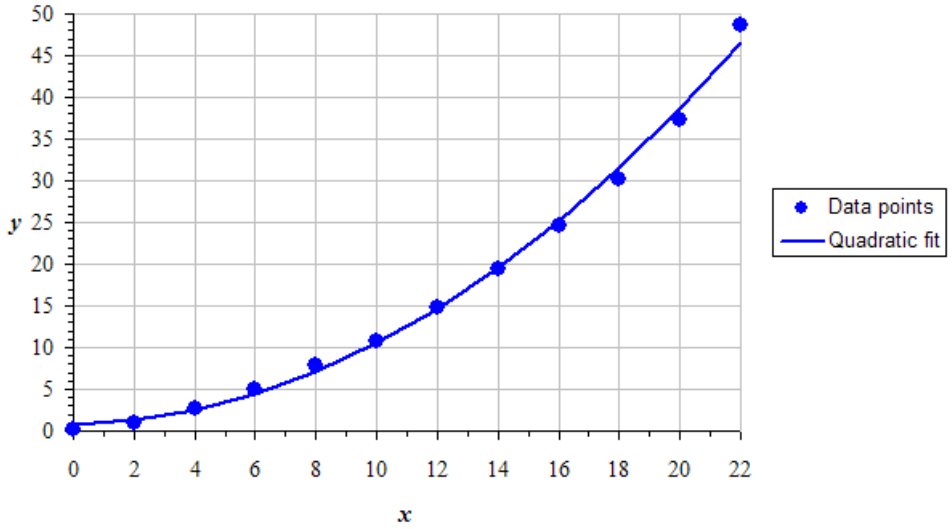


- For the last data point, $e_i / S_{y,x}$ = –2.025. Since |–2.025| > 2, we meet the *first* criterion for an outlier.

- The second criterion is a bit subjective, but the last data point *is* consistent with its neighbors (the data are *smooth* and follow a recognizable pattern). The second criterion is *not* met for this case.
- Since both criteria are not met, we say that **the last data point is *not* an outlier**, and we cannot justify removing it.

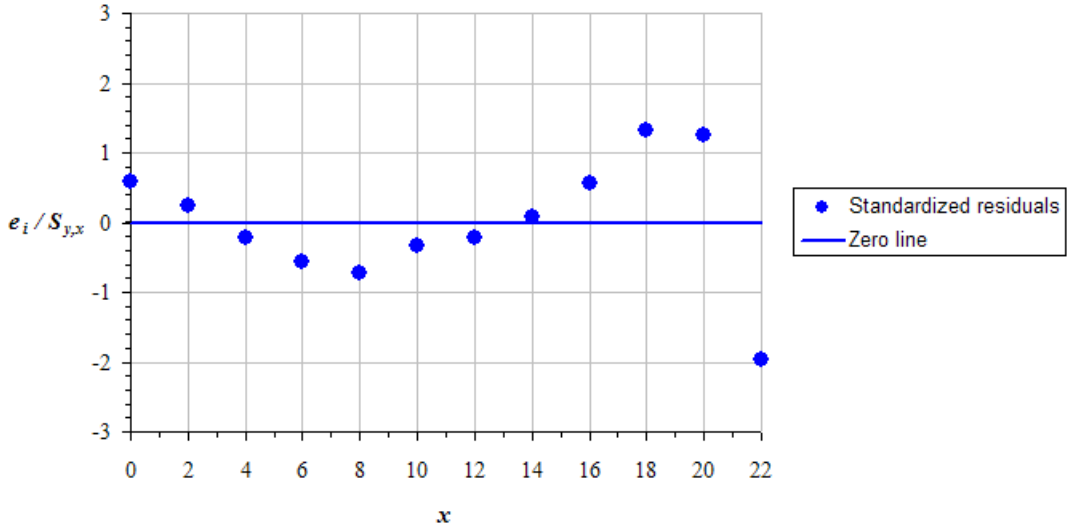
Discussion: When the standardized residual plot is *not* random, but instead, the data follow a pattern, as in this example, it often indicates that we should *use a different curve fit*.

- Here, we apply a second-order polynomial fit to the data, and re-plot the results. We plot the data pairs and perform a quadratic (2nd-order) regression analysis. Standard error = $S_{y,x}$ = 1.066 (much lower than the previous value of 4.385). *The quadratic fit is much better than the linear fit!*



- By eye, we see that the 2nd-order fit is indeed much better – the fitted curve passes through nearly all the data points.
- However, we still suspect that the last data point may be an outlier. Let's check.
- First criterion: Using the new value of $S_{y,x}$, we calculate the standardized residual at the last data point to be $e_i / S_{y,x}$ = –1.976. Since |–1.976| < 2, we do *not* meet the *first* criterion for an outlier.

  There are no outlier points for this quadratic fit.
- Finally, we plot the standardized residuals, just for fun:



- There still appears to be somewhat of a pattern in the standardized residuals, but the last data point is inconsistent with the pattern of the other points. Thus, we would say that the second criterion *is* met. However, since the first criterion failed, we cannot claim legitimately that the last data point is an outlier.
- Since the standardized residuals still show somewhat of a pattern, we might want to try a different kind of curve fit. However, comparing the fitted curve to the data, the agreement is excellent, so further analysis is probably not necessary in this example.