BY JULIA E. SEAMAN
AND I. ELAINE ALLEN

Outlier Options

Consider simple parametric tests to find an outlier's significance

EVEN IN THE MOST basic introductory statistics courses, we teach students that outliers in a data set can pose significant problems. We often teach that visually examining the data can help identify outliers. Beyond detection, however, few statistics textbooks devote much time to the subject of statistically assessing outliers and their effect on final analyses.

Students and researchers alike find outliers difficult to recognize. When outliers are identified, there is no clear set of statistical tools or tests available to find an outlier's significance. There are several outlier tests for parametric data, and we've applied them to chemical assay data to showcase the method.

Keep or remove?

Upon discovering a suspected outlier, the initial temptation has always been to eliminate the points from the data—with appropriate rationale—and simplify the analyses to make the results easy to explain. This method can be subjective, however, and may miss intricacies of the data. When there is more than one outlier or more than two variables in the analysis, the problem becomes more complex. Removing an outlier also can have large effects on any analysis of the data.

A good example to illustrate how outliers can affect an analysis, and even go undetected in an analysis, is Francis J. Ansombe's regression models, which are almost identical for four data sets that are markedly different.[1] Table 1 shows all analyses of the model.

Plotting the individual data sets shows how different they are, even with the same regression number. In particular, the plots show how much one outlier point can influence the analysis. From the plots of two of the data sets in Figures 1 and 2, it is clear the same regression line should not fit the points equally well, as they do, and that an outlier is evident in each plot.

It should be noted that some statistical software programs (for example, Minitab) report outliers in linear regression through the identification of highly standardized residual values as a default in their standard output for regression. Some software programs also have an option to provide plots of residuals versus the dependent values and probability plots of standardized residuals. This can help further identify outliers, but it may not be enough to statistically justify removing the data and may still miss some outliers.

Simple outlier tests

The majority of parametric outlier tests look at some measure of the relative distance of a particular data point to the mean of all the data points and assesses what the probability is that a particular piece of data occurred by chance. Most tests are designed to look at individual or specific points, but several can be generalized to examine multiple data points, usually pairs. In addition, pairs of points or n-tuples of points may represent combinations of variables and may be difficult to identify with a simple test.[2]

Most parametric tests are generalizations or extensions of the original work by F.E. Grubbs,[3] who derived several simple parametric tests that are used most frequently when testing for outliers. Grubbs tests can be given as follows (in which x_i denotes an individual data point, s is the sample standard deviation and n is the sample size):

$G_1 = \frac{|\overline{x} - x_i|}{s}$, looks for outliers in single points of data,

$G_2 = \frac{x_n | x_i}{s}$, finds outliers at the minimum and maximum of a distribution, and

$G_3 = 1 - \frac{(n-3) \times s^2_{n-2}}{(n-1) \times s^2}$, finds pairs of outliers at either extreme.

Dixon's Q test is similar to G_2 for a small number of observations (between 3 and 25), and Rosner's test is a generalization of Grubbs test to detect up to k outliers when the sample size is 25 or more.[4]

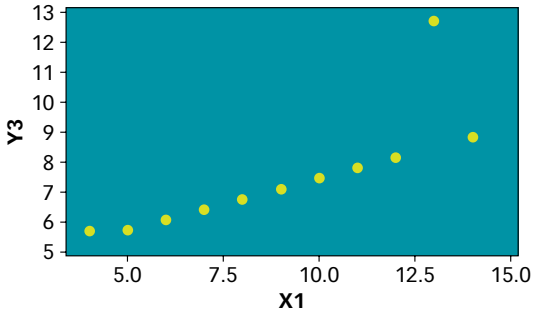
Grubbs test (G_1) example

We used the simplest form of a Grubbs test to remove outliers in infrared (IR) spectroscopy research data. IR spectroscopy was taken from mixtures of three organic compounds in solution, and the outliers

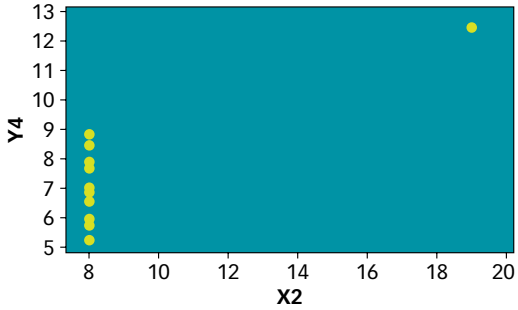
Regression model / TABLE 1

The regression equation is Y1 = 3.00 + 0.500 X1				
Predictor	Coefficient	Standard error	T-statistic	p-value
Constant	3.000	1.125	2.67	0.026
X1	0.5001	0.1179	4.24	0.002
Syx = 1.23660	R-squared = 66.7%			

Scatterplot of Y3 vs. X1 / FIGURE 1



Scatterplot of Y4 vs. X2 / FIGURE 2



needed to be removed before using the results in later chemometric analysis.

The purpose was to create a statistical model based on the spectra that can be used to determine unknown concentrations of compounds from an IR spectroscopy. By systematically removing the outliers, we start with cleaner data that will give us a better model and, ultimately, better results.

The mixtures were run in triplicate and included 1,501 data points of the IR spectrum. The samples were scanned in 2 cm^{-1} increments from 450 cm^{-1} to 4,400 cm^{-1}. The analyzed spectral region for all samples was 600 cm^{-1} to 3,500 cm^{-1}. Prior to making a chemometric model to predict unknown concentration values, the spectra sets were validated and examined for outliers. All the spectra in the updated data set were mean-centered before analysis.

Outliers were identified using the Grubbs test. As shown in G_1 above, this was done by finding the standard deviation for each data point between the triplicate spectroscopy values and then calculating the overall average standard deviation and the overall standard deviation of the data points' standard deviations for the triplicate. For each group of triplicates, these overall standard deviations were used in the Grubbs test.

When an overall triplicate standard deviation was rejected, the three runs within the triplicate were analyzed using a jackknife technique. A single run was removed from the triplicate of the outlier group if it significantly lowered the overall standard deviation of the group. The Grubbs test was repeated as needed. All statistical tests were done at the 95% confidence level.

In our IR data, the overall average for one triplicate group was 2.653, with a standard deviation of 2.888, and we calculated a Grubbs test statistic, G_1, of 5.22. With a G_{crit} from a Grubbs table of 1.91, G_1 is greater than G_{crit}, the null hypothesis is rejected, and the sample is declared an outlier.

We recalculated the overall standard deviation with one spectrum removed to find which remaining two reduced it the most. After finding and removing the most different triplicate, the overall standard deviation of the sample dropped to 0.04, confirming the outlier behavior of the eliminated spectrum.

Options for analyses

Examination and detection of outliers is a key part of any data analysis. Analyses that include data that are unusually large or small compared to the rest of the data set run the risk of estimating models that are not representative or that introduce variability. Analyses that exclude these values without testing their significance as outliers may seriously bias a model.

Parametric tests should be used when there are sufficient data available, sufficient precision in the data and no genuinely long tails on the distribution that would identify successive outliers when a Grubbs test is applied. A Grubbs test is easy to use and apply and, along with the graphical display of the data, can identify whether extreme data should be examined separately. QP

REFERENCES AND NOTE

1. Francis J. Anscombe, "Rejection of Outliers," *Technometrics*, Vol. 2, 1960, pp.123-147.
2. Vic Barnett and Toby Lewis, *Outliers in Statistical Data*, J. Wiley & Sons, 1984, offers a complete statistical overview of all outlier tests.
3. F.E. Grubbs, "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics,* Vol. 21, 1950, pp. 27-58.
4. Robert D. Gibbons, *Statistical Methods for Groundwater Monitoring,* J. Wiley & Sons, 1994.

BIBLIOGRAPHY

Belsley, D.A., E. Kuh and R.E. Welsch, *Regression Diagnostics,* J. Wiley & Sons, 1980.

Burke, S., "Missing Values, Outliers, Robust Statistics and Non-parametric Methods," *Scientific Data Management,* Europe online supplement, 2001, pp. 19-24.

Meijer, Rob R., "Outlier Detection in High Stakes Certification Testing," *Journal of Educational Measurement*, Vol. 39, No. 3, 2002, pp. 219-233.

Quesenberry, C.P., and H.A. David, "Some Tests for Outliers," *Biometrika*, Vol. 48, 1961, pp. 379-390.

Zhang, Jing, "Tests for Multiple Upper or Lower Outliers in an Exponential Sample," *Journal of Applied Statistics*, Vol. 25, No. 2, 1998, pp. 245-255.

JULIA E. SEAMAN is a researcher at Genentech in South San Francisco, CA. She earned a bachelor's degree in chemistry and mathematics from Pomona College in Claremont, CA.

I. ELAINE ALLEN is director of the Babson Survey Research Group and professor of statistics and entrepreneurship at Babson College in Wellesley, MA. She earned a doctorate in statistics from Cornell University in Ithaca, NY. Allen is a member of ASQ.