Messy Data, Analysis of

SAIF SHAHIN
Bowling Green State University, USA

Statistical analyses rely on some assumptions about the distribution of observations in a dataset as well as the relationships among the variables being studied. If the raw data is "messy" and does not meet one or more of such assumptions, the reliability of the analyses and the generalizability of the results comes into doubt. It is necessary, therefore, to check for messy data, identify the ways in which sample distributions and variable relations violate basic assumptions, and take ameliorative measures. Detecting and dealing with messy data is the very first step of a quantitative research project after data collection—conducted before analyses meant to address the objectives of the research are carried out.

The presence of univariate or multivariate outliers, skewness or kurtosis in a distribution, and heteroscedasticity or multicollinearity among variables are all examples of messy data. Scholars have outlined mathematical techniques to detect such problems in a dataset, determine the extent to which they could compromise analysis, as well as methods to address these issues.

Outliers

Data samples often have a few observations with extreme values on one variable, or with an irregular combination of values on two or more variables. Such observations are called outlying observations, or outliers, as they lie outside the normal distribution of the sample. While outliers represent "uncommon" cases in a sample, finding outliers is quite common and most medium- to large-sized randomized samples will have a few outliers. However, as statistical analyses typically assume a normal distribution of observations, the presence of outliers can potentially lead to erroneous interpretations and false generalizations (Bradley, 1984).

Outliers can creep into a dataset in a number of ways. Mistakes in data entry, failure to specify missing value codes, or errors in experimentation or instrumentation can lead to the presence of outlying observations. Therefore, every outlier should be checked carefully to ensure that data has been entered correctly, that missing values, if any, have been imputed properly, and that the instruments and procedures of experimentation worked as they were supposed to. A second reason could be that the outlier does not belong to the sampled population. Such an observation may simply be excluded from the sample. Finally, outliers may be present in a sample because the population does indeed have cases with extreme values or a combination of values on some variables.

The International Encyclopedia of Communication Research Methods. Jörg Matthes (General Editor), Christine S. Davis and Robert F. Potter (Associate Editors).

DOI: 10.1002/9781118901731.iecrm0152

While there are no strict mathematical rules for classifying particular observations as outliers, they have been viewed as a problem since the mid-18th century (Dodge, 2008). Early studies on identifying and dealing with outliers include the works of Adrien M. Legendre, Benjamin Peirce, and George B. Airy in the early to mid-19th century (see Stigler, 1973). Scholars follow certain norms for identifying outliers, which depend on the type of data a sample has and the type of outlier.

Univariate outliers

Univariate outliers are observations that register extreme values on one variable when compared with the rest of the sample. For example, in a sample of 50 elementary school students, if all students are aged 5 to 12 except one who is aged 17, then the oldest student may be considered an outlier on the age variable. As age is measured here in terms of open-ended consecutive numbers, this is also an example of univariate outlier in continuous variables. Mathematically, in a sample of n observations of variable X such that $x_1 < x_2 < x_3 < \ \ldots \ < x_n$, the biggest observation of x_n may be deemed an outlier if its value is exceptionally higher than all other values. As a rule of thumb, outliers in continuous variables are cases with large standardized scores, or z scores—typically in excess of 3.29 (Tabachnick & Fidell, 2013). Graphical methods such as histograms and normal probability plots may also be used to identify outliers.

Univariate outliers may also be present in categorical variables, that is, if the data is coded into two (dichotomous) or more categories. If the frequency of distribution among the categories is highly skewed and there are very few observations in a particular category, then those observations may be considered outliers. In the example presented earlier, if age is measured in categories of 0–4, 5–10, 11–15, and 16–20 years, we will again identify the same student as the outlier as all other students would fall in the second and third categories. For dichotomous variables, a 90:10 split in the number of cases generally implies that the smaller group is comprised of outliers.

Scholars have recommended several ways of dealing with univariate outliers, depending on why they occur as well as the purposes of the data analysis. The most basic is to make sure there were no mistakes in data entry, no errors in experimentation or instrumentation, and that missing value codes were imputed properly. Next, the researcher needs to consider if the outlying cases are indeed from the sampled population. If they are not, they can simply be deleted. Third, the researcher should check if most of the outliers are occurring on a single variable. If that is the case, one may consider eliminating the variable, especially if it is not vital to the analysis or is highly correlated with other variables.

If the outlying cases or the responsible variable(s) cannot be eliminated, the researcher should take steps to reduce their impact. There are two strategies to achieve this. The first is through variable transformation, or using a statistical technique that brings the shape of the distribution closer to normal. Several techniques may be used for variable transformation, depending on how different the sample is compared with normal distribution. A square root transformation is appropriate for a moderately different distribution. If the sample is substantially different from

normal, logarithmic transformation may be tried. For severely different distributions, an inverse transformation is required (see Box & Cox, 1964; Bradley, 1984). All these techniques will bring the outlying observations closer to the mean, thus reducing their impact. A second strategy for reducing the impact of univariate outliers is altering the scores of outlying observations so that they are no longer as deviant. Tabachnick and Fidell (2013) suggest assigning a raw score a unit larger or smaller than the next most extreme score.

Multivariate outliers

Some cases are outliers not because their values on a particular variable are far from normal but because their combination of values on two or more variables is unusual. In the sample mentioned earlier, suppose all students—after excluding the univariate outlier—are also measured for height and the range is found to be 40–60 inches. In this sample, all observations with age 6 will seemingly be normal, as will be all observations with a height of 58 inches. However, as younger students would likely be shorter and older students taller, if we find that a case that has both an age of 6 and a height of 58 inches, then that may be an outlier—a student who is too tall for his age.

Mahalanobis distance, a measure of how far an observation is from the intersection of the means of all variables in multivariate space, is commonly used to identify such outliers. It conceptually replicates the idea of *z* scores in multivariate space. Each observation in the dataset occupies a unique position in multivariate space by virtue of its unique combination of values on all variables. Typically, the observations cloud around the centroid, or the intersection of means of all variables. A multivariate outlier would occupy a point outside the cloud and can be distinguished using Mahalanobis distance. For an observation $x_i = (x_1, x_2, x_3, \ \dots \ , x_n)^{\mathrm{T}}$ in a sample with a mean of $m_i = (m_1, m_2, m_3, \ \dots \ , m_n)^{\mathrm{T}}$ and covariance matrix S, the Mahalanobis distance is measured as

$$\mathrm{MD}(x_i) = [(x_i - m_i)^{\mathrm{T}} \mathrm{S}^{-1} (x_i - m_i)]^{1/2}$$

The Mahalanobis distance gives lower weight to groups of highly correlated variables as well as to variables with large variances. The square of Mahalanobis distance corresponds to a chi-square distribution with the same degrees of freedom as the number of variables in the dataset. A conservative significance test of a case being an outlier—$p < .001$ for the X^2 value—is therefore appropriate with Mahalanobis distance (Tabachnick & Fidell, 2013).

Variable transformations or score alteration of outliers may be used to reduce the impact of multivariate outliers on statistical analyses, just as they are with univariate outliers. If a few outliers are still left after transformations or alterations, they may be deleted.

Nonnormality

When observations for a random variable are normally distributed, as represented by the Bell curve, the mean and the median of the distribution converge and the frequency

of observations across the distribution follow an expected pattern. But distributions could be distorted in a number of ways. The mean may diverge from the median, or the frequency of observations could be higher than expected closer to the mean or in the tails. Depending on the sample size, these distortions may affect statistical analyses and their interpretation.

Skewness

Skewness pertains to the horizontal symmetry of a distribution. Conceptually, it measures discrepancies in the values of observations in a distribution compared with a normal distribution. A skewed variable's mean does not coincide with its median. When a distribution has some observations with values that are quite high, the right tail of the Bell curve becomes longer. For instance, the values 3, 5, 5, 5, 7 represent a fairly normal distribution with both the mean and the median being 5. If we add another observation of 11 to the distribution, it will elongate the right tail of the curve. This is known as a *positive skew*. Conversely, when a distribution has some observations with values that are quite low, it is the left tail of the Bell curve that is elongated—constituting a *negative skew*. An often-followed rule of thumb for differentiating between the two kinds of skewness is that the mean lies to the right of the median in a positive skew, and to the left in a negative skew.

Skewness is zero for a normal distribution. Positive skewness is a lot more common because the lowest value of many distributions representing natural or social phenomena is fixed at zero while higher values are not bounded (von Hippel, 2011). Examples include distributions of age, income, time elapsed, and so on. Negative skewness is less common but occurs when observations tend to be closer to the maximum than the minimum value, such as the scores of an easy test.

Traditionally, skewness has been understood as the difference between the mean (m) and the median (v), divided by its standard distribution (s). This formula—$(m - v)/s$, attributed to Karl Pearson—conforms to the rule of thumb differentiation of positive and negative skewness. As the median is being subtracted from the mean in the numerator, a positive skew would imply the mean is to the right of the median and a negative skew would imply the mean is to the left of the median. Pearson later introduced a coefficient of skewness viewed in terms of the third standard moment, a more descriptive measure better suited for larger datasets (Dodge, 2008). Skewness for random variable X is thus measured as

$$\mu_3 = \mathrm{E}[\{(X - m)/\mathrm{s}\}^3]$$

Von Hippel (2005) warns that the rule of thumb does not always hold true with this definition, especially for categorical data. Variable transformations such as square-root or log transformation may be considered to deal with highly skewed data. But scholars have cautioned that transformed variables are difficult to interpret (Levine, Liukkonen, & Levine, 1996) and transformations may also alter the relationship among different variables (von Hippel, 2011). Decisions about transformation to normalize skewed data, therefore, should not be taken lightly—especially as the impact of skew is limited by the

size of the sample. In large samples, skewness does not make a substantive difference to the analysis (Tabachnick & Fidell, 2013).

Kurtosis

Kurtosis pertains to the vertical contours of a distribution—its "peakedness" near the mean or flatness in the tails. Conceptually, it measures discrepancies in the frequency of observations in a distribution compared with a normal distribution. Following Pearson, kurtosis is mathematically defined as the fourth standard moment. Thus, for a random variable X with mean m and standard deviation s, the coefficient of kurtosis is measured as

$$\mu_4 = \mathrm{E}[\{(X - m)/\mathrm{s}\}^4]$$

This is technically a measure of the "tailedness" of a distribution. The kurtosis of a normal distribution is 3. Distributions with kurtosis higher than 3 have thick, short tails and are known as *leptokurtic*. The frequency of observations is concentrated in the center of such a distribution, stretching its peak. Variables with this kind of nonnormality would have low variance. Distributions with kurtosis lower than 3 have thin, long tails and are known as *platykurtic*. As a kurtosis coefficient of 3 is normal, nonnormality is traditionally discussed in terms of "excess kurtosis." Some scholars, however, prefer to subtract 3 from the formula to bring the kurtosis coefficient of a normal distribution to zero. They also use positive and negative kurtosis to refer to leptokurtosis and platykurtosis, respectively (e.g., Tabachnick & Fidell, 2013).

As with skewness, the impact of kurtosis on statistical analyses diminishes with the size of the sample. Waternaux (1976) suggested that the impact of leptokurtosis disappears with 100 or more cases, while platykurtosis has little effect when the sample is in excess of 200 cases.

Heteroscedasticity

Linear regression models assume that the variance at each value point of the outcome variable corresponds to the variance of the explanatory variable. When this does not happen—a condition of heteroscedasticity—the model's predictive power is reduced. A common example of heteroscedasticity is the positive relationship between age and income. Teenagers typically do not earn a lot of money. But as they grow into their 20s, 30s, and 40s, their careers—and earnings—increase disproportionately. Some become lowly-paid school teachers or clerks, others become corporate executives or basketball coaches earning much higher salaries. Thus, even as income rises with age, the variance in income levels is much higher by comparison. The regression model's ability to predict income by age is, therefore, limited.

In other words, the random errors of such a regression model—residuals—will not belong to the same probability distribution with consistent variance. Mathematically, therefore, heteroscedasticity is said to occur when residuals have different probability distributions and different variances. Often, the variance increases proportionally to the

square of some factor, *F*, which could be an explanatory variable. For residuals e_i in an ordinary least squares (OLS) linear regression model with common variance σ_i^2,

$$\text{var}(e_i) = \sigma_i^2 F_i^2$$

Heteroscedasticity can emerge from a number of reasons. Nonnormality of variables is a common one, as apparent from the earlier above in which income is positively skewed. Another reason could be the difference in the sample sizes of variables. Such differences increase the probability that residuals will have different variances, leading to heteroscedasticity in the regression model. Finally, heteroscedasticity can occur if the regression model itself is not specified correctly and a significant explanatory variable is not included.

Several statistical tests are recommended for detecting heteroscedasticity. The Park test comprises regressing the natural logarithm of squared OLS residuals—$\ln[e_i^2]$—on the natural logarithm of the squared proportionality factor—$\ln[F^2]$. A statistically significant relationship between them would indicate heteroscedasticity. More commonly used is the White test, in which the squared residuals are regressed on all explanatory variables, their squares and cross-products. If the product of the R^2 and the sample size (n) is large, it indicates heteroscedasticity.

A popular way of dealing with heteroscedasticity in a regression model is to use "weighted" instead of ordinary least squares. When the residuals increase proportionally to F_i, all the variables in the model are divided by the "weight" F_i and the regression analysis is carried out again. The error terms would now have constant variance. When the residuals do not increase proportionally, the variables should be divided by $1/(e_i)^{1/2}$.

Multicollinearity

Multicollinearity occurs when two or more explanatory variables in a linear regression model are highly correlated. In other words, there exists linear dependence among the explanatory variables. This is a problem because OLS regression presumes there is no significant linear relationship among explanatory variables and they only predict the outcome variable with a high degree of certainty—not each other. If they are correlated, they predict the "same part" of the outcome variable, leading to redundancy. The presence of multicollinearity does not bias the overall regression model for a given sample. But it shows large standard errors for the correlated variables. That means it is not reliable for specific calculations involving these variables. Also, if the coefficients from the model are applied to another sample from the same population, it could lead to erroneous predictions.

Multicollinearity can occur in all kinds of datasets. A common reason is the presence of too many dummy variables. Smaller sample size can also be a cause. Minor levels of multicollinearity are acceptable. The problem arises when the correlation between two variables is .70 or above—a conservative rule of thumb (Tabachnick & Fidell, 2013). Bivariate correlation among explanatory variables is thus easy to detect. Measures such as variance inflation factor (VIF) and tolerance (TOL) are used to discern if

the multivariate correlation is too high. For a linear regression model with R_i^2 as the coefficient of determination,

$$\text{VIF} = 1/(1 - R_i^2)$$

$$\text{TOL} = 1/\text{VIF}$$

Conservative rules of thumb for identifying multicollinearity are VIF > 5 or TOL < .02.

There are several ways to deal with multicollinearity. If the problem is bivariate, one of the two variables may be omitted from the model. When the problem is multivariate, increasing the sample size of the dataset can help reduce standard errors and bring down the correlation among explanatory variables.

SEE ALSO: Coding; Confoundation Check; Data Imputation; Data Recording; Mean Centering; Measurement Error; Missing Values and Missing Data; Sampling, Random

References

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, *26*, 211–243. Retrieved from http://pegasus.cc.ucf.edu/~lni/sta6236/BoxCox1964.pdf (accessed March 6, 2017).

Bradley, J. V. (1984). The complexity of nonrobustness effects. *Bulletin of the Psychonomic Society*, *22*(3), 250–253. doi:10.3758/BF03333824

Dodge, Y. (2008). *The concise encyclopedia of statistics*. New York: Springer.

Levine, A., Liukkonen, J., & Levine, D. W. (1996). Equivalent inference using transformations. *Communications in Statistics, Theory and Methods*, *25*(5), 1059–1072. doi:10.1080/03610929608831748

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, *68*(344), 872–879. doi:10.1080/01621459.1973.10481439

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, *13*(2), n2. Retrieved from http://ww2.amstat.org/publications/jse/v13n2/vonhippel.html (accessed March 6, 2017).

Von Hippel, P. T. (2011). Skewness. In *International encyclopedia of statistical science* (pp. 1340–1342). Berlin/Heidelberg: Springer.

Waternaux, C. M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika*, *63*(3), 639–645. doi:10.1093/biomet/63.3.639

Further reading

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data*. New York: John Wiley & Sons.

Park, R. (1966). Estimation with heteroscedastic error terms. *Econometrica*, *34*(4), 888.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, *185*, 71–110. doi:10.1098/rsta.1894.0003

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. Retrieved from http://www.jstor.org/stable/1912934 (accessed January 29, 2017).
Yule, G. U. (1911). *Introduction to the theory of statistics*. London: Griffith.

Saif Shahin is an assistant professor in the School of Media and Communication, Bowling Green State University. His research focuses on news, technology, and politics. His articles have been published in refereed journals such as *Journalism & Mass Communication Quarterly*, *The International Journal of Press/Politics*, *Journalism: Theory, Practice & Criticism*, *Journalism Practice*, and *Communication Methods and Measures*. He has also authored chapters in books on social media and international politics.