TEACHER’S CORNER

Smoking and Cancers: Case-Robust Analysis of a Classic Data Set

Peter M. Bentler
University of California, Los Angeles

Albert Satorra
Universitat Pompeu Fabra

Ke-Hai Yuan
University of Notre Dame

A typical structural equation model is intended to reproduce the means, variances, and correlations or covariances among a set of variables based on parameter estimates of a highly restricted model. It is not widely appreciated that the sample statistics being modeled can be quite sensitive to outliers and influential observations, leading to bias in model parameter estimates. A classic public epidemiological data set on the relation between cigarette purchases and rates of 4 types of cancer among states in the United States is studied with case-weighting methods that reduce the influence of a few cases on the overall results. The results support and extend the original conclusions; the standardized effect of smoking on a factor underlying deaths from bladder and lung cancer is .79.

Although there are exceptions, correlation and covariance structure methods involve modeling standard product–moment correlations and covariances, and when they involve mean structures, also model the unweighted sample mean. However, these sample statistics are well known to be highly sensitive to outliers and influential observations; that is, the sample statistics being modeled could be substantially biased by such observations. This implies that various statistics associated with a resulting structural equation model also could be biased, and hence, might compromise scientific conclusions. They might also lose power (Wilcox, 2003). Informally, we have observed that most research reports provide no evidence regarding whether or not any

Correspondence should be addressed to Peter M. Bentler, Departments of Psychology and Statistics, UCLA, Box 951563, Los Angeles, CA 90095-1563. E-mail: bcntler@ucla.edu
such outlier cases have influenced their results, and Cohen, Cohen, West, and Aiken (2003) noted that “There are currently few published applications of robust statistics in the behavioral sciences” (p. 418). This article illustrates the problem and provides practical advice on how to use weighting procedures that downweight problematic cases in structural equation modeling (SEM). In terms of methodology, it is a review article. Our model of the classic smoking and cancer data of Fraumeni (1968) is, as far as we know, new.

Even a single case can destroy an otherwise excellent model. It is known that the influence function associated with the sample covariance is quadratic, so that a few influential cases or outliers can lead to inappropriate solutions for virtually all standard statistical methods that rely on sample covariances (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Zimmerman & Williams, 2000), including SEM. A theoretical understanding of how this happens was provided by Poon and Poon (2002) for means and covariances and by Yuan and Bentler (2001) for structural models.

Methods for handling data in nonstandard situations, such as methods that have bounded influence functions and can tolerate a high proportion of bad data before breaking down, have existed for a long time (e.g., Hoaglin, Mosteller, & Tukey, 1983; Maronna, 1976). However, in the past these methods have been presented primarily as exploratory and graphical methods, with little attention paid to standard problems of inference in the multivariate case (Huber, 1981). This is still true today (Filzmoser, Maronna, & Werner, 2008). A pioneering application to covariance structure analysis was published quite early (Huba & Harlow, 1987), but no general statistical development was proposed for another decade. Yuan and his colleagues (Yuan & Bentler, 1998a, 1998b, 2000; Yuan, Bentler, & Chan, 2004; Yuan, Chan, & Bentler, 2000) developed several methods to weight cases or observations differentially along with a complete statistical rationale for correct inference in the context of SEM. A review is provided by Yuan and Bentler (2007). Finding an appropriate weight to give each case makes the influence of outliers on a case-robust procedure minimal. Because there are in principle a lot of potential weight functions (see e.g., Table 11-1 of Hoaglin et al., 1983), there are also potentially a lot of different case weight vectors that could be used. Examples are Huber (1977) type weights, multivariate t weights (Lange, Little, & Taylor, 1989), descending weights (Hampel, 1974), and Campbell’s (1980) weights. The Yuan and Bentler (1998b) methodology based on Campbell’s weights has been available in EQS (Bentler, 2002–2009), for about 10 years and is illustrated later. An overview of a wide variety of modern methods for handling outliers and influential cases in a wide range of statistical contexts is provided by Wilcox (2005).

SMOKING AND CANCER

Fraumeni (1968) investigated the relation between the per-capita sales of cigarettes and variation in mortality from bladder, lung, kidney, and leukemia cancers. The original data (available at http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html) represent smoking rates in 44 states in the United States and the associated age-adjusted death rates for the four cancers. Fraumeni’s main interest was in evaluating whether smoking was a risk factor for bladder cancer, where “Respiratory tract cancer was selected for study because of the definite causal relation established with cigarette smoking, kidney cancer for its urinary tract localization, and leukemia for its lack of reported association with smoking” (p. 1206). His main method involved bivariate regression and correlation between smoking and a given cancer, and also included partialing
out the effects of urbanization (not given at the Web site). The scatter plot for smoking and bladder cancer is given in Figure 1.

It will be obvious that there is a strong relation between smoking and deaths from bladder cancer. It is also obvious that there are two outlying cases on the abscissa; that is, in the rate of smoking as reflected in cigarette sales. As noted by the CMU (Carnegie Mellon University) researchers, "Nevada and the District of Columbia are outliers in the distribution of cigarette consumption (sale) per capita by states in 1960. . . . The ready explanation for the outliers is that cigarette sales are swelled by tourism (Nevada) and tourism and commuting workers (District of Columbia)." Because these outliers are also along the regression line, their elimination only marginally influences the bivariate correlation (reduced from .68 to .60). There is also a strong association of smoking and lung cancer, a lower association with kidney cancer, and, as expected, no association with leukemia.

A model for these data is given in Figure 2. It is run in EQS with the usual model setup (not shown). Estimates in the table are the results of a confirmatory regression analysis of kidney cancer on the four independent variables of cigarette consumption, lung cancer, and bladder cancer with body mass index, age, and education as covariates. Although we have data on cancer cases from the National Cancer Institute (1968), the covariates have not been included.

Figures 1 and 2 show that states with higher cigarette sales had higher cancer rates for bladder and lung cancer and lower rates for kidney cancer. The scatter plot shows a high correlation between cigarette sales and bladder cancer rates (r = 0.75).

We have included a model of cigarette sales and cancer rates.
shown) with an added command CROBUST=2.00,1.25, which instructs EQS to do case-robust estimation with the given two parameters set at default values. Details on this method are given in the Appendix. The model introduces a latent variable $F_1$ to potentially account for the strong relation between bladder and lung cancer noted by Fraumeni (1968). It also allows prediction of kidney cancer by smoking, but no effect of smoking on leukemia. It was expected that this model might be a bit too parsimonious, for example, in not allowing leukemia cancer rates to correlate with any other variable. The final model is shown in Figure 3, which gives the standardized solution.

Two additional residual correlations were added to the model to obtain a well-fitting model (Yuan–Bentler scaled $χ^2 = 6.0, p = .2$). These account for unexplained correlations among leukemia and kidney cancers, and kidney and lung cancers, which result from variation of variables outside the system. Bladder and lung cancer death rates do indeed define a strong factor, which in turn is strongly predicted ($\hat{β} = .79$) by rates of cigarette sales. As Fraumeni (1968) expected, kidney cancer was marginally related to smoking, and leukemia not at all. Although it can be dangerous to draw inferences about causal effects at the level of individuals from aggregate data (i.e., committing the ecological fallacy; Robinson, 1950), Fraumeni’s hypothesis that smoking is a very strong risk factor for bladder cancer has more recently been confirmed (e.g., Brennan et al., 2000; Brennan et al., 2001).

From a technical point of view, the important result involves the case weighting. Forty cases received a weight of 1.0; that is, they were treated as usual. The downweighted cases were District of Columbia (.87), Pennsylvania (.34), Nevada (.01), and Alaska (.00). Nevada and District of Columbia, which were noted to be univariate outliers in the distribution of cigarette sales, are treated differently in this case-robust weighting based on the multivariate distribution.

**DISCUSSION**

We have illustrated the influence of outliers in the typical multivariate SEM context, as applied to the classic Fraumeni (1968) smoking and cancer data. We noted that a few problematic cases
received weights far below 1.0, thus reducing their influence on the means and covariances, and hence, on the SEM results. An alternative method to a full case-weighting method would be to give all acceptable observations a weight of 1.0, and a weight of 0.0 to a few problematic cases (i.e., to delete them from the data set). This simpler approach might be an appropriate solution when there are only a few clear-cut cases that are problematic. Methods for identifying such cases are widely known and implemented in standard packages (e.g., Cohen et al., 2003; Tabachnick & Fidell, 2007) as well as in most SEM packages. For example, EQS identifies cases that contribute maximally to Mardia's (1970) normalized coefficient of multivariate kurtosis and offers a simple DEL command to delete cases. A good discussion comparing outlier deletion and more general case weighting is given in Cohen et al. (2003, section 10.4).

However, complicated outlier detection methods are not necessarily needed, as often simple data inspection will reveal a problem. For example, consider the data provided by Chatterjee and Yilmaz (1992) in their review of regression diagnostics. They reported data on 24 patients and showed how various multivariate measures could identify a particular case as being problematic. If we take a simpler approach and evaluate the scatter plot of variables 2 and 3 (patient's age vs. an illness severity index), we obtain Figure 4 which also provides a bivariate regression line.

The individual who is 65 years old but is healthier than all younger patients is unusual. The correlation between these two variables, including this case is .39. The correlation, and hence the regression line, changes substantially when this case is deleted, as can be seen in Figure 5.

Now the regression line is quite different, reflecting a correlation between these variables of .79. For structural modeling, the question is whether the correlation of .39 or that of .79 would more accurately represent the population from which this sample is drawn. Models that incorporate these variables—or at least their parameter estimates—could easily be very

---

**FIGURE 4** Regression of severity of illness on age (Chatterjee & Yilmaz, 1992).
Regression of severity of illness on age with one case deleted.

Figure 5

Different if they are meant to reproduce a correlation of .39 versus one of .79. Case-weighting
procedures provide a more nuanced methodology to that of outlier removal, although simple
outlier removal also can be a result of case weighting. For example, when the case-robust
method is applied with a saturated model to the preceding two variables from Chatterjee and
Yilmaz, the outlying case is given a weight of 0.0 and all other cases are given a weight
of 1.0.

Of course, identified outliers also should be intensively studied. In the age versus illness
example, data on the older individual who is especially healthy might have been miscoded; or,
on the contrary, perhaps this person could provide an insight into the basis for good health at
an older age.

The only approach to case-weighting that is currently available in SEM programs (and in
EQS) is that based on downweighting observations that are far from the main cloud of the data.
Although this approach provides a useful tool for data analysis, it is not fully adequate because
does not differentiate among various kinds of problematic cases, including influential cases
whose inclusion or exclusion might have a great effect on an estimated parameter in the model,
and good or bad leverage cases that might help to improve or degrade estimation precision.
A thorough discussion of how to identify and deal with various types of problematic cases
is given in Yuan and Zhong (2008), who also relate their approaches to standard concepts
in robust regression as well as the wider SEM literature (e.g., case-level residuals; Bollen &

The problem of outliers and influential cases is different from the problem of nonnormal

distributions. Methods devised for dealing with nonnormal distributions, such as transforming
variables or using an asymptotically distribution free method (Browne, 1984) or the Satorra and
Bentler (1994; Chou, Bentler, & Satorra, 1991) scaled or adjusted test statistics and sandwich
standard errors, are designed to deal with smooth but nonnormal distributions. Although these
methods are widely known as “robust” in the SEM literature, they treat all cases equally and do not downweight problematic observations. In the statistical literature, the phrase “robust statistics” is limited to case-weighting methods, which we have here called case-robust to distinguish them from the broader set of robust methods used in SEM. In conclusion, although we have emphasized case-robust procedures for general SEM models, they are also available for two closely related procedures: principal components and exploratory factor analysis. We refer the interested reader to Cui, He, and Ng (2003), Ibáñez and Dauxois (2003), Hubert, Rousseeuw, and vanden Branden (2005), and Serneels and Verdonck (2008) for principal components; and to Yuan, Marshall, and Bentler (2002) and Pison, Rousseeuw, Filzmoser, and Croux (2003) for factor analysis.

ACKNOWLEDGMENTS

This research was supported in part by grants DA00017 and DA01070 from the National Institute on Drug Abuse (to Peter M. Bentler), grants SEJ2006-13537 and PR2007-0221 from the Spanish Ministry of Science and Technology (to Albert Satorra), and grant DMS04-37167 from the National Science Foundation (to Ke-Hai Yuan).

REFERENCES


APPENDIX
SOME DETAILS ON CASE-ROBUST ESTIMATION

Using case weights \( w_i \), unstructured case-robust means and covariances are computed as

\[
\hat{\mu}_i = \frac{\sum_{i=1}^{N} w_i z_{ii}}{\sum_{i=1}^{N} w_i}
\]

and

\[
\hat{\sigma}_{ij} = \frac{\sum_{i=1}^{N} w_i^2 (z_{ii} - \hat{\mu}_i) (z_{ij} - \hat{\mu}_j)}{\sum_{i=1}^{N} w_i^2 - 1}
\]

These define the unstructured estimators \( \hat{\mu}, \hat{\Sigma} \) used as input to obtain a mean or covariance structure. If the \( w_i \) all equal 1.0, these formulae give the usual unbiased sample covariance matrix with denominator \( (N - 1) \). In case-robust methods, the weights \( w_i \) are not known a priori and are iteratively updated, depending on the distance that a case is from the robust mean of the distribution in the metric of the inverse robust covariance matrix; that is,

\[
d_i = \{(z_i - \hat{\mu})' \hat{\Sigma}^{-1} (z_i - \hat{\mu})\}^{\frac{1}{2}}.
\]

These distances are translated into weights using Campbell's (1980) suggestion that

\[
w_i = \begin{cases} 
1 & \text{if } d_i \leq d_0 \\
\frac{d_0}{d_i} \exp\{-0.5(d_i - d_0)^2/b_2^2\} & \text{if } d_i > d_0
\end{cases}
\]

where \( d_0 = \sqrt{p + b_1/\sqrt{2}} \), \( p \) is the number of variables, and \( b_1 \) and \( b_2 \) are constants. If \( d_i \) is small, the case receives a weight of 1; otherwise the weight is smaller than 1. On the basis of empirical experience, Campbell gave three recommendations for choices of constants: (a) \( b_1 = \infty \), corresponding to the usual sample covariance; (b) \( b_1 = 2 \) and \( b_2 = \infty \), corresponding to a Huber-type M-estimator (Huber, 1981); and (c) \( b_1 = 2 \) and \( b_2 = 1.25 \), corresponding to a Hampel-type redescending M-estimator (Hampel, 1974). EQS uses case (c) as its default. Yuan and Bentler (1998b) proved that this methodology provides correct asymptotic behavior of the estimators, standard errors, and test statistics. The specific test statistic used in the example is the scaled test statistic, but EQS provides additional statistics such as the residual-based tests \( T_{\text{RES}}, T_{\text{YB(RES)}}, \) and \( T_{\text{F(RES)}} \), which have not yet been studied in detail.