Published in final edited form as:

Struct Equ Modeling. 2009 April 1; 16(2): 382–390. doi:10.1080/10705510902751382.

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

Peter M. Bentler,
University of California, Los Angeles

Albert Satorra, and
Universitat Pompeu Fabra

Ke-Hai Yuan
University of Notre Dame

Abstract

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 four types of cancer among states in the USA 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, i.e., the sample statistics being modeled may be substantially biased by such observations. This implies that various statistics associated with a resulting structural equation model (SEM) also may be biased, and hence, may compromise scientific conclusions. They may also lose power (Wilcox, 2003). Informally, we have observed that most research reports provide no evidence regarding whether or not any such outlier cases have influenced their results, and Cohen, Cohen, West, and Aiken (2003, p. 418) noted that “There are currently few published applications of robust statistics in the behavioral sciences.” This paper illustrates the problem and provides practical advice on how to use weighting procedures that downweight problematic cases in SEM. In terms of methodology, it is a review paper. 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

*Research supported in part by grants DA00017 and DA01070 from the National Institute on Drug Abuse (Bentler), grant SEJ2006-13537 from the Spanish Ministry of Science and Technology (Satorra), and grant DMS04-37167 from the National Science Foundation (Yuan).
Contact information: Peter M. Bentler, Departments of Psychology and Statistics, UCLA, Box 951563, Los Angeles CA 90095-1563.
bentler@ucla.edu Phone: 310 825 3831.

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). 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, 1998ab, 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 statistical inference in the context of SEM. A review is provided by Yuan and Bentler (2007). By finding an appropriate weight to give each case, the influence of outliers on a case-robust procedure is minimal. Since 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-Bentler (1998b) methodology based on Campbell's weights have been available in EQS for about 10 years and is illustrated below. 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, given at <http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html>, represent smoking rates in 44 states in the USA 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 cmu 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, i.e., in the rate of smoking as reflected in cigarette sales. As noted by the CMU 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) 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 F1 to potentially account for the strong

relation between bladder and lung cancer noted by Fraumeni. 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, e.g., 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.

It is seen that two additional residual correlations were added to the model to obtain a well-fitting model (Yuan-Bentler scaled $\chi^2_4=6.0$, $p=.2$). These account for unexplained correlations among leukemia and kidney cancers, and kidney and lung cancers, that 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 ($\beta=.79$) by rates of cigarette sales. As Fraumeni 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, 2001).

From a technical point of view, the important result involves the case weighting. Forty cases received a weight of 1.0, i.e., they were treated as usual. The downweighted cases were: District of Columbia (.87), Pennsylvania (.34), Nevada (.01), 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 may 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 such as SPSS and SAS (see e.g., Cohen, Cohen, West, & Aiken, 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, 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 subject 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, more accurately would represent the population from which this sample is drawn. Models that

incorporate these variables – or at least their parameter estimates -- could easily be very 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 above 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 vs. illness example, data on the older subject who is especially healthy might have been miscoded; or, on the contrary, perhaps this subject 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 structural modeling programs (and in EQS) is that based on downweighting observations that are far from the main cloud of the data. While this approach provides a useful tool for data analysis, it is not fully adequate since it does not differentiate among various kinds of problematic cases, including influential cases whose inclusion or exclusion may have a great effect on an estimated parameter in the model, and good or bad leverage cases that may 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 (in press), who also relate their approaches to standard concepts in robust regression as well as the wider SEM literature (e.g., case-level residuals, Bollen & Arminger, 1991).

The problem of outliers and influential cases is different from the problem of non-normal distributions. Methods devised for dealing with non-normal distributions, such as transforming variables or using an asymptotically distribution free method (Browne, 1984) or the Satorra-Bentler (1994; Chou, Bentler & Satorra, 1991) scaled or adjusted test statistics and sandwich standard errors, are designed to deal with smooth but non-normal 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 He and Ng (2003), Ibazizen 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.

References

- Bollen, K.A.; Arminger, G. Observational residuals in factor analysis and structural equation models. In: Marsden, P., editor. *Sociological Methodology*. Cambridge, MA: Blackwell; 1991. p. 235-262.
- Brennan P, Bogillot O, Cordier S, Greiser E, Schill W, et al. Cigarette smoking and bladder cancer in men: A pooled analysis of 11 case-control studies. *International Journal of Cancer* 2000;86:289–294.
- Brennan P, Bogillot O, Greiser E, Chang-Claude J, Wahrendorf J, et al. The contribution of cigarette smoking to bladder cancer in women (pooled European data). *Cancer Causes Control* 2001;12:411–417. [PubMed: 11545456]
- Browne MW. Asymptotically distribution-free methods for the analysis of co-variance structures. *British Journal of Mathematical and Statistical Psychology* 1984;37:62–83. [PubMed: 6733054]
- Campbell NA. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics* 1980;29:231–237.

- Chatterjee, Yilmaz. A review of regression diagnostics for behavioral research. *Applied Psychological Measurement* 1992;16:209–227.
- Chou C-P, Bentler PM, Satorra A. Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology* 1991;44:347–357. [PubMed: 1772802]
- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum; 2003.
- Cui H, He X, Ng KW. Asymptotic distributions of principal components based on robust dispersions. *Biometrika* 2003;90:953–966.
- Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Computational Statistics & Data Analysis* 2008;52:1694–1711.
- Fraumeni JF. Cigarette smoking and cancers of the urinary tract: Geographic variations in the United States. *Journal of the National Cancer Institute* 1968;41:1205–1211. [PubMed: 5688094]
- Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 1974;69:383–393.
- Hampel, FR.; Ronchetti, EM.; Rousseeuw, PJ.; Stahel, WA. *Robust statistics: The approach based on influence functions*. New York: Wiley; 1986.
- Hoaglin, DC.; Mosteller, F.; Tukey, JW. *Understanding robust and exploratory data analysis*. New York: Wiley; 1983.
- Huba GJ, Harlow LL. Robust structural equation models: Implications for developmental psychology. *Child Development* 1987;58:147–166.
- Huber, PJ. Robust covariances. In: Gupta, SS.; Moore, DS., editors. *Statistical decision theory and related topics*. Vol. Vol. 2. New York: Academic Press; 1977. p. 165–191.
- Huber, PJ. *Robust statistics*. New York: Wiley; 1981.
- Hubert M, Rousseeuw PJ, vanden Branden K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 2005;47:64–79.
- Ibrazzen M, Dauxois J. A robust principal component analysis. *Statistics* 2003;37:73–83.
- Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* 1989;84:881–896.
- Mardia KV. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 1970;57:519–530.
- Maronna RA. Robust M-estimators of multivariate location and scatter. *Annals of Statistics* 1976;4:51–67.
- Pison G, Rousseeuw PJ, Filzmoser P, Croux C. Robust factor analysis. *Journal of Multivariate Analysis* 2003;84:145–172.
- Poon W-Y, Poon YS. Influential observations in the estimation of mean vector and covariance matrix. *British Journal of Mathematical and Statistical Psychology* 2002;55:177–192. [PubMed: 12034019]
- Robinson WS. Ecological correlations and the behavior of individuals. *American Sociological Review* 1950;15:351–357.
- Satorra, A.; Bentler, PM. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A.; Clogg, CC., editors. *Latent variables analysis: Applications for developmental research*. Thousand Oaks, CA: Sage; 1994. p. 399–419.
- Serneels S, Verdonck T. Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis* 2008;52:1712–1727.
- Tabachnick, BG.; Fidell, LS. *Using multivariate statistics*. 5th Ed.. Boston, MA: Pearson; 2007.
- Wilcox, RR. Power: practical problems, and possible solutions. In: Weiner, IB.; Schinka, JA.; Velicer, WV., editors. *Handbook of psychology*. Vol. Vol 2. Hoboken, NJ: Wiley; 2003. p. 65–85. Research methods in psychology
- Wilcox, RR. *Introduction to robust estimation and hypothesis testing*. 2nd Ed.. Burlington, MA: Elsevier Academic; 2005.
- Yuan K-H, Bentler PM. Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology* 1998a;51:63–88. [PubMed: 9670817]

- Yuan, K-H.; Bentler, PM. Structural equation modeling with robust covariances. In: Raftery, A., editor. *Sociological methodology*. Malden, MA: Blackwell; 1998b. p. 363-396.
- Yuan K-H, Bentler PM. Robust mean and covariance structure analysis through iteratively reweighted least squares. *Psychometrika* 2000;65:43–58.
- Yuan K-H, Bentler PM. Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical & Statistical Psychology* 2001;54:161–175. [PubMed: 11393898]
- Yuan, K-H.; Bentler, PM. Robust procedures in structural equation modeling. In: Lee, S-Y., editor. *Handbook of latent variable and related models*. Amsterdam: North-Holland; 2007. p. 367-397.
- Yuan K-H, Chan W, Bentler PM. Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology* 2000;53:31–50. [PubMed: 10895521]
- Yuan K-H, Bentler PM, Chan W. Structural equation modeling with heavy tailed distributions. *Psychometrika* 2004;69:421–436.
- Yuan K-H, Marshall LL, Bentler PM. A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika* 2002;67:95–122.
- Yuan K-H, Zhong X. Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology*. (in press).
- Zimmerman DW, Williams RH. Restriction of range and correlation in outlierprone distributions. *Applied Psychological Measurement* 2000;24:267–280.

Appendix: Some Details on Case-Robust Estimation

Using case weights w_t , unstructured case-robust means and covariances are computed as

$$\hat{\mu}_i = \sum_{t=1}^N w_t z_{ti} / \sum_{t=1}^N w_t$$

and

$$\hat{\sigma}_{ij} = \sum_{t=1}^N w_t^2 (z_{ti} - \hat{\mu}_i)(z_{tj} - \hat{\mu}_j) / \left(\sum_{t=1}^N w_t^2 - 1 \right).$$

These define the unstructured estimators $\hat{\mu}$, $\hat{\Sigma}$ used as input to obtaining a mean and/or covariance structure. If the w_t all equal 1.0, these formulae give the usual unbiased sample covariance matrix with denominator $(N - 1)$. In case-robust methods, the weights w_t 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,

$$d_t = \{(z_t - \hat{\mu})' \hat{\Sigma}^{-1} (z_t - \hat{\mu})\}^{\frac{1}{2}}.$$

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

$$w_t = \begin{cases} 1 & \text{if } d_t \leq d_0 \\ \frac{d_0}{d_t} \exp\{-.5(d_t - d_0)^2 / b_2^2\} & \text{if } d_t > d_0 \end{cases}$$

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_t 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) prove 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_{RES} , $T_{\text{YB(RES)}}$, and $T_{\text{F(RES)}}$, which have not yet been studied in detail.

Figure 1.
Relation between Cigarette Sales and Bladder Cancer for 44 States (Fraumeni, 1968)

Figure 2.
Model of Cigarette Sales and Cancer Rates

Figure 4.
Regression of Severity of Illness on Age (Chatterjee & Yilmaz, 1992)

Figure 5.
Regression of Severity of Illness on Age With One Case Deleted