

Case-Level Residual Analysis in the CALIS Procedure

Catherine Truxillo, SAS

ABSTRACT

This paper demonstrates the new case-level residuals in the CALIS procedure and how they differ from classic residuals in structural equation modeling (SEM). Residual analysis has a long history in statistical modeling for finding unusual observations in the sample data. However, in SEM, case-level residuals are considerably more difficult to define because of 1) latent variables in the analysis and 2) the multivariate nature of these models. Historically, residual analysis in SEM has been confined to residuals obtained as the difference between the sample and model-implied covariance matrices. Enhancements to the CALIS procedure in SAS/STAT® 12.1 enable users to obtain case-level residuals as well. This enables a more complete residual and influence analysis. Several examples showing mean/covariance residuals and case-level residuals are presented.

INTRODUCTION

Structural equation models (SEMs) use information in a covariance matrix (and optionally, a mean vector) to make inferences about a system of directional, regression-like relationships. These models frequently include latent variables, which cannot be directly observed, as well as manifest variables. Researchers express these models as systems of equations, model matrices, and path diagrams. Models are assessed on their overall fit to the data, rather than one-parameter-at-a-time. Until recently, there was not a convenient way to evaluate the impact of a small number of cases on the estimated parameters in the model. Case-level residuals in the CALIS procedure make this problem a thing of the past.

MODEL RESIDUALS

In SEM, a researcher proposes a model or a system of relationships among the variables that you have observed, by hypothesizing constraints on the associations among variables in the analysis. To assess goodness-of-fit, you compare the model-implied variance-covariance matrix and mean vector, with structural constraints (equality, zero, and so on), to the sample-based estimates of location and scale, with no constraints. From this the CALIS procedure can compute residuals based on the moment matrices, hereinafter referred to as *model residuals*.

Raw model residuals are computed as $s_{ij} - \hat{\sigma}_{ij}$, $\bar{x}_i - \hat{\mu}_i$ for covariance and mean residuals, respectively, where s_{ij} is an element of the sample covariance matrix, $\hat{\sigma}_{ij}$ is an element of the model-implied covariance matrix, \bar{x}_i is the vector of sample means, and $\hat{\mu}_i$ is the model-implied mean vector. (In practice, mean parameters are rarely constrained, so the sample and model-implied mean vectors are identical in many cases.) Raw residuals depend on the scale of the variables, and their magnitude is not a useful tool for determining whether the residual is “large.” It is useful to evaluate lack-of-fit with one of the standardized residual methods, such as the normalized residuals. The normalized residuals divide the raw residuals by the square root of elements of the estimated asymptotic covariance matrix of the sample covariances (for covariance residuals) or the estimated asymptotic covariance matrix of sample means (for mean residuals). There are several standardization methods available in PROC CALIS.

Model residuals help identify areas of poor overall model fit and misspecification. However, they do not tell you about the model's ability to predict specific cases, or of the influence those cases exert on parameter estimates in your model.

In a fully saturated model, as is the case with all OLS regression models, model residuals are always 0. It is the set of constraints in the model that enable you to test hypotheses about the fit of the model to the data.

CASE-LEVEL DIAGNOSTICS

Recent developments in SEM literature (Yuan and Hayashi, 2010; Yuan and Zhong, 2008) extend regression diagnostics and robust estimation (Rousseeuw and van Zomeren, 1990) to multivariate models with latent variables.

For an excellent, thorough discussion of how case-level diagnostics in SEM are obtained, and how they differ from standard regression diagnostics, see Yuan and Hayashi (2010). A brief description of the diagnostics available in the CALIS procedure follows.

Latent variables, which are not directly observed, must be estimated from observed data. In order to involve a latent variable in a residual computation, there must be a prediction for the latent variable or variables. It should be noted that this is not a sample versus population issue: even if the population of cases was available, the latent variables would still be estimated, as it is the variable, rather than the observation, that cannot be directly observed. Lawley and Maxwell (1971) discuss two methods for predicting the latent variables (the regression formulations and the Bartlett formulation) of which the Bartlett factor score predictor is considered superior by Yuan and Hayashi (2010, see Appendix A). In brief, the Bartlett factor score is given by a function of the factor loadings and the error covariance matrix derived from the parameter estimates of the SEM, and the manifest indicator variables.

Specifically, the vector of Bartlett factor scores in PROC CALIS are computed as

$$\hat{\mathbf{f}}_i = \left(\hat{\Lambda} \hat{\Psi}^{-1} \hat{\Lambda} \right)^{-1} \hat{\Lambda} \hat{\Psi}^{-1} \mathbf{y}_i$$

where $\hat{\Lambda}$ is the estimated factor matrix, $\hat{\Psi}$ (some authors denote this Θ) is the estimated error covariance matrix, \mathbf{f} are the factor scores, and \mathbf{y} are the manifest indicator variables. These definitions of matrices for Bartlett factor scores work exactly as they do for factor models, but they need to be defined with modifications for general structural equation models. (Yuan and Hayashi, 2010; Yung and Yuan, 2014.) This formula enables the procedure to compute casewise residuals as if all the values of the factors were observed.

In a multivariate model, there is a vector of residuals for each observation. A single summary measure of the magnitude of the residuals (per case) makes diagnostic checking a more manageable task. The problem of computing residuals for a multivariate model (with a vector of dependent variables) is solved by using a Mahalanobis distance (M-distance).

As a reminder, a general formula for Mahalanobis distances is

$$d_i = \sqrt{(x_i - \bar{x})' COV^{-1} (x_i - \bar{x})}$$

where COV could refer to any variance-covariance matrix and i is the index of the i^{th} observation. In the CALIS procedure, you can obtain a form of Mahalanobis distances, the residual M-distances (Yuan and Hayashi, 2010; SAS/STAT[®]: *User's Guide*):

$$d_{ri} = \sqrt{(\mathbf{L}\hat{\mathbf{e}}_i)' (\mathbf{L}\Omega_e \mathbf{L})^{-1} (\mathbf{L}\hat{\mathbf{e}}_i)}$$

which use the independent components of the residual vector $\hat{\mathbf{e}}$ and the covariance matrix of the residuals, Ω_e . \mathbf{L} is a matrix that makes the inversion possible and reduces $\hat{\mathbf{e}}$ to its independent components. The way that they are defined in PROC CALIS, M-distances are always positive, so the larger the M-distance, the further the case is from the origin of the residuals. This makes them useful for detecting outliers. According to Yuan and Hayashi (2010), if the residuals follow a multivariate normal distribution, then the M-distances are distributed as the square root of the corresponding χ^2 variate with df equal to the number of independent components. In other words, you can compare the residuals to quantiles from a χ distribution with r degrees of freedom to identify multivariate outliers as observations whose M-distances differ significantly from the origin,

$$pr(\chi_r \geq d_{ri}) < \alpha \text{ (by default, } \alpha = 0.01)$$

It should be noted that, on average, one percent of the sample will be identified as outliers by chance alone, when assumptions are met. As sample size n increases, the number of outliers increases as well.

GOOD AND BAD LEVERAGE OBSERVATIONS

Outliers are cases that are unusual in the space of the residuals. In other words, they are far from their predicted values. This is analogous to being far from the regression line in OLS regression models. Outliers can impact the precision of parameter estimates, lead to biased parameter estimates, or both. An outlier is not necessarily a leverage observation (nor are leverage observations necessarily outliers).

Leverage observations are far from the predictors in the model. The leverage M-distances in PROC CALIS are a generalization of Yuan and Hayashi (2010).

$$\mathbf{d}_{vi} = \sqrt{\mathbf{v}_i (\Omega_v)^{-1} \mathbf{v}_i'}$$

where $\mathbf{v}_i = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{f}_i \end{bmatrix}$ is a vector of exogenous manifest (\mathbf{x}_i) and latent (\mathbf{f}_i) variables, and $\mathbf{\Omega}_v$ is the covariance matrix

of the exogenous variables. See Yung and Yuan (2014) for the definitions of the formulas and SAS/STAT®: User's Guide for more information. Significant leverage observations are determined in the same way as the outliers: by comparison to a χ distribution. However, not all leverage observations impact models in the same way.

If the leverage distance is in line with the prediction by the model (in other words, it is not an outlier in the residuals), then it is a *good leverage* observation. If the leverage observation is not in line with the model predictions (in other words, it is an outlier and it is far from the predictors), then it is a *bad leverage* observation. Figure 1 shows examples of outlier (non-leverage), bad leverage (outlier), and good leverage observations in a simple linear regression context.

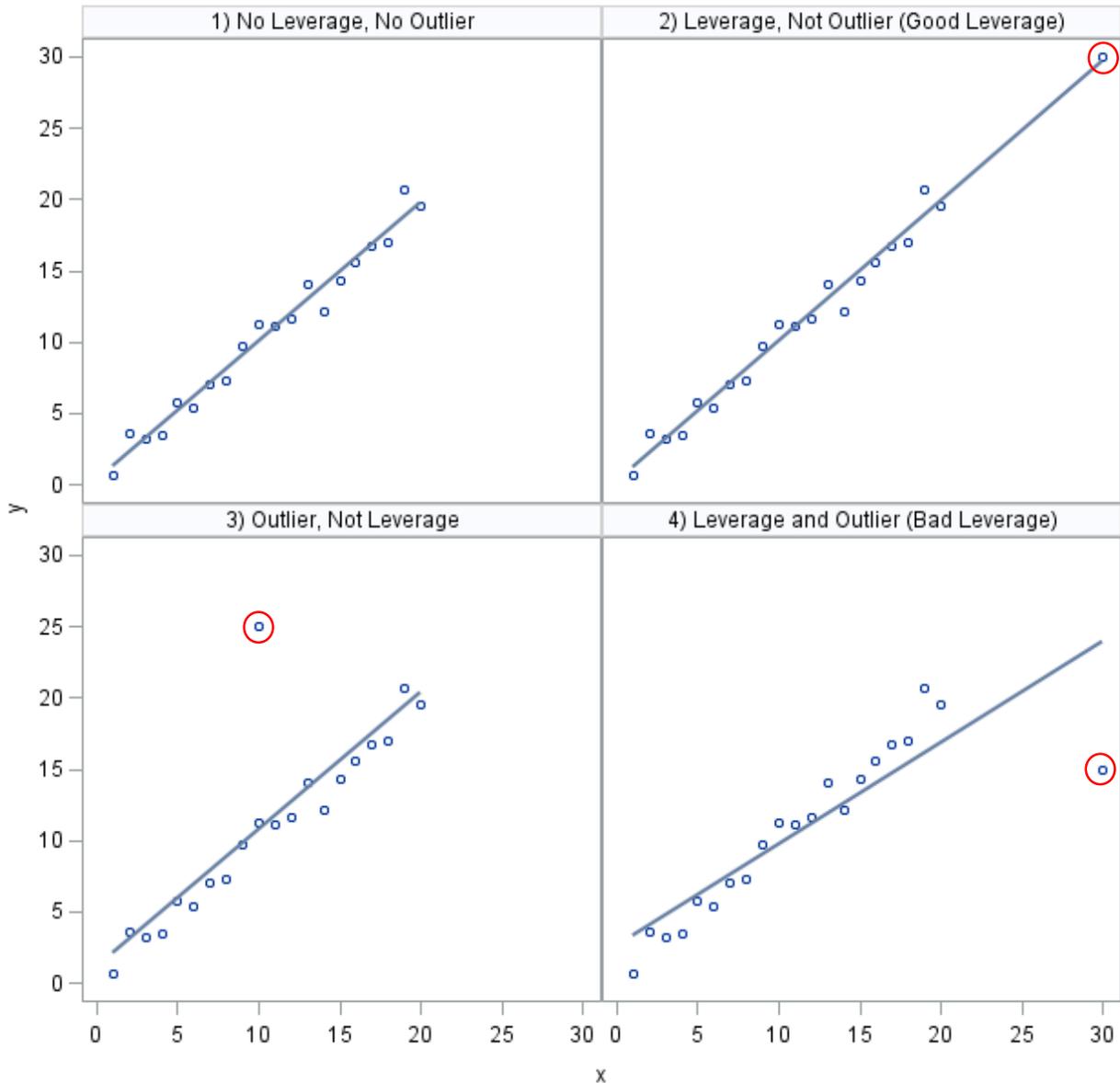


Figure 1: Outlier and Leverage Cases

Consider the consequences of the three types of observations above. Good leverage (non-outlier) observations, much like axial points in designed experiments such as central composite designs, improve the precision of the

model without biasing parameter estimates. As the name implies, good leverage observations are beneficial to the model.

Non-leverage outliers lead to biased parameter estimates and inflated standard errors, although their impact is not as serious as that of bad leverage observations. The most likely consequence of non-leverage outliers is incorrect inference. Bad leverage observations can substantially bias parameter estimates and lead to incorrect inferences.

Once outliers and bad leverage observations have been identified, the analyst must determine the best course of action. It is usually poor practice to delete observations unless they are known to be erroneous. Sometimes outliers give information about effects that are absent from the model, or help to identify a previously undiscovered phenomenon. Increasingly, analysts seek estimation techniques that minimize the impact of outliers.

ROBUST DIAGNOSTICS

When there are multiple outliers or leverage observations in the data, there is a well-known masking effect, which impacts estimation and detection of outliers (Rousseeuw and van Zomeren, 1990). It is possible for one large outlier to make it difficult to detect less dominant, but nonetheless important, outliers in the data. Hand-in-hand with outlier detection is determining a way to unmask outliers and leverage observations. Robust regression methods are popular for their ability to estimate unknown model parameters in the presence of unusual and influential observations, and to make those observations easier to detect.

Robust estimation downweights problematic observations during estimation, effectively minimizing the impact of these problematic observations. Interestingly, robust estimation reduces the masking effect of outliers, but also eliminates the need to remove outliers to obtain good model estimates. The analyst is able to identify outliers, discover new phenomena, and consider alternative models that better explain the outliers, while still obtaining unbiased and efficient parameter estimates for the specified model, keeping those observations in the analysis.

Robust diagnostics are similar to their non-robust counterparts, with some simple but key modifications. The CALIS procedure offers multiple methods to perform robust estimation. With two-stage robust estimation (ROBUST=SAT), observations are downweighted before ML estimation is performed, and a robust mean vector and covariance matrix are used in normal ML estimation. With direct robust estimation (ROBUST=RES(E) or ROBUST=RES(F)), reweighting iteratively occurs within the estimation, which effectively downweights outliers and bad leverage observations (but not good leverage observations).

EXAMPLE 1: A MODEL OF GOOD BEHAVIOR

The first example is a completely mediated model for the relationship between a company's culture of innovation and its market share, as mediated by product quality and marketing activity. For simplicity, these variables are treated as completely observed. Later you will see an example with latent variables.

The PATHDIAGRAM statement below produces the initial path diagram. (See Figure 2.) The PATHDIAGRAM statement requires SAS/STAT® 13.1 or later.

```
proc calis data = share meanstr;
  path
    MarketShare <-- ProductQuality MktInvestment,
    MktInvestment ProductQuality <-- Innovation;
  pathdiagram diagram=initial arrange=flow;
run;
```

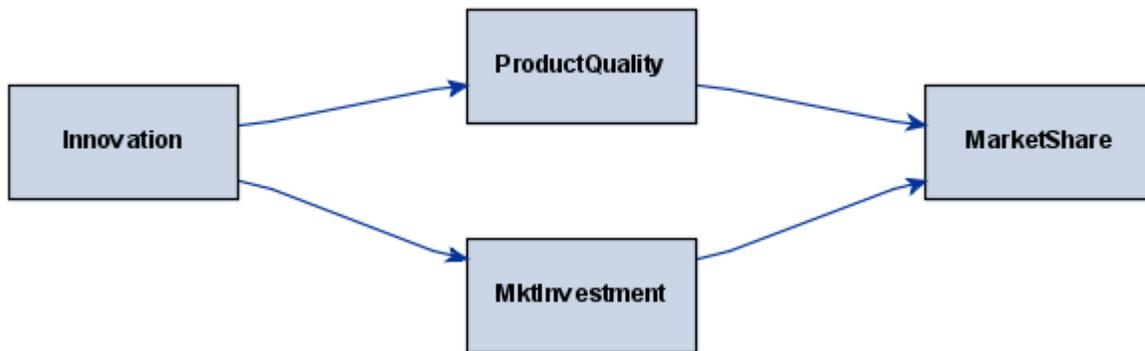


Figure 2. Initial Path Diagram for the Completely Mediated Model

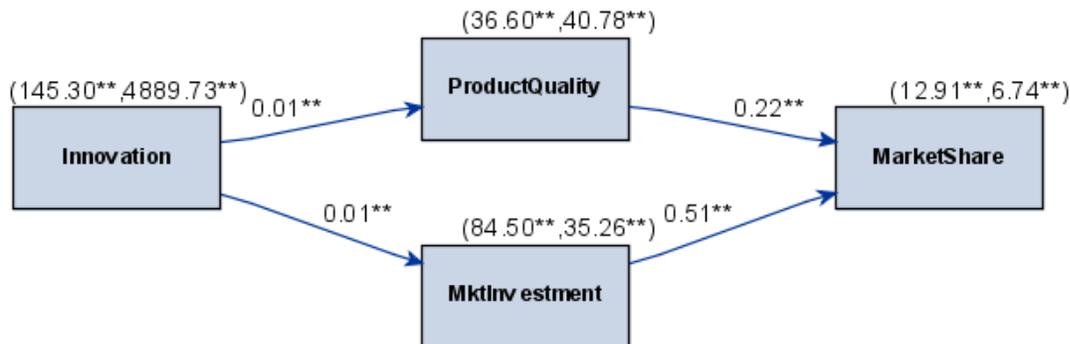
This example uses simulated data. If the model shows poor fit, it is useful to evaluate the model residuals.

To obtain the model residuals for the analysis, add the keyword RESIDUALS to the PROC CALIS statement:

```
proc calis data = share meanstr residuals;
```

To obtain the final path diagram with selected fit statistics in an inset box, add the following PATHDIAGRAM statement. (See Figure 3.)

```
pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] arrange=flow;
```



Chi-sq	1.14
DF	2
Pr > Chi-sq	0.57
CFI	1.00
SRMR	0.01
RMSEA	0.00

Figure 3. Path Diagram and Fit Statistics with Unstandardized Parameter Estimates for the Share Data

This model shows adequate fit to the data. The model residuals are relatively small. (See Figure 4.)

Average Normalized Residual	0.160540
Average Normalized Residual in Covariances	0.224756
Average Off-diagonal Normalized Residual	0.334087
Average Normalized Residual in Means	0

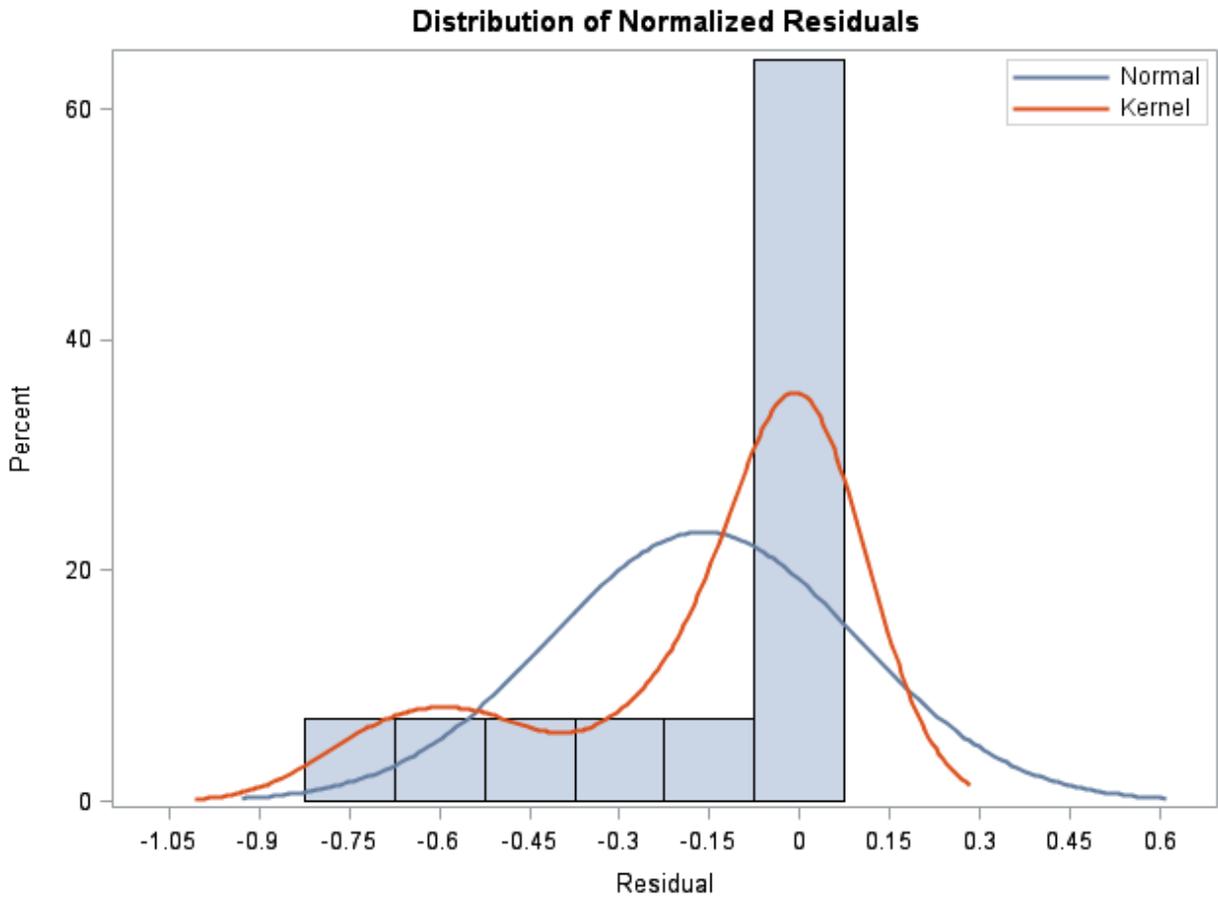


Figure 4: Normalized Model Residuals for the Share Data

The case-level residuals are requested with the following option in the PROC CALIS statement:

```
plots=(caseresidual)
```

This results in the following complete syntax:

```
proc calis data = share meanstr residual=norm plots=(residuals caseresidual);
  path
    MarketShare <-- ProductQuality MktInvestment,
    MktInvestment ProductQuality <-- Innovation;
  pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] arrange=flow;
run;
```

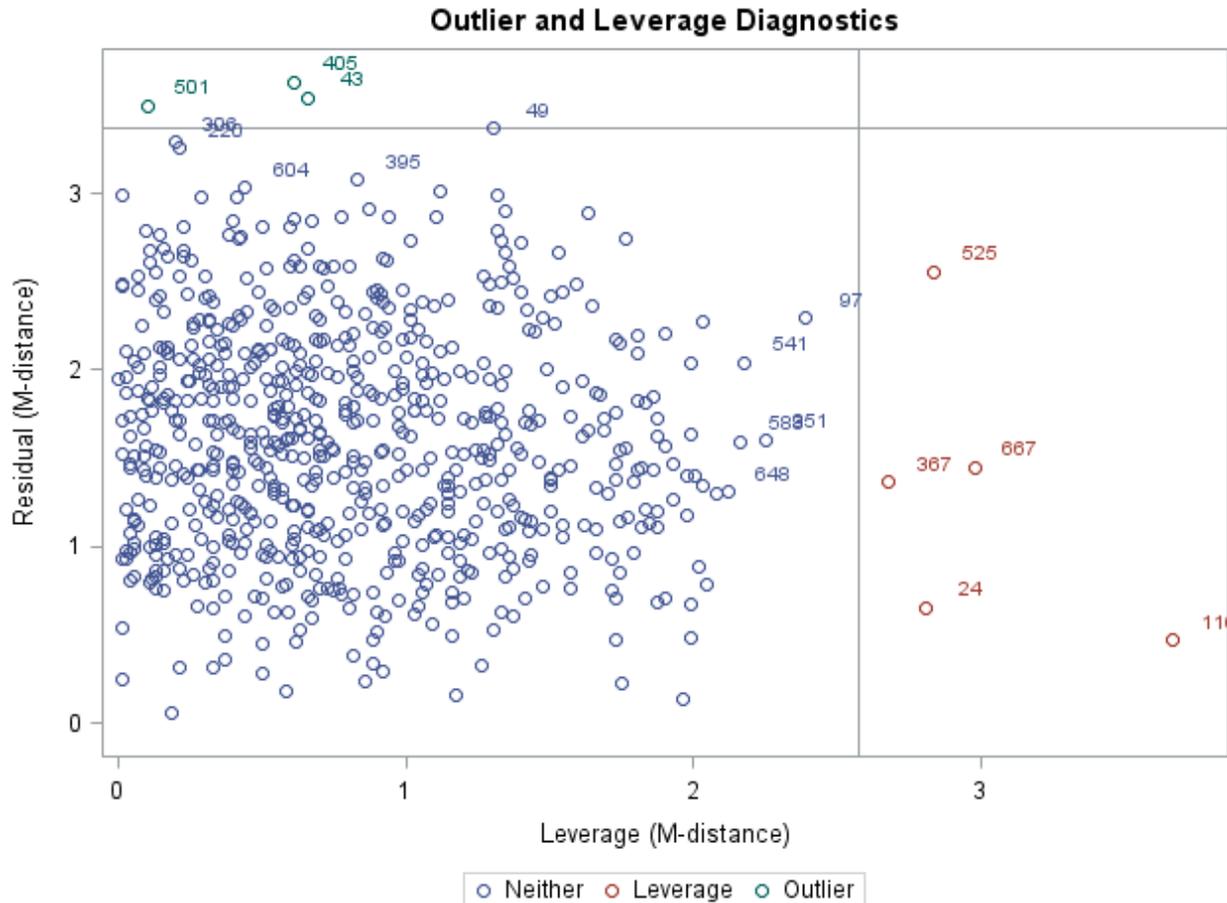


Figure 5: Case-Level Diagnostics for the Share Data

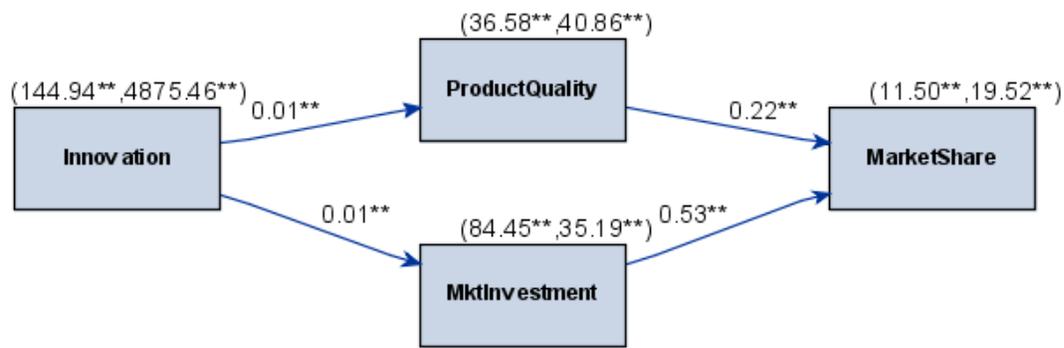
The outlier and leverage diagnostics plot (see Figure 5) gives a complete picture of observations that are residual outliers, bad leverage observations, and good leverage observations. The points on the right (in red) are good leverage. The points near the top (in green) are significant residual outliers. The top right corner, where there are no points, would identify bad leverage observations. There are 695 complete cases in the data, and the default $\alpha = 0.01$ would lead one to expect about 7 outliers and about 7 leverage observations by chance alone. In fact, there are 3 significant outliers and 5 significant good leverage observations.

The remaining diagnostics compare residuals to predicted values and to the theoretical (Chi) distribution. This is useful for identifying the ways in which significant outliers or leverage observations are unusual. Because this data set shows little, if any, evidence of problematic cases, those diagnostics are discussed later.

EXAMPLE 2: A LITTLE BIT OF BAD (LEVERAGE) INFLUENCE

The example below fits the same model but with a slight modification to the data. Can you find the problem?

```
proc calis data = share2 meanstr residual=norm plots=(residuals caseresidual);
  path
    MarketShare <-- ProductQuality MktInvestment,
    MktInvestment ProductQuality <-- Innovation;
  pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] arrange=flow;
run;
```



Chi-sq	10.29
DF	2
Pr > Chi-sq	0.01
CFI	0.98
SRMR	0.03
RMSEA	0.08

Figure 6: Path Diagram and Model Fit Statistics for the Share2 Data

The model fit is not nearly as satisfactory as it had been before. (See Figure 6.) An investigation of the residuals reveals a problem. (See Figures 7 and 8.)

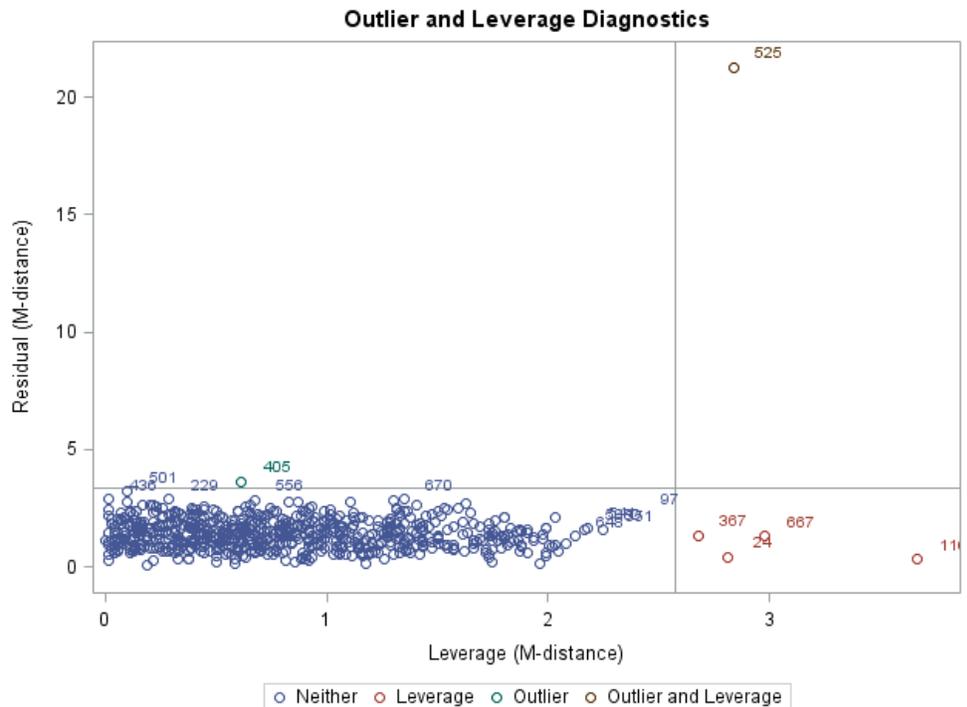


Figure 7: Case-Level Diagnostics for the Share2 Data

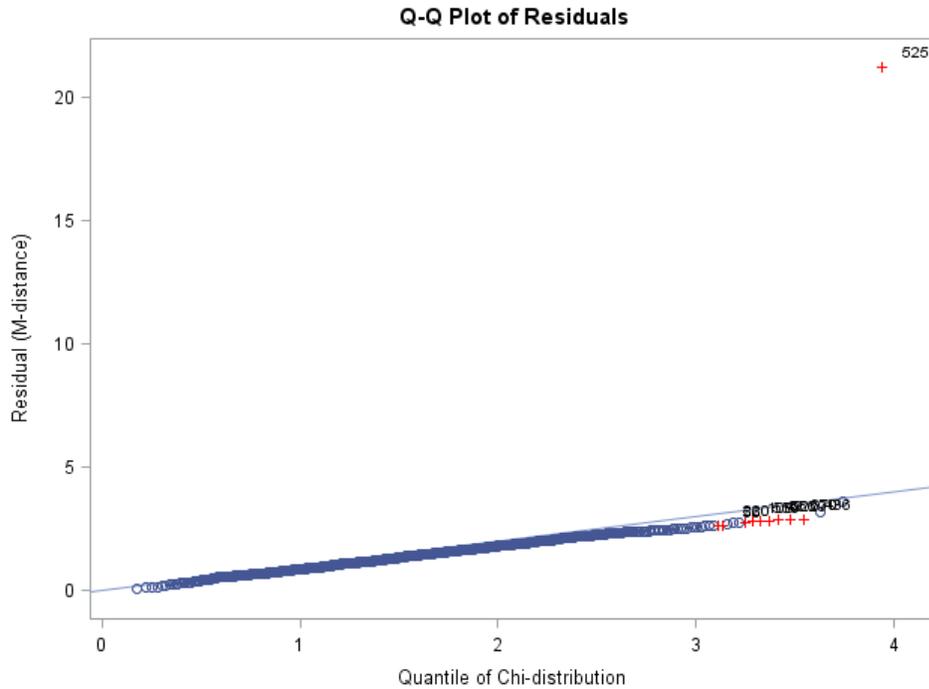
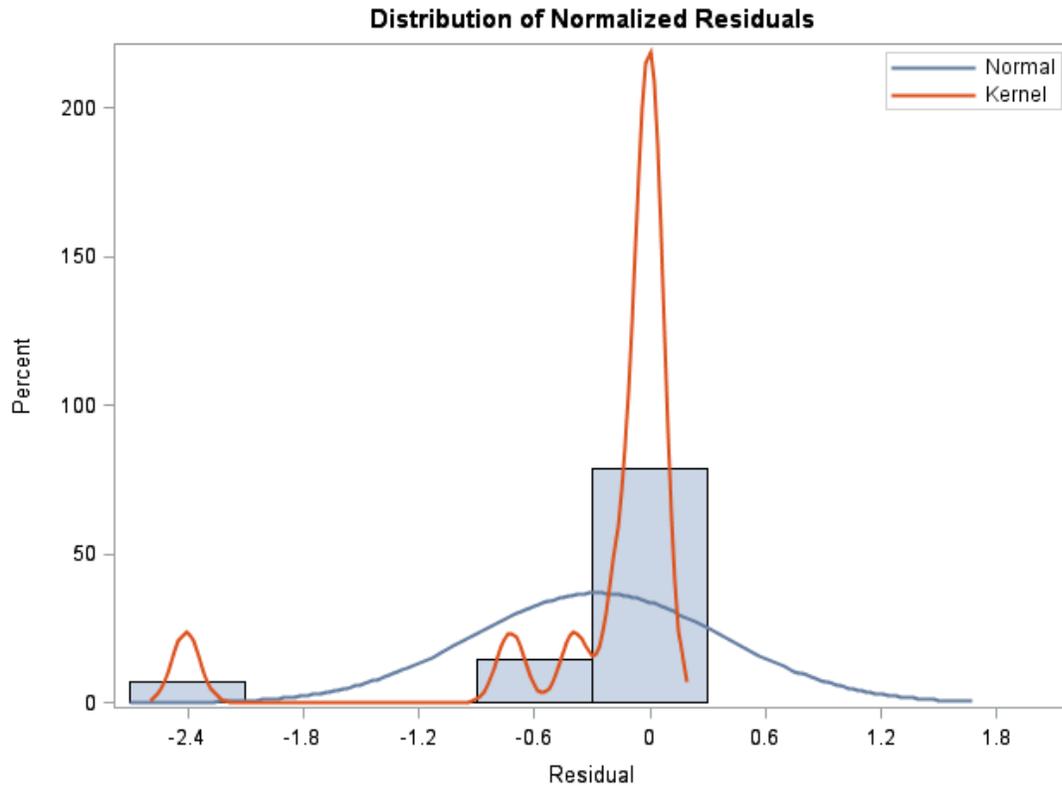


Figure 8: Q-Q Plot of Residuals for the Share2 Data

The observation in row 525 is both an outlier and a leverage observation—it is a bad leverage observation. The CALIS procedure gives quite a large number of graphs to help identify and diagnose problematic cases in the data, as well as tabular summaries of the largest case-level residuals and leverages.

Knowing that observation 525 is unusual, and that it is a bad leverage observation, the model residuals might reveal the nature of the inconsistency.



Rank Order of the 5 Largest Normalized Residuals		
Var1	Var2	Residual
MarketShare	Innovation	-2.40754
MktInvestment	ProductQuality	-0.72372
MarketShare	ProductQuality	-0.39308
MarketShare	MktInvestment	-0.15603
MarketShare	MarketShare	-0.14197

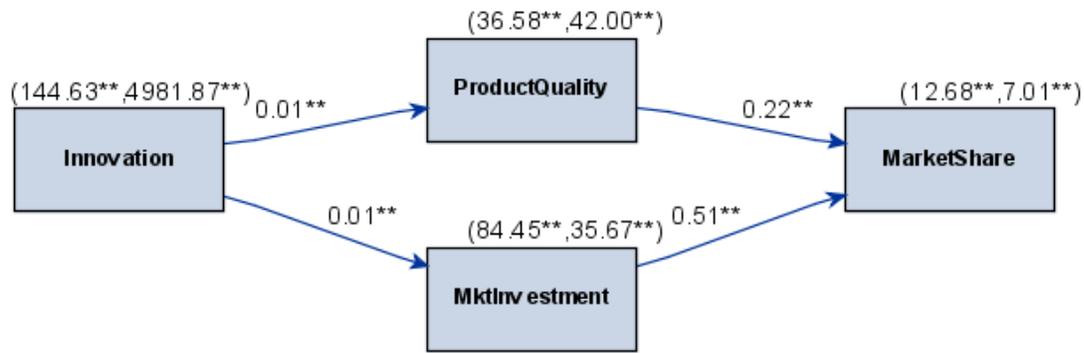
Figure 9: Model Residuals for the Share2 Data

The residual for the covariance between MarketShare and Innovation has the largest normalized residual. (See Figure 9.) A negative model residual indicates underprediction; the model-implied estimate of the covariance is too low, compared to the unrestricted covariance matrix.

This leverage case could be large enough that it masks other leverage cases. It most likely impacts parameter estimates, leading to incorrect conclusions. The problem calls for robust estimation.

Adding the option robust to the PROC CALIS statement results in the following. (See Figure 10.)

```
proc calis data = share2 meanstr robust residual=norm plots=(residuals caseresidual);
  path
    MarketShare <-- ProductQuality MktInvestment,
    MktInvestment ProductQuality <-- Innovation;
  pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] arrange=flow;
run;
```



Chi-sq	2.11
DF	2
Pr > Chi-sq	0.35
CFI	1.00
SRMR	0.01
RMSEA	0.01

Figure 10: Path Diagram and Fit for Robust Estimation of the Share2 Data

The bad leverage case is still evident. (See Figure 11.)

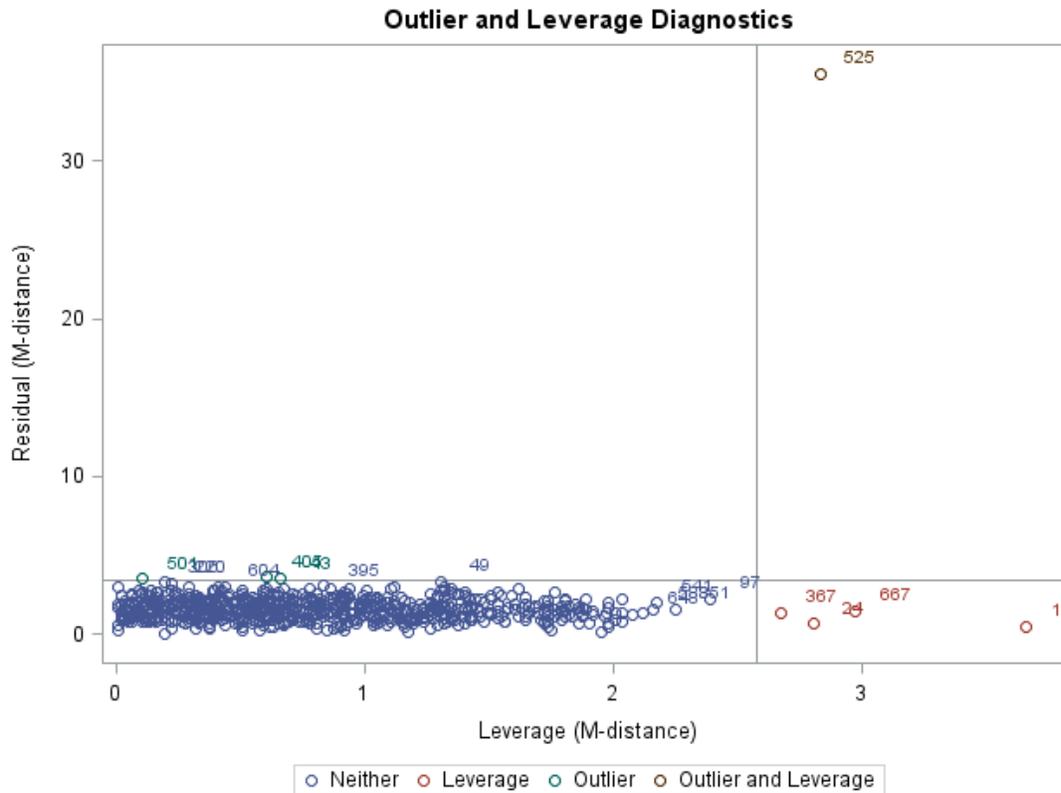


Figure 11: Case-Level Diagnostics for the Share2 Data with Robust Estimation

However, the model residuals give no cause for concern. (See Figure 12.)

Rank Order of the 5 Largest Normalized Residuals		
Var1	Var2	Residual
MarketShare	Innovation	-0.75122
MktInvestment	ProductQuality	-0.72615
MarketShare	ProductQuality	-0.48715
MarketShare	MarketShare	-0.24610
MarketShare	MktInvestment	-0.19938

Figure 12: Model Residuals for the Share2 Data with Robust Estimation

The robust estimation enabled the problematic case to be found while downweighting it in the analysis. It does not need to be removed, but it should certainly be investigated.

EXAMPLE 3: LEVERAGING TEST SCORES IN CONFIRMATORY FACTOR ANALYSIS

The real power and appeal of SEM for most analysts lies in its ability to include latent variables in models. This example shows a relatively simple latent variable model, a confirmatory factor analysis. The scores2 data set contains test scores for three math subtests and three verbal subtests. This confirmatory factor analysis specifies two correlated factors with three indicators each. For the case-level residuals, the outlier and leverage plot (PLOTS=resbylev) is requested.

```
proc calis data=scores2 residual=norm plots=(resbylev);
  path x1 x2 x3 <-- verbal,
        y1 y2 y3 <-- math;
  pvar verbal=1, math=1;
  pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] exogcov;
run;
```

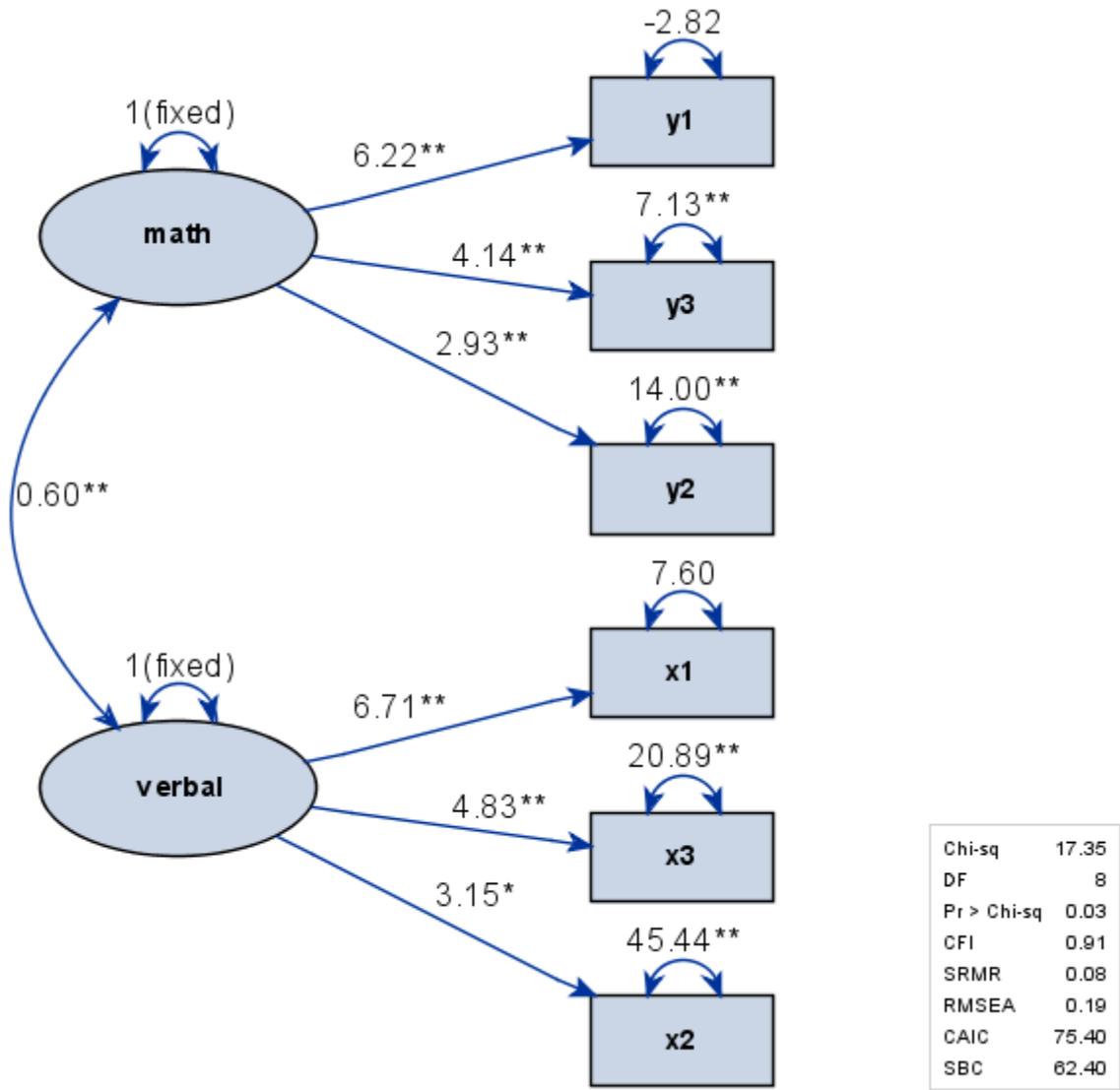


Figure 12: Path Diagram for the Scores2 Data, ML Estimation

The model results in Figure 12 show some evidence of poor fit.

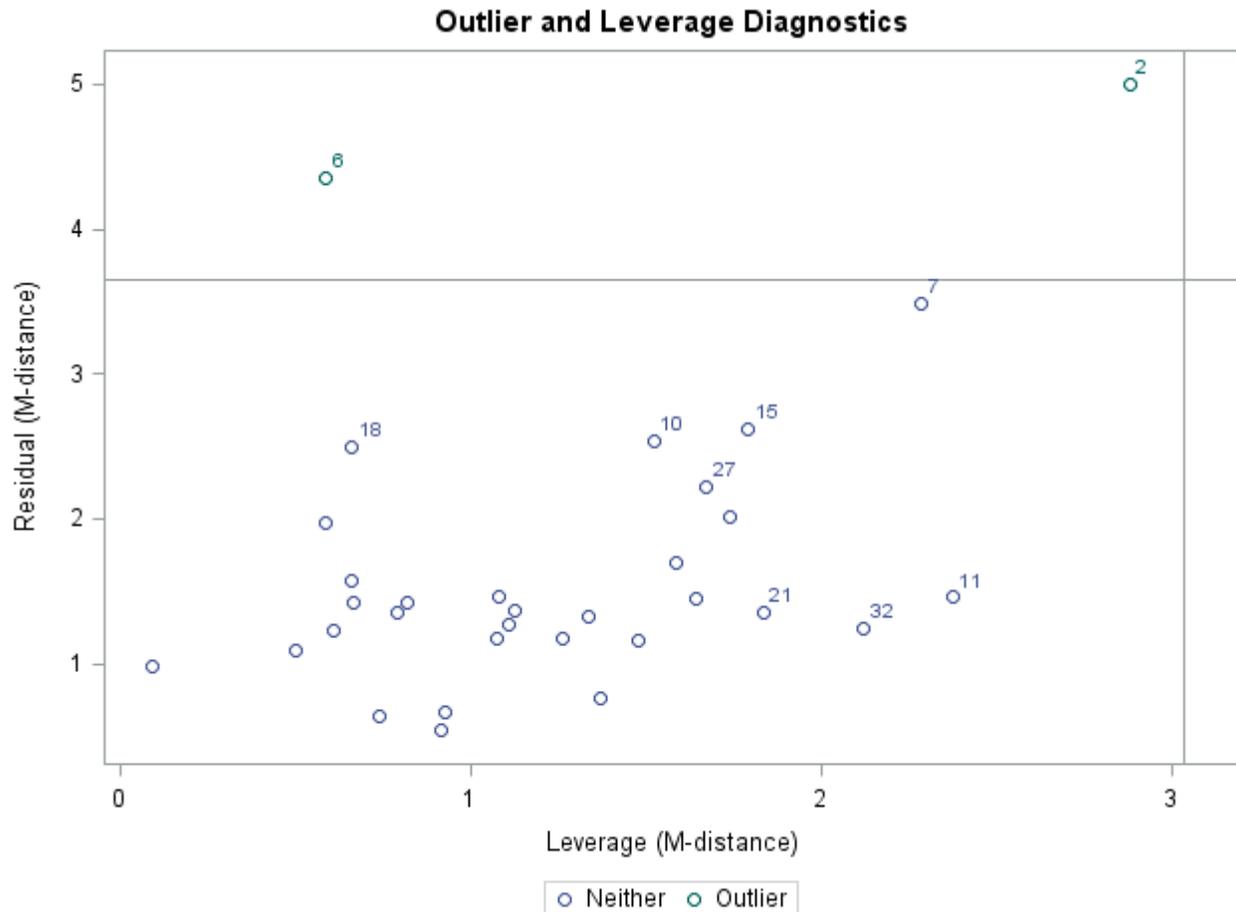


Figure 13: Outlier and Leverage Plot for the Scores2 Data, ML Estimation

Figure 13 shows two potentially problematic cases: row 2 and row 6. Investigation of the data reveals that these students had unusually inconsistent subscale scores, which could possibly indicate data errors, learning difficulty, environmental distraction during the test, or a host of other explanations. The statistical analysis cannot tell you why these cases are unusual, just that they are unusual.

For the purpose of understanding the factors, however, the cases should be downweighted and the factor coefficients estimated using the pattern of the other observations in the data. Rather than deleting cases 2 and 6 (because they could well be legitimate data values), a robust estimation should be requested. The results are displayed in Figure 14.

```
proc calis data=scores2 robust residual=norm plots=(residuals resbylev);
  path x1 x2 x3 <-- verbal,
        y1 y2 y3 <-- math;
  pvar verbal=1, math=1;
  pathdiagram fitindex=[chisq df probchi cfi rmsea srmsr] exogcov;
run;
```

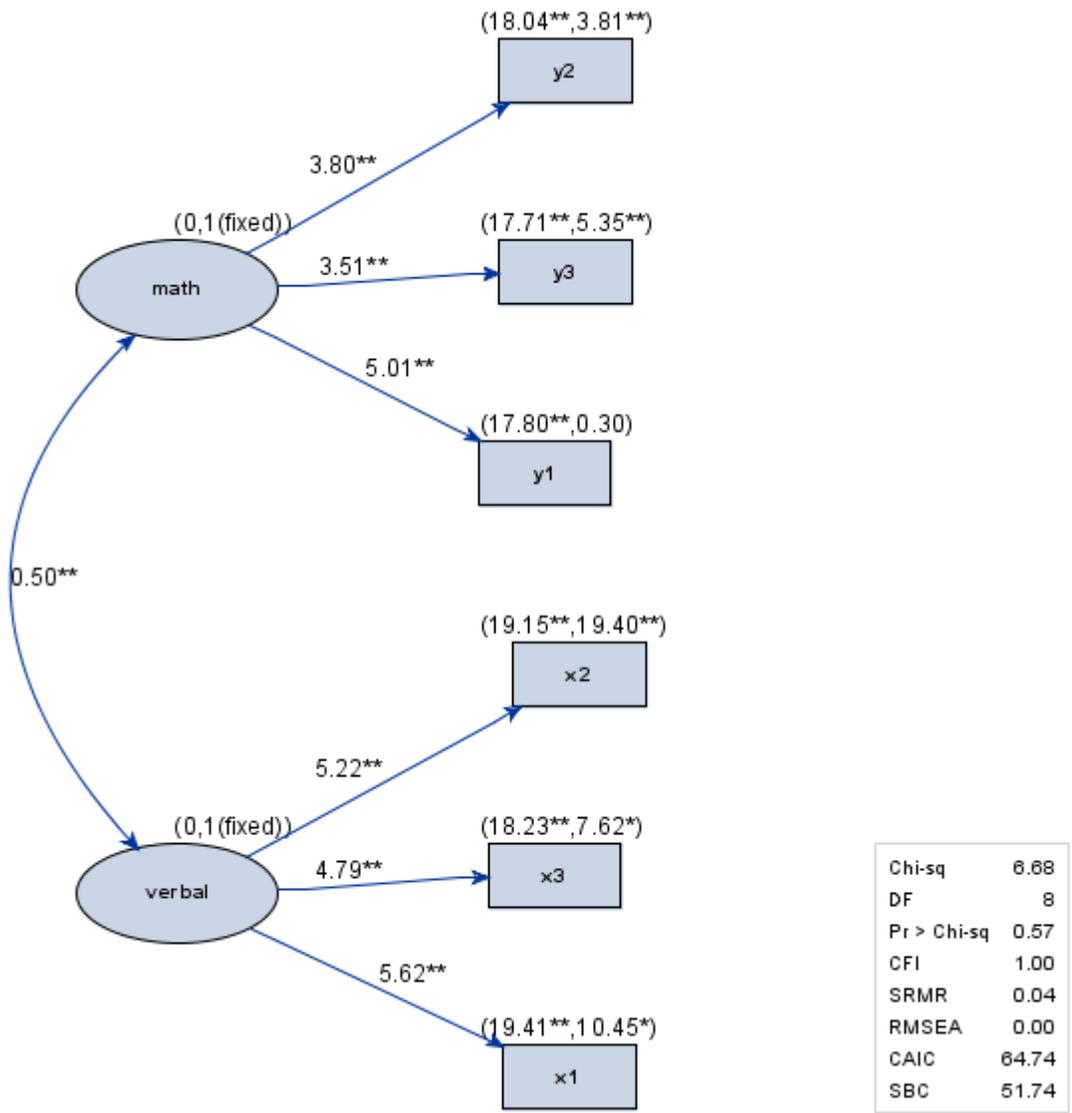


Figure 14: Path Diagram for the Scores2 Data, Robust Estimation

The model shows adequate fit. But more interestingly, the robust estimation has unmasked the effect of the outlier and leverage observations. See Figure 15.

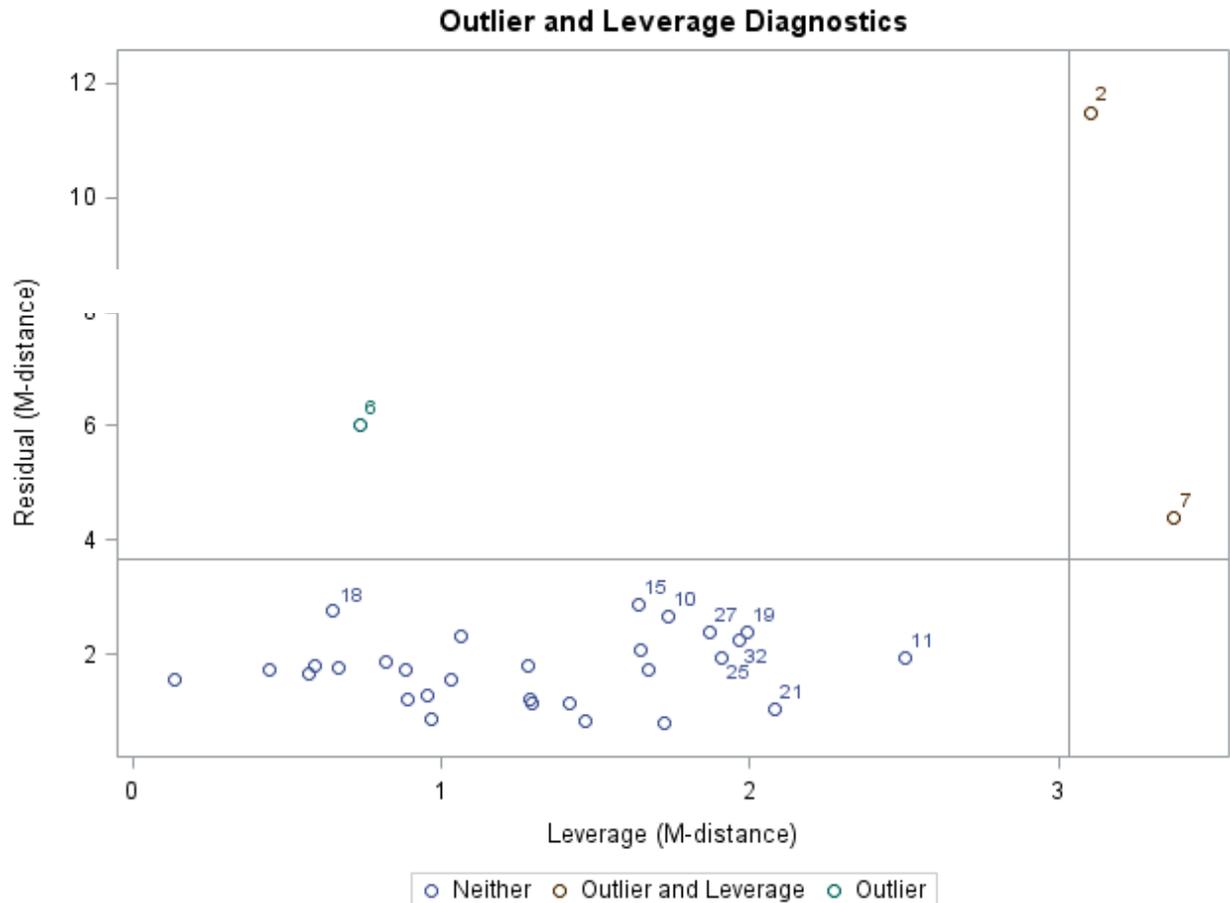


Figure 15: Outlier and Leverage Plot for the Scores2 Data, Robust Estimation

In Figure 15, observation #6 is still shown as an outlier, although observation #2 is shown as a leverage observation. Furthermore, by downweighting the extreme cases in estimation, an additional leverage case is revealed (#7). The researcher can investigate these three observations further and remain secure in the knowledge that these cases, while made easier to identify, did not lead to biased parameter estimates or incorrect inferences in the model because they are downweighted during estimation.

CONCLUSION

The new case-level residuals and robust estimation in the CALIS procedure give analysts an easy way of handling problematic cases in structural equation models. Because linear regression models are a special case of SEM, this approach can be applied to any regression-type model.

REFERENCES

- Lawley, D. N. and A. E. Maxwell. 1971. *Factor Analysis as a Statistical Method*. London: Butterworths.
- Rousseeuw, P. J., and B. C. Van Zomeren. 1990. "Unmasking Multivariate Outliers and Leverage Points." *Journal of the American Statistical Association* 85: 633–639.
- Yuan, K.-H., and K. Hayashi. 2010. "Fitting Data to Model: Structural Equation Modeling Diagnosis Using Two Scatter Plots." *Psychological Methods*. 15: 335–351.
- Yuan, K.-H., and X. Zhong. 2008. "Outliers, Leverage Observations, and Influential Cases in Factor Analysis: Using Robust Procedures to Minimize Their Effect." *Sociological Methodology*. 38: 329–368.
- Yung, Y.-F., and K.-H Yuan, K.-H. 2014. "Bartlett Factor Scores: General Formulas and Applications to Structural Equation Models." *New Developments in Quantitative Psychology: Springer Proceedings in Mathematics & Statistics*. 66: 385-401.

ACKNOWLEDGMENTS

The author would like to acknowledge Stephen Mistler, Jill Tao, and Cynthia Zender for their helpful comments on an earlier draft of this paper, and special gratitude is extended to Yiu-Fai Yung (developer of the CALIS procedure) for his ongoing support of the CALIS procedure, and of this author's efforts to tell the world about it.

RECOMMENDED READING

SAS Institute Inc. 2013. "The CALIS Procedure." *SAS/STAT® 13.1: User's Guide*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Catherine Truxillo
SAS Campus Dr.
Cary, NC 27513
SAS Institute Inc.
catherine.truxillo@sas.com
<http://www.sas.com>
blogs.sas.com/sastraining

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.