

SPSS[®]

white paper

Full-information missing data analysis with Amos¹

by

Werner Wothke, SmallWaters Corp., Chicago

and

James L. Arbuckle, Temple University, Philadelphia

¹ Reprinted from Faulbaum/Bandilla (1996), *SoftStat '95*. (Advances in Statistical Software 5), Stuttgart, Lucius and Lucius Inc., with permission from the publisher. Based on Arbuckle, J.L. (1996). Full information estimation in the presence of incomplete data. In G.A. Marcoulides and R.E. Schumaker [Eds.] *Advanced Structural Equation Modeling Techniques: Issues and Techniques*. Mahwah, NJ: LEA. Tables and figures reproduced with permission by Lawrence Erlbaum Associates.

Most multivariate methods require complete data, but most multivariate data are incomplete. Missing data are usually dealt with by listwise or pairwise deletion methods which aim to fix up the data so they can be analyzed by methods designed for complete data. This kind of approach is ad-hoc and has little theoretical justification. In contrast, a theory-based approach to the treatment of missing data under the assumption of multivariate normality, based on the direct maximization of the likelihood of the observed data, has long been known. The theoretical advantages of this full-information method are widely recognized, and its applicability in principle to structural equation modeling has been noted.

Unfortunately, theory has not had much influence on practice in the treatment of missing data. In part, the under-utilization of maximum likelihood (ML) estimation in the presence of missing data may be due to the unavailability of the method as a standard option in packaged data analysis programs. A (mistaken) belief may also exist that the benefits of using ML, rather than conventional missing data techniques, will in practice be small. In this paper, we will discuss how much more efficient estimation by full-information ML can be, when compared to the listwise or pairwise deletion methods.

Current practice in the treatment of missing data

The most commonly practiced methods for structural equation modeling (SEM) with missing data apply complete-data ML estimation to covariance matrices that have been somehow corrected. Such corrections can be listwise deletion (LD), which excludes from the calculations all records with missing values on any of the variables, and pairwise deletion (PD), by which each sample covariance between two variables is computed from pairwise complete data, excluding cases with missing values on one or both of the variables.

Little and Rubin (1987) reviewed these methods in the general multivariate case, and Brown (1994) studied their performance by Monte-Carlo simulation in a structural equation modeling context. Both references are critical of listwise and pairwise deletion methods, citing biased and/or inefficient estimates as well as the increased potential of obtaining indefinite sample covariance matrices. A third type of correction method, imputation of missing values, is well known in the statistical literature, but rarely used in structural equation modeling (Kim and Curry, 1977; Roth, 1994). We do not discuss imputation methods here, but we do outline a model-based imputation method that uses the results of full-information ML estimation will be outlined at the end of this paper.

Maximum likelihood estimation based on incomplete data

The principles of ML estimation with incomplete data are well known (Hartley and Hocking, 1971; Little and Rubin, 1987, 1989; Rubin, 1976; Wilks, 1932). Allison (1987) and Muthén, Kaplan and Hollis (1987) showed how the method applies to structural equation modeling. Unfortunately, their realizations are unpractical unless the data contain only a few distinct patterns of missing data, thus remaining of limited usefulness. They also require an exceptionally high level of technical expertise in the use of particular SEM programs. Currently, ML estimation with missing data is a standard option in at least two

structural equation modeling programs, Amos (Arbuckle, 1995) and Mx (Neale, 1994). Both maximize the case-wise likelihood of the observed data, computed by minimizing the function

$$C(\gamma) = \sum_{i=1}^N \log |\Sigma_{i,mm}| + \sum_{i=1}^N (\mathbf{y}_{i,m} - \mu_{i,m})' \Sigma_{i,mm}^{-1} (\mathbf{y}_{i,m} - \mu_{i,m}),$$

where $\mathbf{y}_{i,m}$ is the observed (or measured) portion of the data vector for case i , and $\mu_{i,m}$ and $\Sigma_{i,mm}$ are the corresponding mean vector and covariance matrix parameters. Thus, the two programs are not limited by the number of missing data patterns, and do not require the user to take elaborate steps to accommodate missing data.

Missing data mechanisms

In order to state the advantages of ML estimation over PD and LD, it is necessary to consider the mechanisms by which missing data can arise. Rubin (1976) and Little and Rubin (1987) distinguish missing data generating processes with respect to the information they provide about the unobserved data. Missing values of a random variable Y can be missing completely at random (MCAR), missing at random (MAR), or nonignorable.

The advantages of ML estimation over PD and LD are summarized by Little and Schenker (1995). For data that are missing completely at random, PD and LD estimates are consistent, although not efficient. If the data are only MAR, PD and LD estimates can be biased. ML estimates, on the other hand, are already both consistent and efficient when the data are only MAR. In addition, some authors have suggested that ML estimates will tend to show less bias than estimates based on PD or LD even when the data deviate from MAR (Little and Rubin, 1989; Muthén, Kaplan and Hollis, 1987). As a final shortcoming, PD does not provide standard errors of parameter estimates or tests of model fit.

Bootstrap Simulation 1: MCAR data

A bootstrap simulation was conducted to demonstrate the efficiency of ML estimation relative to PD and LD for a single, fairly typical estimation problem when the data are MCAR. The scores of six psychological tests administered to 145 sixth and seventh graders by Holzinger and Swineford (1939) make up the bootstrap population. There are three verbal tests (word meaning, sentence completion and paragraph comprehension) and three tests of spatial ability (lozenges, cubes, and visual perception). The path diagram of Figure 1 displays the postulated factor model and the standardized ML parameters estimated from the original Holzinger and Swineford data. For purposes of the bootstrap simulation, these are the population parameters.

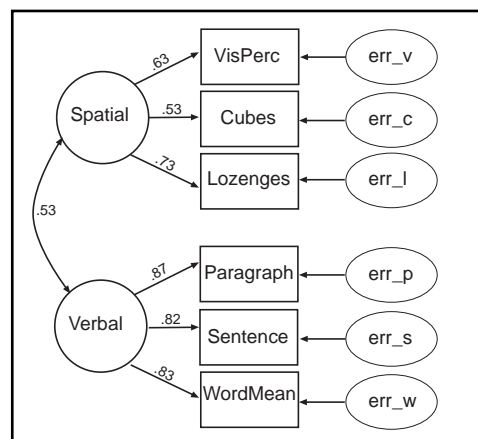


Figure 1. Factor model of Holzinger and Swineford data; standardized parameters.

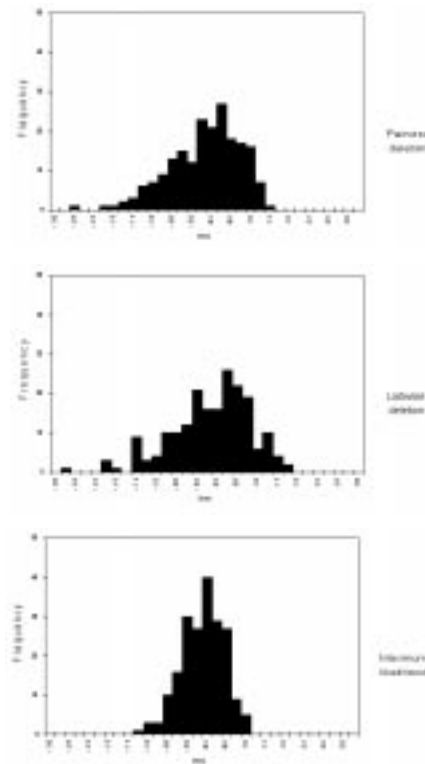


Figure 2. Estimation errors of the standardized regression coefficient for PARAGRAPH (MCAR); N=145, 20% missing data

parameter values. Figure 2 displays the distributions of estimation error for the standardized regression coefficient of PARAGRAPH obtained by each of the three estimation methods². The size of the estimation error varies, with mean square errors (MSEs) of 0.005121 and 0.001859 for PD and ML. Since the distributions are centered almost exactly at zero, we can draw two conclusions. First, estimation bias by all three methods is negligible, as expected with MCAR data. Secondly, the MSEs can be interpreted as sampling variances, indicating differences in the precision (or efficiency) of the estimates. Under the assumption that the sampling variance of means and regression coefficients is inversely related to sample size, the relative mean square errors can be used to express the relative gains in efficiency. In the present case, switching from PD to ML provides an improvement in precision equivalent to increasing the sample size by a factor of $2.75 = 0.005121/0.001859$. That is, in order to estimate the regression weight from VERBAL to PARAGRAPH as precisely with PD as with ML, the sample size would have to be increased from N=145 to N=399.

The simulation study involved two design factors — sample size was varied at two levels (N=145 or N=500), and the missing data rate at five levels (0%, 5%, 10%, 20%, or 30%). Each of the 10 (= 2 x 5) paired conditions was represented with a fresh set of 200 samples from the bootstrap population, drawn randomly with replacement. In each of the bootstrap samples, an MCAR process was employed to set the pre-specified number of values to missing. The factor model shown in Figure 1 was fitted to each bootstrap sample using ML, LD, and PD methods. Altogether, the simulation comprised 2,000 bootstrap samples and 6,000 attempts to fit the factor model.

The performance of a single estimation method (say, ML) under a particular combination (N=145 with 20% missing data, say) is assessed in the following way. First, the method is applied to estimate the factor model for each of 200 bootstrap samples. The accuracy of the (standardized) estimates then is judged by comparing them to the bootstrap population parameters in Figure 1. The question is which, if any, of the three methods comes closest to all the

² The LD method produces model estimates for only 195 of the 200 samples.

Of course, the relative MSE of 2.75 just referred to is specific to a single parameter, a single sample size, and a single missing data rate. Figure 3 shows relative MSEs for all six standardized regression weights under a variety of conditions. One of the points (indicated by an arrow) in Figure 3 represents the relative MSE of 2.75 that was just discussed. The points in Figure 3 are affected by an undetermined amount of sampling error. Nevertheless, some broad trends are apparent. Most importantly, the relative MSE increases as the missing data rate increases. It is also clear that the relative MSE is different for different parameter estimates. In particular, the relative MSEs for the verbal tests are larger than for the spatial tests. One possible explanation is that, by using all information in the data, the ML method yields more precise estimates for parameters associated with highly predictable variables. This has been found in previous simulation studies with complete data (Anderson and Gerbing, 1984; Boomsma, 1985) and seems to be the case here as well. In the (bootstrap) population, the squared multiple correlations range between 0.58 and 0.62 for the three verbal tests, but only between 0.20 and 0.32 for the three spatial tests.

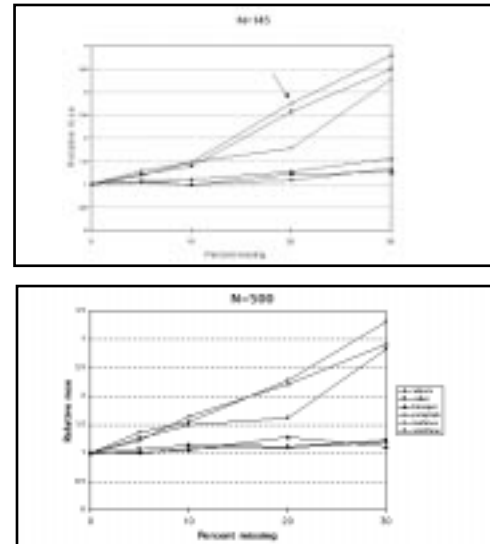


Figure 3. Relative mean square errors (PD vs. ML) for standardized regression weights

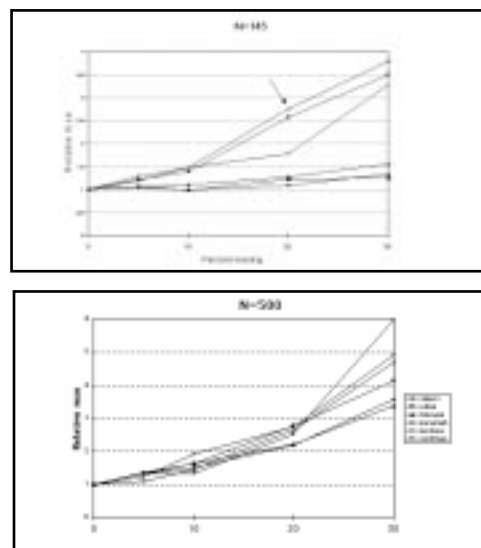


Figure 4. Relative mean square errors (LD vs. ML) for standardized regression weights

Figure 4 shows the relative MSEs for LD compared to ML. The number of cases retained for analysis by LD falls rapidly as the missing data rate increases. For example, with 30% missing data and a sample size of $N=145$, one would expect on average only $145(1-0.3)^6 = 17.06$ cases with complete observations. With such small effective sample sizes, it is not surprising to encounter cases with severe numerical problems during model estimation (Wothke, 1994). With 30% missing data and a sample size of 145, a solution was obtained for only 165 of the 200 samples. With 20% missing data and $N=145$, a solution was obtained for 195 of the 200 samples. In contrast, ML solutions were successfully obtained for every sample. Figure 4 is based only on the samples in which an LD solution was obtained.

Figure 4 indicates even higher gains in efficiency of the ML method when the alternative is LD. The relative MSE stays above 2.0 when 20% of the data are missing and above 3.0 with 30% missing data. The top panel of Figure 4 also shows an unusually large relative MSE for the SENTENCE regression weight with 30% missing data and N=145. This particular result appears to be due to the influence of four samples for which the estimate was exceptionally far from its population value.

Discussion

It is impossible to put a single figure on the gain in accuracy of estimation to be had by abandoning PD and LD in favor of ML. The advantage of ML depends on the missing data rate, the size of the sample, and it differs from one parameter to another. Nevertheless, the MCAR simulation demonstrates that ML can be superior to PD and LD by a wide margin.

Bootstrap Simulation 2: Extreme Case of MAR data

A second simulation was performed to illustrate the benefits of ML with data that are MAR but not MCAR. The same Holzinger and Swineford (1939) data were used for the simulation, except that a different missing data process was employed: SENTENCE and PARAGRAPH scores were set to missing whenever the WORDMEAN score was 12 or less, and retained for examinees with WORDMEAN scores greater than 12. The threshold of 12 was chosen so as to delete approximately 30% of the SENTENCE and PARAGRAPH score. No scores were deleted on the VISPERC, CUBES, LOZENGES or WORDMEAN tests.

This type of MAR process emulates the situation in which an examiner administers the WORDMEAN test first, and then administers the other verbal tests only to the examinees who “passed” some minimum score requirement on the WORDMEAN test. It is particularly easy to see how LD would lead to biased estimates in this situation, as selectively removing records with low WORDMEAN scores would affect both means and covariances of the remaining sample. The effect of the missing data pattern on PD is not so clear cut. We have already observed the efficiency of ML estimates in Simulation 1 where, because the data were MCAR, bias in the estimates was relatively small. By contrast, because the present data are only MAR, estimation bias is now of central concern.

Table 1 shows the error (and its standard error in parentheses) for several estimates averaged across 200 samples of size 145. For almost every estimate, ML has the smallest mean error. In rare cases when ML has a numerically higher mean error than PD or LD, the size of the standard error indicates that the difference is due to sampling variation. For several parameters, the mean error is dramatically smaller under ML estimation than under PD or LD.

Parameter	Mean error (s.e.)		
	ML	LD	PD
Regression Weights			
visperc	.000 (.000)	.000 (.000)	.000 (.000)
cubes	.000 (.010)	.040 (.010)	.000 (.010)
lozenges	.070 (.007)	.330 (.000)	.114 (.007)
paragraph	.000 (.000)	.000 (.000)	.000 (.000)
sentence	-.100 (.010)	-.100 (.010)	-.100 (.010)
wordmean	.001 (.000)	-.107 (.000)	-.077 (.000)
Intercepts			
visperc	-.014 (.040)	.000 (.000)	-.014 (.040)
cubes	.000 (.000)	.000 (.000)	.000 (.000)
lozenges	-.010 (.040)	1.007 (.000)	-.010 (.040)
paragraph	-.000 (.000)	1.040 (.000)	1.040 (.000)
sentence	.000 (.000)	1.070 (.000)	1.070 (.000)
wordmean	-.017 (.040)	3.440 (.000)	-.017 (.040)
Covariance			
	-.000 (.000)	-1.070 (.000)	-.010 (.000)
Standardized Regression Weights			
visperc	-.011 (.007)	-.004 (.000)	-.010 (.007)
cubes	.000 (.000)	-.000 (.000)	-.010 (.000)
lozenges	-.004 (.000)	-.001 (.000)	-.004 (.000)
paragraph	-.000 (.000)	-.001 (.000)	-.001 (.000)
sentence	-.000 (.000)	-.000 (.000)	-.040 (.000)
wordmean	.010 (.000)	-.000 (.000)	-.100 (.000)
Correlation			
	.010 (.000)	-.001 (.000)	.000 (.000)

Table 1. Mean error (s.e.) of parameter estimates under MAR sampling from the Holzinger and Swineford population, N=145

Although the present simulation demonstrates that ML estimates are less biased than PD or LD estimates, it is also a fact that some of the mean errors under ML remain large compared to their standard errors. Thus, while giving less biased estimates than PD and LD, ML apparently did not completely compensate for the bias introduced by the missing data process.

Inspection of the Holzinger and Swineford sample revealed that the data deviate from multivariate normality. In particular, the relation between SENTENCE and WORDMEAN is distinctly nonlinear. It was therefore suspected that the significant mean errors for ML in Table 1 are due to the inability of the factor model to accommodate nonlinearities in the data. To confirm this explanation, the simulation was repeated by parametric bootstrap from the multivariate normal distribution, using the parameters obtained by fitting the factor model of Figure 1 to the Holzinger and Swineford data. In this additional simulation, the ML estimates came out virtually unbiased, while the PD and LD estimates remained substantially biased.

Imputation of missing values

It is not necessary either to impute values for missing data or to estimate the population moments as a prerequisite to model fitting by ML. These are optional steps which, if they are performed at all, are best done after the model is fitted, not before. Estimates of population means, variances and covariances, calculated from parameter estimates under the assumption of a correct model, are reported by most structural modeling programs.

Let μ^* and Σ^* be the population means and covariances of all variables in the model, both measured and unmeasured, and let $\hat{\mu}^*$ and $\hat{\Sigma}^*$ be their estimates assuming a correct model. For an individual case i , let the partitioned data vector \mathbf{y}_i contain all unobserved and observed data of the model, with unobserved variables ordered first

$$\mathbf{y}_i = (u_{i,1} \ u_{i,2} \ \dots \ u_{i,p} \ | \ m_{i,1} \ m_{i,2} \ \dots \ m_{i,q}) = [\mathbf{y}_{i,u} \ | \ \mathbf{y}_{i,m}] . \quad (1)$$

The subvectors $\mathbf{y}_{i,u}$ and $\mathbf{y}_{i,m}$ will be of different sizes for different missing data patterns. Arranging the values in $\hat{\mu}^*$ and $\hat{\Sigma}^*$ in the same order as in \mathbf{y}_i yields the partitioned implied mean vector

$$\hat{\mu}_i = (\hat{\mu}_{i,u} \ | \ \hat{\mu}_{i,m}) \quad (2)$$

and covariance matrix

$$\hat{\Sigma}_i = \begin{pmatrix} \hat{\Sigma}_{i,uu} & | & \hat{\Sigma}_{i,um} \\ \hline \hat{\Sigma}_{i,mu} & | & \hat{\Sigma}_{i,mm} \end{pmatrix} \quad (3)$$

Under normality, the expectation of the missing data, conditional on the observed values, is estimated as

$$\overline{\mathbf{E}(\mathbf{y}_{i,u} \ | \ \mathbf{y}_{i,m})} = \hat{\mu}_{i,u} + \hat{\Sigma}_{i,um} \hat{\Sigma}_{i,mm}^{-1} (\mathbf{y}_{i,m} - \hat{\mu}_{i,m}) , \quad (4)$$

and their conditional covariance matrix as

$$\overline{\text{Cov}(\mathbf{y}_{i,u} \ | \ \mathbf{y}_{i,m})} = \hat{\Sigma}_{i,uu} - \hat{\Sigma}_{i,um} \hat{\Sigma}_{i,mm}^{-1} \hat{\Sigma}_{i,mu} . \quad (5)$$

These statistics can be used to impute the model-based means and confidence intervals of the missing data, given the observed portion of the data. With complete data, using (4) produces the usual regression estimates of factor scores provided by many structural modeling programs. Stochastic regression imputation (Little and Rubin, 1989) and multiple imputation (Little and Rubin, 1987) are important variants of this imputation method.

Conclusion

Maximum likelihood estimation with incomplete data is feasible and should be the preferred method of treating missing data when the alternative is pairwise or listwise deletion. Maximum likelihood's lack of reliance on the MCAR requirement is a feature that remains to be fully exploited. It should not be overlooked that structural modeling with the Amos program can be used to solve missing data problems that arise in conventional analyses, such as regression with observed variables or the simple estimation of means and variances.

About the authors

James L. Arbuckle is an Associate Professor of Psychology at Temple University. Professor Arbuckle is the author of the Amos structural equation modeling program.

Werner Wothke is president of SmallWaters Corp., a statistical software and consulting firm located in Chicago, Illinois. SmallWaters publishes Amos, and offers a line of supporting books, training and consulting services in structural equation modeling. SmallWaters and SPSS maintain a cooperative agreement, so that Amos can be licensed from either company. Additional information and an Amos demo version can be obtained from the world wide web at <http://www.smallwaters.com>.

References

- Allison, P.D. (1987) Estimation of linear models with incomplete data. In C.C. Clogg [Ed.] *Sociological Methodology, 1987*. San Francisco: Jossey-Bass, 71–103
- Anderson, J.C. and Gerbing, W.D. (1984) The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173
- Arbuckle, J.L. (1995) *Amos for Windows. Analysis of moment structures*. Version 3.5. Chicago: SmallWaters Corp.
- Boomsma, A. (1985) Nonconvergence, improper solutions, and starting values in Lisrel maximum likelihood estimation. *Psychometrika, 50*, 229–242
- Brown, R.L. (1994) Efficacy of the indirect approach for estimating structural equation model with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal, 1*, 287–316

- Hartley, H.O. and Hocking, R.R. (1971) The analysis of incomplete data. *Biometrics*, 27, 783–823
- Holzinger, K.J. and Swineford, F.A. (1939) A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monographs, No. 48*. Chicago: University of Chicago, Department of Education
- Kim, J.-O. and Curry, J. (1977) The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, 6, 215–240
- Little, R.J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data*. New York: Wiley
- Little, R.J.A. and Rubin, D.B. (1989) The analysis of social science data with missing values. *Sociological Methods and Research*, 18, 292–326
- Little, R.J.A. and Schenker, N. (1995) Missing data. In G. Arminger, C.C. Clogg and M.E. Sobel [Eds.] *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum, 39–75
- Muthén, B., Kaplan, D. and Hollis, M. (1987) On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462
- Neale, M.C. (1994) *Mx: Statistical modeling, 2nd Edition*. Box 710 MCV, Richmond, VA 23298: Department of Psychiatry
- Roth, P.L. (1994) Missing data: A conceptual view for applied psychologists. *Personnel Psychology*, 47, 537–560
- Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 61, 581–592
- Wilks, S.S. (1932) Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, 3, 163–195
- Wothke, W. (1994) Nonpositive definite matrices in structural modeling. In K.A. Bollen and J.S. Long [Eds.] *Testing Structural Equation Models*. Newbury Park, CA: Sage Publications

About SPSS

SPSS Inc. is a multinational software products company that delivers statistical product and service solutions for survey research, marketing and sales analysis, quality improvement, scientific research, government reporting and education. Primary product lines include: SPSS for a variety of business solutions, SYSTAT and BMDP for scientific analysis, and QI Analyst for manufacturing and quality improvement applications. More than 2 million people worldwide use SPSS products.

Chicago-based SPSS has sales and support offices and distributors worldwide. In 1995, SPSS completed the best year in its 28-year history with total revenues of \$63 million.

SPSS software operates on most models of all major computers. It is widely used on personal computers running Microsoft[®] Windows[®] and Windows 95. Versions for the Power Macintosh[®] and many UNIX[®] platforms are also available. In addition, many products are offered in Catalan, French, German, Italian, Japanese, Spanish and traditional Chinese.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at <http://www.spss.com>

SPSS Inc. United States and Canada	+1.312.329.2400 Toll-free: 1.800.543.2185	SPSS Italia srl	+39.51.252573
SPSS Federal Systems (U.S.)	+1.703.527.6777	SPSS Japan Inc.	+81.3.5474.0341
SPSS Argentina srl.	+541.816.4086	SPSS Korea	+82.2.552.9415
SPSS Asia Pacific Pte. Ltd.	+65.3922.738	SPSS Latin America	+1.312.494.3226
SPSS Australasia Pty. Ltd.	+61.2.9954.5660 Toll-free: 1800.024.836	SPSS Malaysia Sdn Bhd	+60.3.704.5877
SPSS Belgium	+32.162.389.82	SPSS Mexico Sa de CV	+52.5.575.3091
SPSS Benelux	+31.183.636711	SPSS Middle East and Southeast Asia	+971.4.525536
SPSS Central and Eastern Europe	+44.(0)1483.719200	SPSS Scandinavia AB	+46.8.102610
SPSS East Mediterranean and Africa	+972.9.526700	SPSS Schweiz AG	+41.1.201.0930
SPSS France SARL	+33.1.4699.9670	SPSS Singapore Pte.	+65.2991238
SPSS Germany	+49.89.4890740	SPSS Taiwan Corp.	+886.2.5771100
SPSS Hellas SA	+30.1.7251925	SPSS UK Ltd.	+44.1483.719200
SPSS Hispanoportuguesa S.L.	+34.1.447.37.00		
SPSS Ireland	+353.1.66.13788		
SPSS Israel Ltd.	+972.9.526700		

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.