Tilburg University

lavaan.survey

Oberski, D.L.

lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models

Daniel Oberski

Tilburg University

Abstract

This paper introduces the R package **lavaan.survey**, a user-friendly interface to design-based complex survey analysis of structural equation models (SEMs). By leveraging existing code in the **lavaan** and **survey** packages, the **lavaan.survey** package allows for SEM analyses of stratified, clustered, and weighted data, as well as multiply imputed complex survey data. **lavaan.survey** provides several features such as SEMs with replicate weights, a variety of resampling techniques for complex samples, and finite population corrections, features that should prove useful for SEM practitioners faced with the common situation of a sample that is not *iid*.

Keywords: complex survey analysis, structural equation modeling, clustering, stratification, sampling weights, multiple imputation, resampling, jackknife, bootstrap, replicate weights, R.

1. Introduction

Structural equation models (SEMs) constitute a popular framework for formulating, fitting, and testing an abundant variety of models for continuous interval-level data in a wide range of fields. Special cases of structural equation modeling include factor analysis, (multivariate) linear regression, path analysis, random growth curve and other longitudinal models, errors-in-variables models, and mediation analysis (Bollen 1989; Kline 2011). The main development of structural equation modeling has been in social science fields such as psychology (Ullman and Bentler 2003), education (Kaplan 2008), and sociology (Duncan 1975; Saris and Stronkhorst 1984), while more recently structural equation modeling is finding applications in other fields such as ecology and biology (Grace 2006) and neuroscience (McIntosh and Gonzalez-Lima 1994; Roelstraete and Rosseel 2011).

While classical SEM theory assumes independently and identically distributed (*iid*) observations (Bollen 1989), applications often require the analysis of data from complex surveys

that may involve stratification, clustering, and unequal selection probabilities, violating this assumption (Skinner, Holt, and Smith 1989; Muthén and Satorra 1995, p. 281). For example, Marsh and Hau (2004) explained the relations between academic self-concepts and achievements in a 26-country complex multistage survey. Outside of the realm of complex surveys clustering may also occur, for instance in Byrnes *et al.* (2011)'s analysis of the effect of storms on kelp forest food webs, where variables such as kelp density and species richness are likely correlated across sites that are geographically close to each other. It is well-known that under complex sampling, both point and variance estimators derived under *iid* assumptions may produce biased and inconsistent estimates (Cochran 1977; Skinner *et al.* 1989). This finding was reproduced for SEM parameter estimates by Kaplan and Ferguson (1999) and Asparouhov and Muthén (2005). Hahs-Vaughn and Lomax (2006) analyzed student data from the Beginning Postsecondary Students Longitudinal study to explain college experiences and learning outcomes with pre-college traits, showing that SEM parameter estimates, standard errors, and fit measures can change dramatically when complex sampling is taken into account.

Adjustments to point and variance estimators for SEMs under complex sampling were discussed by Muthén and Satorra (1995) and Stapleton (2006), and estimation using pseudo-maximum likelihood procedures by Asparouhov (2005, 2006) and Asparouhov and Muthén (2005). For an overview of literature related to complex sampling in structural equation modeling, see Bollen, Tueller, and Oberski (2013). These procedures have since been implemented in standard closed-source commercial software for SEMs: **LISREL** (Jöreskog and Sörbom 2006), **Mplus** (Muthén and Muthén 2012), **EQS** (Bentler 2008), and Stata (Stata-Corp. 2011a,b). Another popular commercial program, **AMOS** (Arbuckle 2011), does not implement complex sampling estimation at the date of writing.

None of the open-source SEM packages, **sem** (Fox 2006; Fox, Nie, and Byrnes 2012), **OpenMx** (Boker *et al.* 2011), and **lavaan** (Rosseel 2012), directly implement complex survey adjustments. These packages do provide enough flexibility to allow for such adjustments through resampling methods if the user is willing to program these (the **sem** manual provides some guidance to this effect). More user-friendly interfaces are currently not available. Furthermore, with the exception of Stata and **Mplus**, the commercial packages that do implement estimation procedures for complex sampling still omit features dealing with several complications that may arise in the analysis of complex surveys:

- Some secondary data sources such as the OECD's Programme for International Student Assessment (PISA) do not provide the sampling design variables directly, but instead provide a set of so-called "replicate weights" (OECD 2009). In principle this represents a considerable simplification of highly complex survey analysis (Brick, Morganstein, and Valliant 2000). Currently, however, not all SEM software allows for adjustments of SEM estimators using replicate weights;

- More generally, variance estimation of SEM parameters with complex sampling using resampling methods such as the jackknife and bootstrap are not implemented directly but require additional programming on the part of the user (see Stapleton 2008, for a discussion of these methods in the context of SEMs);

- Structural equation modeling is primarily an analytic method, so that finite population corrections may not usually be relevant (e.g., Fuller 2009, p. 342). However, structural equation modeling is also a flexible method of reformulating several descriptive methods

for which the finite population may be of interest, such as domain mean and model-based small area estimation. Currently finite population corrections, which may be relevant for these purposes, are not available in all SEM programs.

The purpose of this article is to introduce the **lavaan.survey** package (Oberski 2013a) for the R environment (R Core Team 2013), which serves to bring user-friendly complex survey SEM analysis to the open source SEM implementation **lavaan**. In addition, by leveraging the many features of the **survey** package (Lumley 2004, 2010, 2012b) it provides users with the above features currently omitted from some commercially available SEM software packages. Thanks to code reuse and the flexibility of the **survey** and **lavaan** packages, the **lavaan.survey** package is able to provide an extremely flexible, user-friendly, and open source framework for design-based analysis of complex survey data using SEM. It also allows for the analysis of multiply imputed complex survey data (Little and Rubin 1987; Graham and Hofer 2000). At the time of writing, a limitation of the package is that it deals with the continuous case only. The package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=lavaan.survey`.

Section 2 discusses the theory of structural equation modeling in general and SEM under complex sampling in particular. After a brief overview of the package in Section 3, Sections 4.1, 4.2, 4.3, and 4.4 demonstrate the usage of the package by applying it to SEM analyses arising from the literature.

2. Technical explanation

Different methods have been suggested to deal with complex sampling in SEMs. In this article we will only deal with "aggregate" design-based methods (see Skinner *et al.* 1989, p. 8; Muthén and Satorra 1995). "Design-based" refers to the fact that inferences are based on the theoretical distribution of all possible samples under a particular survey design. Such a basis for inference stands in contrast to the "model-based" approach, which derives point and variance estimators from the assumed model. In practice, the two may sometimes coincide (see Sterba 2009, for an overview). Three aggregate design-based point estimators have been suggested in the literature: adjustment of the weights or sample size to an effective sample size (Stapleton 2002), pseudo-maximum likelihood (Muthén and Satorra 1995; Asparouhov 2005, 2006), and weighted least squares estimation (Skinner *et al.* 1989, p. 86; Vieira and Skinner 2008); see Stapleton (2006) for an overview of these approaches. For these point estimators, different variance estimation methods are possible, including linearization (Skinner *et al.* 1989, p. 83; Muthén and Satorra 1995, p. 279) and a range of resampling methods (Stapleton 2008). This article and the **lavaan.survey** package adopt a framework due to Muthén and Satorra (1995) that encompasses pseudo-maximum likelihood (PML) or weighted ("generalized") least squares (WLS) point estimation, and variance estimation by linearization or resampling. The option of which combination of methods to employ is left to the user, the default being PML, the *de facto* standard for SEMs at the time of writing (Asparouhov 2005).

The framework adopted here starts from the observation (Skinner *et al.* 1989, p. 78) that the problem of the estimation of SEM parameters under complex sampling can be simplified to the usual problem of estimation of means under complex sampling through a classical three-step device (e.g., Fuller 1987, Appendix 4.B). The current discussion of this remarkable observation is necessarily more condensed than that found in the comprehensive discussion by

Muthén and Satorra (1995), but, following the design principle of **lavaan.survey**, also slightly more general in that it allows one to take into account all complex survey design aspects allowed for in the **survey** package. I focus on explaining the three steps which comprise the basic logic behind complex survey analysis of SEMs followed by **lavaan.survey**:

1. All SEM parameter estimates are implicit nonlinear functions of the vector of variances and covariances or, more generally, moments of the observed variables;

2. The moments themselves are linear estimators of the mean vector of a redefined vector of variables d;

3. Therefore, after fitting a SEM using the estimation method of choice, the usual theory for variance estimation of means under complex sampling can be applied to the (co)variances and projected back into the parameter space.

This simple logic produces an incredibly flexible framework for SEM estimation incorporating sampling weights, stratification, and clustering, but also resampling methods and multiple imputation.

2.1. Structural equation models

Given a p-vector of observed variables y, let Σ denote its population covariance matrix, and S_n a sample estimator such that $\mathsf{E}_\pi(S_n) = \Sigma$, where E_π denotes expectation under the sampling design. A SEM is a covariance structure model $\Sigma = \Sigma(\theta)$ expressing the population covariances Σ as a function of a parameter vector θ, an often used parameterization of SEM being the "**LISREL** all-y" model also used by **lavaan**,

$$\begin{aligned} \eta &= B\eta + \zeta \\ y &= \Lambda\eta + \epsilon, \end{aligned} \tag{1}$$

where η is a vector of latent variables, or, equivalently, random effects, and ζ and ϵ are vectors of (latent) residuals. This model implies the covariance structure

$$\Sigma = \Lambda(I - B)^{-1}\Phi\left((I - B)^{-1}\right)^\top \Lambda^\top + \Psi, \tag{2}$$

where $\Phi := \mathsf{VAR}(\zeta)$ and $\Psi := \mathsf{VAR}(\epsilon)$[1]. The model encompasses well-known methods such as factor analysis ($B = I$), random effects modeling ($B = I$, $\Lambda = 1_p$, Ψ is diagonal and $\mathrm{dg}(\Psi) = \psi$), and path analysis ($\Lambda = I$, $\Psi = 0$), as well as any combinations that might be identified. Typically the model parameters are not identifiable without further restrictions; indeed it is customary to impose more restrictions than necessary for identification, allowing for a test of these restrictions. In that case the model degrees of freedom are usually taken to equal $df = p^* - q$, where q is the number of free parameters and $p^* = p(p + 1)/2$, the number of unique (co)variances. For clarity the mean structure is ignored, though the present treatment is easily extended to means and other moments (Satorra 1992).

SEM parameter estimates $\hat{\theta}_n$ are obtained by minimizing a discrepancy function $F(s_n, \sigma(\theta))$, where $s_n := \mathrm{vech}(S_n)$, $\sigma := \mathrm{vech}(\Sigma)$, and the vech operator denotes columnwise stacking of

---

[1]Note that in standard **LISREL** notation, $\mathsf{VAR}(\zeta)$ is usually denoted Ψ and $\mathsf{VAR}(\epsilon)$ is denoted Θ. To avoid confusion with the parameter space Θ, we use Neudecker and Satorra (1991)'s notation here.

the non-redundant moments (Magnus and Neudecker 2007). The most common choice for F is the maximum likelihood (ML) discrepancy function,

$$F_{\mathrm{ML}} = \log|\Sigma(\theta)| + \mathrm{trace}(S_n[\Sigma(\theta)]^{-1}) - \log|S_n| - p. \tag{3}$$

It is straightforward to show (e.g., Bollen 1989, Chap. 4 Appendix) that minimizing F_{ML} maximizes the likelihood of the data under multivariate normality. Under the model (see Fuller 1987, pp. 334-5) and as the sample size increases, the F_{ML} becomes asymptotically equal to the weighted ("generalized") least squares (WLS) fitting function (Browne 1984),

$$F_{\mathrm{WLS}} = [s_n - \sigma(\theta)]^\top V_n[s_n - \sigma(\theta)], \tag{4}$$

where V_n consistently estimates a symmetric estimation weight matrix V. For the asymptotic equality to normal-theory ML to hold, the estimation weights V_n are chosen as the inverse of the normal-theory sampling variance of s_n, denoted Γ_{NT} (Fuller 1987, Appendix 4.B): $V_{\mathrm{NT}} = \Gamma_{\mathrm{NT}}^{-1} = 2^{-1}D^\top(\Sigma^{-1} \otimes \Sigma^{-1})D$, where D is the duplication matrix (Magnus and Neudecker 2007). When the data do not follow a multivariate normal distribution, both F_{ML} and F_{WLS} still provide consistent estimates. The asymptotically optimal estimator is obtained by replacing V with a consistent estimate of $\mathsf{VAR}(s_n)^{-1}$, a method sometimes called "asymptotic distribution free" (ADF) estimation (Browne 1984; Satorra 1989). In spite of its asymptotic optimality, the ADF estimator has performed very badly in simulation studies of its finite sampling behavior (Hu, Bentler, and Kano 1992; Bentler and Yuan 1999; Chou, Bentler, and Satorra 2011).

The choice of discrepancy function and estimation weight matrix V_n thus determines the precise form of the estimator. Regardless of this choice, unless the model is just-identified, $\hat{\theta}_n(s_n)$ is neither a linear estimator nor an explicit function of s_n. However, $\hat{\theta}_n(s_n)$ is the solution to the equation $\partial F[s_n, \sigma(\theta)]/\partial\theta = 0$, so that under mild regularity conditions (Satorra 1989), $\hat{\theta}_n(s_n)$ can be viewed an implicit function of s_n.

2.2. Estimation of (co)variances under complex sampling

Since the $\hat{\theta}_n$ are determined entirely by the s_n, it follows that the complex sampling properties of SEM parameter estimates depend on those of the variances and covariances. These can be easily studied by redefining them as a linear estimator. Suppose a complex sample is obtained by sampling, not necessarily with equal probability, primary sampling units (PSU's) within strata, after which second and third stages are sampled. For instance, in the British sample of the European social survey (ESS) round four, 232 postcode sectors (PSU's) were sampled within strata, 20 delivery points (2SU's) sampled within postcode sectors, and for each delivery point, one person aged 15 or over was sampled (3SU's).

Let $\bar{x}$ denote a design-consistent estimator of $\mathsf{E}_\pi(x)$. The estimator $\bar{x}$ possibly but not necessarily involves weighting. Define

$$d_{hi} := \sum_{ct} \mathrm{vech}[(y_{hict} - \bar{y})(y_{hict} - \bar{y})^\top],$$

where y_{hict} is the vector associated with the tth third-stage unit of the cth second-stage unit of the ith PSU of stratum h, with the summation going over all the units within the ith

PSU (Satorra 1992, p. 260). This device essentially redefines the observed data matrix to d, simplifying the estimation of the (co)variances s to that of estimating the mean vector

$$s_n = \bar{d}_n.$$

This simplification of the problem to that of estimating a mean implies that the usual theory of estimators for means may be applied. Assuming that $\Gamma := \mathsf{VAR}_\pi(s_n)$ is finite, the variance estimator $\hat{\Gamma}$ can be obtained by "nonparametric" Taylor linearization (Skinner *et al.* 1989, p. 48; Muthén and Satorra 1995, p. 279), or by resampling methods including jack-knife, balanced repeated replicates, bootstrap, and half-sample methods (Wolter 2007). The R implementation of these methods is described by Lumley (2004, 2010). These variance estimators $\hat{\Gamma}$ are also consistent under nonnormality, so that all complex sampling analyses also take into account any effects of nonnormality of the observed variables. In smaller samples, non-parametric estimators of Γ may become unstable; since, under the model, $\hat{\Gamma}$ also estimates $\mathsf{VAR}_\pi[s_n - \sigma(\theta)]$, Yuan and Bentler (1998, p. 293) suggested using residuals in a model-based adjustment to $\hat{\Gamma}$ that was found to stabilize the estimator in small samples; **lavaan.survey** allows this optionally.

A common problem in surveys is that of missing data, either through item or unit nonresponse. A common solution under the assumption of missingness at random given covariates is to multiply impute the missing values (Little and Rubin 1987; Rubin 2004). For M imputations this yields M estimates s_{mn} for $m = 1, \ldots, M$. The point estimate is then simply the average over imputations, $\bar{s}_{.n}$. The variance Γ of these estimates can be estimated by (Schafer 1997)

$$\hat{\Gamma}_{\mathrm{MI}} = M^{-1} \sum_{m=1}^{M} \hat{\Gamma}_m + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (s_{mn} - \bar{s}_{.n})(s_{mn} - \bar{s}_{.n})^\top. \tag{5}$$

In **lavaan.survey**, this procedure is applied while also taking into account the complex sampling design within imputations whenever multiply imputed data sets are given as data. This is the approach taken for other analysis types in many software packages including, for instance, the **survey** package. It should be noted, however, that any survey weights should be included in the imputation models, and Equation 5 may not consistently estimate the variance if the response mechanism is not at random with respect to the weights (Kott 1995; Kim, Brick, Fuller, and Kalton 2006). Some care should therefore be taken with this approach when weights are involved.

2.3. SEM under complex sampling

Complex sampling impacts a SEM analysis in two ways:

1. The conventional estimator of the covariance matrix may be biased and inconsistent. This, in turn, causes bias in the SEM parameter estimates;

2. The sampling variance Γ of consistent estimates of the (co)variances may be affected by the design. This will affect standard errors and test(s) of model fit.

The first point suggests simply that the design-consistent estimator of the (co)variances $\bar{d}$ should be used for s_n in the fitting function. This will then guarantee consistency of the estimator $\hat{\theta}(s_n)$, at least for the population value $\hat{\theta}(\sigma)$, when the model is misspecified. It

can be shown that minimizing F_{ML} with $\bar{d}$ as an estimate of s_n is equivalent to the PML estimator introduced by Skinner *et al.* (1989, pp. 80–3).

The second point means that a design-consistent estimate $\hat{\Gamma}$ of the sampling variance of the (co)variances under the complex sampling scheme needs to be taken into account. Wolter (2007, Eq. 6.2.2) notes that

$$\mathsf{VAR}_\pi(\hat{\theta}_n) = \mathsf{E}_\pi[(\hat{\theta}_n - \mathsf{E}_\pi(\hat{\theta}_n))^2] \quad = \quad \left(\frac{\partial \hat{\theta}_n}{\partial s_n}\right) \Gamma \left(\frac{\partial \hat{\theta}_n}{\partial s_n}\right)^\top + O_\pi(r_n^3), \tag{6}$$

where r_n is a term that converges to zero as the sample size increases (see also Lumley 2004, p. 3–4). When the model is identified, $\partial \hat{\theta}_n / \partial s_n$ can be obtained by invoking the implicit function theorem: since $\hat{\theta}_n$ is the solution to $\partial F[s, \sigma(\theta)]/\partial \theta = 0$,

$$\frac{\partial \hat{\theta}_n}{\partial s_n} = \left(\frac{\partial^2 F}{\partial \theta \partial \theta^\top}\right)^{-1} \left(\frac{\partial^2 F}{\partial \theta \partial s_n^\top}\right) = [\Delta^\top V \Delta + o_\theta(n^{1/2})]^{-1} \Delta^\top V, \tag{7}$$

where $\Delta := \partial F / \partial \sigma(\theta)$, and $o_\theta(n^{1/2})$ is a term that, under the model, converges to zero as the sample size increases (see Neudecker and Satorra 1991, for the precise form of both quantities). Thus, when the model is correct, the asymptotic variance of the parameter estimates is

$$\mathsf{AVAR}_{\theta,\pi}(\hat{\theta}) = \lim_{n\to\infty} \mathsf{E}_\theta[\mathsf{VAR}_\pi(\hat{\theta}_n)] = (\Delta^\top V \Delta)^{-1} \Delta^\top V \Gamma V \Delta (\Delta^\top V \Delta)^{-1}, \tag{8}$$

where E_θ denotes expectation under the model. Equation 8 can be recognized as the "sandwich" estimator of variance, which is well-known in econometrics. When F_{ML} is used and the data truly have an *iid* multivariate normal distribution, but also when F_{WLS} is used and V_n is chosen such that $V\Gamma V = V$, the asymptotically optimal estimator (AO) is obtained. Its asymptotic variance can then be seen from Equation 8 to reduce to

$$\mathsf{AVAR}_{\theta,\pi,\mathrm{AO}}(\hat{\theta}) = (\Delta^\top V \Delta)^{-1}. \tag{9}$$

For F_{ML} this corresponds to the inverse of the Fisher information.

Two strategies can now be followed for estimation of SEM parameters under complex sampling:

WLS: Fit the model using WLS with data $s_n = \bar{d}$, and the (Moore-Penrose) inverse $\hat{\Gamma}^+$ as the estimation weight matrix V_n in Equation 4. In this case, after fitting the model, the simple form of Equation 9 can be used as a variance estimator;

Robust ML (PML): Fit the model using ML with data $s_n = \bar{d}$, and estimate the variance with Equation 8, setting $V = \Gamma_{\mathrm{NT}}^{-1}$ and plugging in the design-consistent $\hat{\Gamma}$ matrix.

WLS estimation is asymptotically optimal and similar to the commonly employed complex sampling estimator for multiple regression (Skinner *et al.* 1989; Skinner and de Toledo Vieira 2007; Fuller 2009, Section 6.3). However, it can lead to unstable estimates and has been found in simulation studies to have larger mean square error than the robust ML method, as well as producing test statistics that do not approach their nominal distribution (Hu *et al.* 1992; Bentler and Yuan 1999; Vieira and Skinner 2008; Chou *et al.* 2011). For this reason the robust ML method is the default in **lavaan.survey**, though both methods are implemented.

2.4. Goodness-of-fit testing of the restrictions

Under the null hypothesis of model correctness, the residual covariances should approach zero as the sample size increases. A chi-square statistic for a test of this hypothesis when the estimation procedure is AO is $\chi^2_{\mathrm{AO}}(df) = nF$. A large number of other fit indices exist, all of which are derived either from this model chi-square statistic or from the residuals directly (Kline 2011). When the robust ML procedure is used, nF no longer follows a chi-square distribution and an adjustment to the test statistic is necessary. The most commonly used adjustment matches the first moment of the test statistic to that of the true distribution:

$$\chi^2_{\mathrm{SB}}(df) = \chi^2_{\mathrm{AO}}(df)/\bar{\delta}, \tag{10}$$

where $\bar{\delta}$ is the average generalized design effect (Rao and Scott 1984, p. 53; Skinner *et al.* 1989, pp. 43-4). This adjustment is known in the SEM literature as the "Satorra-Bentler" (SB) chi-square (Satorra and Bentler 1994). In **lavaan**, robust ML estimation using the mean generalized design effect adjustment is called "MLM" estimation, and this is the default used in **lavaan.survey**. Equivalently, the options `se = "robust"` and `test = "Satorra.Bentler"` may be passed to **lavaan**. The mean generalized design effect is given in the output as the "scaling correction factor for the Satorra-Bentler correction". If one defines $n_{\mathrm{eff}} := n/\bar{\delta}$, Equation 10 provides a rationale for the effective sample size method discussed by Stapleton (2002), although the method of obtaining n_{eff} is different, and there is no guarantee that standard errors will be adequately corrected just by replacing n by n_{eff} in variance estimators.

Other adjustments include the "mean-and-variance" (T_3) adjustment of Asparouhov and Muthén (2010, p. 4) and the Satterthwaite (1941) adjustment, which adjusts the degrees of freedom in addition to the value of the test statistic itself. These options are available in **lavaan.survey** by making use of their implementations in **lavaan**. For contingency table tests, Thomas and Rao (1987) found that the Satterthwaite adjustment had a good overall performance, whereas the mean-adjusted test statistic required the coefficient of variation of the generalized design effects to be small. Although the SEM literature on complex sampling (see Bollen *et al.* 2013) has mostly focused on the Satorra-Bentler adjustment, it is therefore possible that the Satterthwaite adjustment may actually be preferable. Currently, I am not aware of any simulation studies investigating this issue explicitly in SEM; therefore **lavaan.survey** follows the SEM literature in choosing the Satorra-Bentler adjustment by default but allows the Satterthwaite adjustment optionally.

Instead of a chi-square fit statistic, one may also consider an F reference distribution, where the denominator degrees of freedom are chosen as the survey design degrees of freedom. This adjustment was found by Thomas and Rao (1987) to perform well when the number of PSU's was small (thanks are due to an anonymous reviewer for this suggestion). The function `pval.pFsum` allows the user to obtain p values for the global test statistic from the F reference distribution with df and survey design degrees of freedom.

In small samples or with few clusters relative to the number of observed covariances to be estimated, it may happen that $\hat{\Gamma}$ is singular. This is not a problem in itself, as the model is typically restricted so that the parameters may still be identified from the observed moments even when some of these moments are collinear. However, for robust ML there may also be cases where the effective degrees of freedom for the goodness-of-fit test are smaller than $df = p^* - q$, which is the default. After using robust ML with a singular $\hat{\Gamma}$, **lavaan.survey** therefore checks whether the degrees of freedom for the goodness-of-fit test are still valid.

Argument	Description	Default
`lavaan.fit`	Output from a conventional SEM analysis using **lavaan**.	–
`survey.design`	Either a 'svydesign' or `svrepdesign`' object produced by a function from package **survey**.	–
`estimator`	Point and variance estimator to be used.	Robust ML (PML).
`estimator.gamma`	Any adjustments/smoothing of $\hat{\Gamma}$.	No adjustments.

Table 1: Overview of the parameters of a `lavaan.survey` call.

If they are not, a warning is issued with the advice to use the Satterthwaite adjustment of degrees of freedom. **lavaan.survey** also implements Yuan and Bentler (1998)'s model-based smoothing of $\hat{\Gamma}$ which may represent an alternative way of alleviating this problem; in the future we plan to add more smoothing estimators to the program.

3. About the package

lavaan.survey is a concise package written entirely in interpreted R code and published under the GPL 2. It links the **survey** and **lavaan** packages, thus providing an interface to a great variety of structural equation model analyses under complex sampling. Its aim is to provide sensible defaults while allowing for flexibility and to check for any estimator-specific problems where possible. In addition, it pools the mean and covariance estimates from data sets obtained by multiple imputation to allow for complex sampling analyses with missing data. The general workflow of a **lavaan.survey** analysis is:

- Create **lavaan** fit object to specify the model;
- Create **survey** 'svydesign' object to specify the sampling design;
- Call the `lavaan.survey` function with the fit and design objects as arguments.

The order of the first two items does not matter. Table 1 gives an overview of the arguments taken by the function `lavaan.survey`.

4. Applications

This section demonstrates the use and features of **lavaan.survey** by discussing four example applications. Code and data for the examples are available as supplementary material from the journal web page.

4.1. Replicate weights analysis of math ability in PISA

Ferla, Valcke, and Cai (2009) present a SEM analysis of academic self-efficacy and academic self-concept's effects on children's math ability in Belgium. Particularly, they were interested in how these variables mediated effects on math ability of other variables such as the level of the school, as well as in the interrelation between efficacy and self-concept. The authors analyzed data from the OECD's 2003 Programme for International Student Assessment (PISA), a
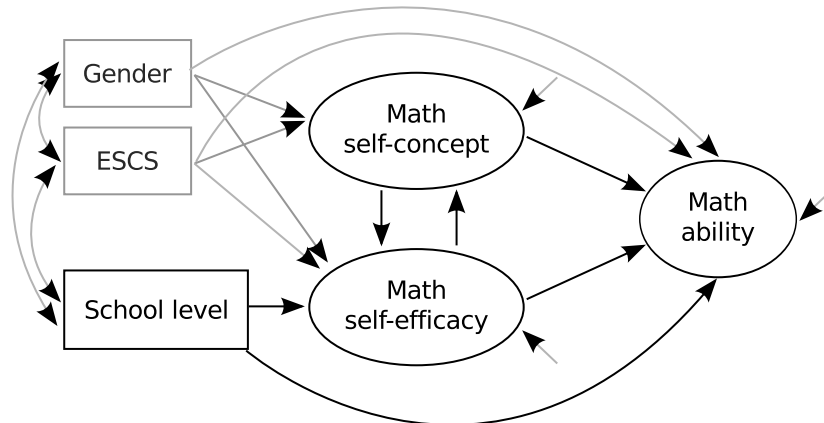
Figure 1: Path diagram for a simplified model of children's math ability in the PISA study. Observed indicators for the three latent variables are omitted from the picture for clarity.

large multinational survey that employs multistage stratified sampling (OECD 2009). Due to the high complexity of PISA's sampling design as well as for purposes of nondisclosure, the OECD does not provide the original design variables, but rather a set of 80 replicate weights generated by the closed-source program **WesVar** (**?**). To take the sampling design into account in SEM analyses of PISA data, these replicate weights need to be included in the analysis, a feature not available in any standard SEM software. This section shows how such an analysis may be performed using **lavaan.survey**. The 2003 PISA data are freely available from the OECD website (http://www.oecd.org/pisa/); here I follow the original authors in analyzing a subset of these data containing the Belgian sample and variables measuring students' math ability in four domains, self-concept of their math ability, self-efficacy, and school level. For the precise definitions of these variables and the indicators used please see the appendix to Ferla *et al.* (2009). In addition, gender and socio-economic status of the parents will be used as fixed covariates. The subset analyzed is included in the online supplement and can be loaded onto the R workspace with the command

```
R> data("pisa.be.2003")
```

In addition to the observed variables, the raw data also contain 80 replicate weights generated by balanced-repeated replication using Fay (1989)'s method with $\rho = 0.5$ (OECD 2009). Using the svrepdesign function from the **survey** package, a survey design object is defined taking this into account:

```
R> des.rep <- svrepdesign(ids = ~1, weights = ~W_FSTUWT, data = pisa.be.2003,
+    repweights = "W_FSTR[0-9]+", type = "Fay", rho = 0.5)
```

It may be of interest to educational researchers that the options used here, weights = ~W_FSTUWT, repweights = "W_FSTR[0-9]+", type = "Fay", and rho = 0.5, are applicable to any analysis of the 2003 PISA data, not just the one at hand.

Having defined the sampling design, the next step is to perform a conventional SEM analysis without taking this design into account. Figure 1 shows a simplified version of the model analyzed by Ferla *et al.* (2009) as a path diagram. The figure shows that a reciprocal effect

between self-concept and self-efficacy is specified, which is identifiable due to the absence of a direct effect of school level on self-concept. Self-concept and efficacy affect math ability and are also partially mediating variables for the effect of school level on math ability. All structural relationships are controlled for gender and socio-economic status of the parents. For clarity, Figure 1 omits the indicators of the three latent variables self-concept, self-efficacy, and ability; these are assumed to be measured by five, eight, and four indicators respectively in a simple factor structure. This model can be specified in **lavaan** syntax as:

```
R> model <- "
+    math = ~ PV1MATH1 + PV1MATH2 + PV1MATH3 + PV1MATH4
+    neg.efficacy = ~ ST31Q01 + ST31Q02 + ST31Q03 + ST31Q04 +
+                     ST31Q05 + ST31Q06 + ST31Q07 + ST31Q08
+    neg.selfconcept = ~ ST32Q02 + ST32Q04 + ST32Q06 + ST32Q07 + ST32Q09
+
+    neg.selfconcept ~ neg.efficacy + ESCS + male
+    neg.efficacy ~ neg.selfconcept + school.type + ESCS + male
+    math ~ neg.selfconcept + neg.efficacy + school.type + ESCS + male
+ "
```

In this syntax, `= ~` indicates "measured by" and `~` "regressed on". Means and variances are freed in the `lavaan` function call. For more information on the precise working and syntax of **lavaan**, please see Rosseel (2012). A conventional SEM analysis on the raw data is then performed:

```
R> fit <- lavaan(model, data = pisa.be.2003, auto.var = TRUE, std.lv = TRUE,
+    meanstructure = TRUE, int.ov.free = TRUE, estimator = "MLM")
R> fit


lavaan (0.5-10) converged normally after 161 iterations

                                              Used         Total
  Number of observations                      7785          8796

  Estimator                                     ML        Robust
  Minimum Function Chi-square             8088.256      7275.544
  Degrees of freedom                           158           158
  P-value                                    0.000         0.000
  Scaling correction factor                                1.112
    for the Satorra-Bentler correction
```

By specifying `estimator = "MLM"`, this conventional analysis uses the option of calculating nonnormality-robust standard errors and chi-square, yielding a "scaling correction" (average generalized "design" effect of nonnormality) of 1.1. This serves to make the conventional analysis more comparable to the complex sampling analysis, which can be expected to increase the scaling correction relative to the value after taking nonnormality into account.

Now that a **survey** design object and a **lavaan** fit object have been obtained, the complex sampling analysis can be performed using `lavaan.survey`:

	Estimate		S.E.		
	Naive	PML	Naive	PML	creff
`neg.selfconcept ~ neg.efficacy`	−0.021	−0.050	0.032	0.046	1.415
`neg.efficacy ~ neg.selfconcept`	0.568	0.609	0.046	0.065	1.421
`neg.efficacy ~ school.type`	0.530	0.518	0.022	0.022	1.009
`math ~ neg.selfconcept`	−0.179	−0.177	0.015	0.021	1.362
`math ~ neg.efficacy`	−0.239	−0.237	0.015	0.018	1.216
`math ~ school.type`	−0.606	−0.596	0.019	0.035	1.858

Table 2: Point and standard error estimates using robust ML with and without correction for the sampling design using BRR replicate weights.

```
R> fit.surv <- lavaan.survey(lavaan.fit = fit, survey.design = des.rep)

lavaan (0.5-10) converged normally after 193 iterations

  Number of observations                          7785

  Estimator                               ML        Robust
  Minimum Function Chi-square          8187.514    5642.873
  Degrees of freedom                        158         158
  P-value                                 0.000       0.000
  Scaling correction factor                           1.451
    for the Satorra-Bentler correction
```

This example call uses all the defaults, i.e., robust ML estimation without model-based smoothing; this is equivalent to pseudo-maximum likelihood (PML) estimation. The average generalized design effect taking into account both nonnormality and the sampling design is 1.45, which is 31% higher than that for the conventional analysis only taking nonnormality into account.

Table 2 gives the point and standard error estimates for the parameters of primary interest, corresponding to the black arrows in Figure 1. For comparison, both the results from the "naive" conventional SEM analysis and from the `lavaan.survey` analysis employing the replicate weights are given. Table 2 shows that the differences in point estimates are relatively small. The differences in standard error estimates, however, are considerable. The average ratio between the standard errors from the complex sampling and the conventional analysis, termed "conditional relative efficiency" (creff) in the table (Oberski 2011, Chap. 3, Oberski 2013b), is 1.38.

The model fits very badly, even after taking the scaling due to complex sampling into account. One method of investigating which restrictions are especially offensive is to examine the "modification indices" (also known as "score" or "Lagrange multiplier" tests) for restricted parameters. Under the null hypothesis of a correct restriction, these will follow a chi-square distribution with one degree of freedom. **lavaan** allows the user to obtain modification indices with the command `modificationIndices`, which are adjusted to the complex sampling design after the call to `lavaan.survey`.

```
R> head(arrange(modificationIndices(fit.surv)[, -c(7, 9)], mi.scaled,
```

```
+    decreasing = TRUE))

              lhs op      rhs   mi mi.scaled    epc sepc.all
1        ST31Q05 ~~ ST31Q07 1964      1354  0.289    0.355
2        ST31Q05 ~~ ST31Q08  566       390 -0.153   -0.201
3 neg.selfconcept =~ PV1MATH1  357      246 -0.019   -0.091
4        ST31Q03 ~~ ST31Q08  348       240  0.122    0.158
5           math =~ ST31Q08  340       235  0.149    0.229
6        ST31Q03 ~~ ST31Q05  332       229 -0.111   -0.144
```

The **plyr** (Wickham 2011) function `arrange` was used to give a concise syntax. The modification indices adjusted for complex sampling are shown as `mi.scaled`. The most problematic restrictions appear to be zero error correlations among items in the self-efficacy construct. This may indicate common method variance or multidimensionality of the latent self-efficacy variable.

4.2. Confirmatory factor analysis (CFA) of welfare state attitudes

Roosma, Gelissen, and van Oorschot (2013) discuss an analysis of citizens' attitudes toward the welfare state. They used data from the 2008 (fourth) round of the ESS to compare factor means that represented whether respondents thought the welfare state was legitimate and achieved its stated goals across countries. An additional goal of the study was the investigation of the relationship between the factors. The ESS is a multinational survey in which each country has its own sampling design – a design that can vary in complexity from simple random sampling from a population register (e.g., Denmark) to four-stage stratified cluster sampling (e.g., Turkey). This section analyzes the United Kingdom (UK) sample, focusing on two of the factors investigated by Roosma *et al.* (2013).

The ESS data for round four are publicly downloadable online (http://www.europeansocialsurvey.org/data/). The UK subset analyzed here additionally includes information on strata and primary sampling units that is absent from the public database. The subset is included in the online appendix.

```
R> data("ess4.gb")
```

Focusing on two factors representing "range" and "outcomes goals" (Roosma *et al.* 2013, Table 1), a two-factor model is formulated using **lavaan** syntax as:

```
R> model.cfa <-
+    "range = ~ gvjbevn + gvhlthc + gvslvol + gvslvue + gvcldcr + gvpdlwk
+     goals = ~ sbprvpv  +  sbeqsoc  +  sbcwkfm"
```

The "range" factor represents the opinion that government should be responsible for various outcomes associated with the welfare state and has six observed indicators. The "outcome goals" factor represents the respondent's opinion of whether these goals are actually reached, and is measured by three observed variables. Of particular interest here are the covariance between the two factors as well as the factor variances. For the precise question formulations and rationale behind these definitions of the factors, please see the ESS website and the original article respectively. The factor model can be estimated with **lavaan** using

```
R> fit.cfa.ml <- lavaan(model.cfa, data = ess4.gb, estimator = "MLM",
+    meanstructure = TRUE, int.ov.free = TRUE, auto.var = TRUE,
+    auto.fix.first = TRUE, auto.cov.lv.x = TRUE)
R> fit.cfa.ml

lavaan (0.5-10) converged normally after  51 iterations
```

	Used	Total
Number of observations	2108	2273
Estimator	ML	Robust
Minimum Function Chi-square	483.564	379.983
Degrees of freedom	26	26
P-value	0.000	0.000
Scaling correction factor		1.273
for the Satorra-Bentler correction		

This shows again that the nonnormality of the data have a considerable effect on the standard errors and chi-square test of model fit. Since the original authors were satisfied with the attained model fit and the focus is here on the estimation of the relationship between the factors, we shall ignore the issue of model fit in this application.

The UK sample was stratified based on 37 regions (`stratval`). Within each region, postcode sectors (`psu`) were listed in increasing order of population density and tenure; sectors with fewer than 500 delivery points were combined. In the first stage a systematic sample of 232 sectors (225 in Great Britain and 7 in Northern Ireland) was then drawn with probability proportional to postal delivery point count. The second and third stages were simple random sampling of 20 postal delivery points within the sector, and selection by Kish grid of one person aged 15 or over at the selected address. In some cases there was an intermediate stage in which a dwelling required selection from an address before a person could be selected within the dwelling. The final sampling weights (`dweight`) were constructed by the ESS sampling team by multiplying all selection probabilities together, normalizing to the nominal sample size, and finally trimming the weights at 4. This rather complicated design can be neatly summarized in a **survey** design object:

```
R> des.gb <- svydesign(ids = ~psu, strata = ~stratval, weights = ~dweight,
+    data = ess4.gb)
```

After the definition of the sampling design, the confirmatory factor analysis taking it into account using robust ML is again performed using **lavaan.survey**:

```
R> fit.cfa.surv <- lavaan.survey(fit.cfa.ml, survey.design = des.gb)
R> fit.cfa.surv

lavaan (0.5-10) converged normally after  50 iterations
```

Number of observations	2108

```
Estimator                                       ML        Robust
Minimum Function Chi-square                 513.094       333.119
Degrees of freedom                              26            26
P-value                                      0.000         0.000
Scaling correction factor                                  1.540
   for the Satorra-Bentler correction
```

The mean generalized design effect taking both nonnormality and the sampling design into account is 21% higher than that taking only nonnormality into account.

An alternative to the default robust ML estimator is WLS using the (generalized) inverse of $\hat{\Gamma}$ as a weight matrix. This can be accomplished in **lavaan.survey** by changing the estimator to `"WLS"`.

```
R> fit.cfa.surv.wls <- lavaan.survey(fit.cfa.ml, survey.design = des.gb,
+     estimator = "WLS")
```

Since, as remarked above, this method was found unstable in a range of simulation studies and applications, a possible adjustment is to smooth the estimation weights for WLS using the model-based smoothing method suggested by Yuan and Bentler. This can be done by changing the setting for `estimator.gamma` to `"Yuan-Bentler"`.

```
R> fit.cfa.surv.wls.yb <- lavaan.survey(fit.cfa.ml, survey.design = des.gb,
+     estimator = "WLS", estimator.gamma = "Yuan-Bentler")
```

To estimate the covariances used as input for the SEM analysis and their $\hat{\Gamma}$ matrix, it is also possible to use the various resampling methods available in the **survey** package:

```
R> des.gb.rep <- as.svrepdesign(des.gb, type = "JKn")
```

In this call to the **survey** function `as.svrepdesign`, the jackknife for stratified designs (`"JKn"`) is specified, which is the default. The confirmatory factor analysis can then be performed on the jackknifed covariances using `lavaan.survey`:

```
R> fit.cfa.surv.rep <- lavaan.survey(fit.cfa.ml, survey.design = des.gb.rep)
R> fit.cfa.surv.rep
```

```
lavaan (0.5-10) converged normally after  50 iterations

  Number of observations                        2108

  Estimator                                       ML        Robust
  Minimum Function Chi-square                513.094       332.248
  Degrees of freedom                              26            26
  P-value                                      0.000         0.000
  Scaling correction factor                                  1.544
     for the Satorra-Bentler correction
```

	$\widehat{\mathrm{VAR}}$(range)			$\widehat{\mathrm{VAR}}$(goals)			$\widehat{\mathrm{COV}}$(range, goals)		
	Est.	S.E.	creff	Est.	S.E.	creff	Est.	S.E.	creff
ML, Naive	1.961	0.162		0.188	0.026		−0.111	0.024	
ML, Taylor	1.893	0.170	1.046	0.186	0.030	1.175	−0.115	0.033	1.373
ML, jackknife	1.893	0.170	1.050	0.186	0.030	1.177	−0.115	0.033	1.379
WLS	1.063	0.105	0.649	0.056	0.023	0.914	0.031	0.014	0.573
WLS, Y-B	2.370	0.203	1.250	0.160	0.027	1.062	−0.146	0.035	1.458

Table 3: Factor variances and covariance of interest for attitudes to the welfare state in the UK sample of the European Social Survey (2008). Point and standard error estimates using robust ML without correction for the sampling design, by Taylor linearization, by jackknifing, using WLS, and using WLS with the Yuan-Bentler correction.

In this case, the results of complex sampling CFA using the jackknife gives results very similar to the default method using Taylor linearization.

Table 3 presents point and standard error estimates, as well as a relative efficiency compared with the naive method for three parameters of interest, namely the factor variances and the covariance. Table 3 gives results using the five different methods discussed above: ML not taking the sampling design into account ("Naive"), the default robust ML method using linearization to estimate $\hat{\Gamma}$ ("Taylor"), the robust ML method using the stratified jackknife to estimate $\hat{\Gamma}$, WLS using the linearized $\hat{\Gamma}^{-1}$ matrix as weights ("WLS") and the same method using the model-based smoothing estimate of $\hat{\Gamma}$ suggested by Yuan & Bentler ("WLS, Y-B").

Table 3 again shows that the point estimates for different versions of ML are very similar. As the conditional relative efficiencies indicate, the standard error estimates using both Taylor linearization and the jackknife are substantially larger than those obtained under the naive method; these two methods give very similar results in all respects. Unadjusted WLS estimation gives point estimates that are wildly different from those obtained by all of the other methods: most strikingly, the relationship between the factors is estimated to be positive rather than negative using this method, with z-values larger than 2 for both WLS and the other methods. However, when the Yuan-Bentler smoother is applied to the $\hat{\Gamma}$ matrix, point estimates are obtained that are much more similar to those obtained with ML.

Although it is possible that WLS is the only method indicating the correct direction of the relationship, cautions in the literature on this estimator would suggest that the ML or Yuan-Bentler smoothed WLS estimators are likely to be preferable. A caveat on this last estimator is that it relies on the correctness of a model for which the fit statistic indicates significant misspecification, so that the stability it introduces relative to the WLS estimator may be paid for with some amount of bias. This trade-off may work out well in some applications, however.

4.3. Multiple imputation of dropouts in the LISS panel

The longitudinal internet studies for the social sciences (LISS) panel is a web survey panel recruited by probability sampling. A random sample of households from the Dutch population register was asked to participate in the panel, and all household members were then asked to participate. To prevent undercoverage problems, the panel organizers provided internet connections and computers to those who did not have them. For more details on the design
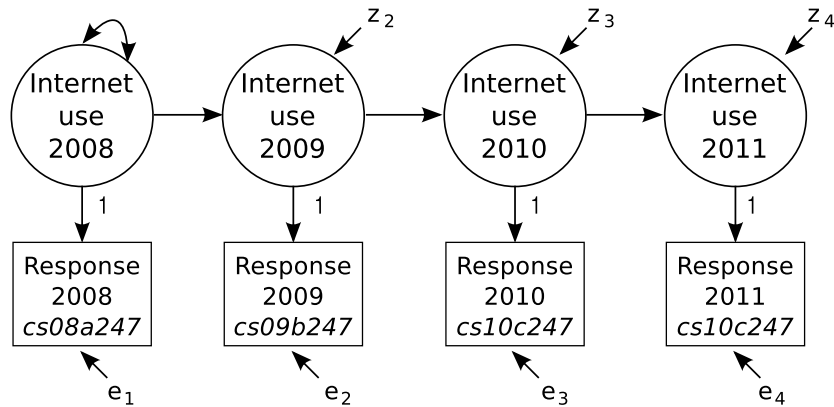
Figure 2: The quasi-simplex model for the evaluation of measurement error in the question "How many hours per week do you use the internet at home?".

of the LISS panel and recruitment efforts, see Scherpenzeel (2011). The LISS panel measures a wide range of variables and allows external researchers to submit proposals as well.

One question of interest is whether these questions have sufficient reliability to be of use for substantive research. Thanks to the longitudinal design, this can be investigated by the so-called "quasi-simplex" model (Alwin 2011), which Figure 2 represents as a SEM for the variable "internet use". The model in Figure 2 is only identified by the additional restriction $\mathsf{VAR}(e_t) = \vartheta$, i.e., equality of measurement error variances (Jöreskog 1970). Parameters of interest could then be the error variance ϑ itself, but also the reliability ratio at a time point, for example $\rho_1 := \vartheta/\mathsf{VAR}(\texttt{cs08a247})$.

The data for estimating the model in Figure 2 can be loaded by:

```
R> data("liss")
```

This data set contains the answers 7369 respondents gave to the question "How many hours per week, on average, do you use the internet at home?" when asked in 2008, 2009, 2010, and 2011, as well as the household identifier. The model in Figure 2 can be written in **lavaan** syntax as:

```
R> model.liss <- "
+    cs08 = ~ 1 * cs08a247
+    cs09 = ~ 1 * cs09b247
+    cs10 = ~ 1 * cs10c247
+    cs11 = ~ 1 * cs11d247
+
+    cs09 ~ cs08
+    cs10 ~ cs09
+    cs11 ~ cs10
+
+    cs08a247 ~~ vare * cs08a247
+    cs09b247 ~~ vare * cs09b247
+    cs10c247 ~~ vare * cs10c247
+    cs11d247 ~~ vare * cs11d247
```

```
+
+    cs08 ~~ vart08 * cs08
+
+    reliab.ratio : =  vart08 / (vart08 + vare)
+ "
```

The last line defines the reliability ratio as `reliab.ratio`. **lavaan** will automatically output the point estimate for `reliab.ratio` as well as its standard error (using the delta method). As before, the model accounting for household clustering (`nohouse_encr`) can be estimated with `lavaan.survey`:

```
R> fit.liss <- lavaan(model.liss, auto.var = TRUE, meanstructure = TRUE,
+    int.ov.free = TRUE, data = liss)
R> des.liss <- svydesign(ids = ~nohouse_encr, prob = ~1, data = liss)
R> fit.liss.surv <- lavaan.survey(fit.liss, des.liss)
R> fit.liss.surv

lavaan (0.5-10) converged normally after  26 iterations

  Number of observations                            3374

  Estimator                                   ML       Robust
  Minimum Function Chi-square               2.496       1.836
  Degrees of freedom                           2           2
  P-value                                  0.287       0.399
  Scaling correction factor                            1.360
    for the Satorra-Bentler correction
```

The reliability estimate itself and its standard error and 95% confidence interval can be inspected by

```
R> parameterEstimates(fit.liss.surv)[24, ]

          lhs op                  rhs        label   est    se    z pvalue
1 reliab.ratio := vart08/(vart08+vare) reliab.ratio 0.622 0.017 37.5      0
  ci.lower ci.upper
1    0.589    0.654
```

As shown in the **lavaan** output, although there are 7369 respondents in the data set, after listwise deletion only 3374 complete observations are left to estimate the reliability. Figure 3 shows that this large amount of missing data is mostly due to panel attrition (dropouts) over time. The attrition is considerable, reaching 46% in the 2011 wave.

One method of dealing with this large amount of missing data is multiple imputation. As with many missing data methods, the core assumption is that the data are missing at random, given the covariates used for imputation. In this application the covariates for a missing answer by a respondent are that respondent's answers at previous time points, so that this assumption is not entirely implausible. To create multiply imputed data sets, the R packages
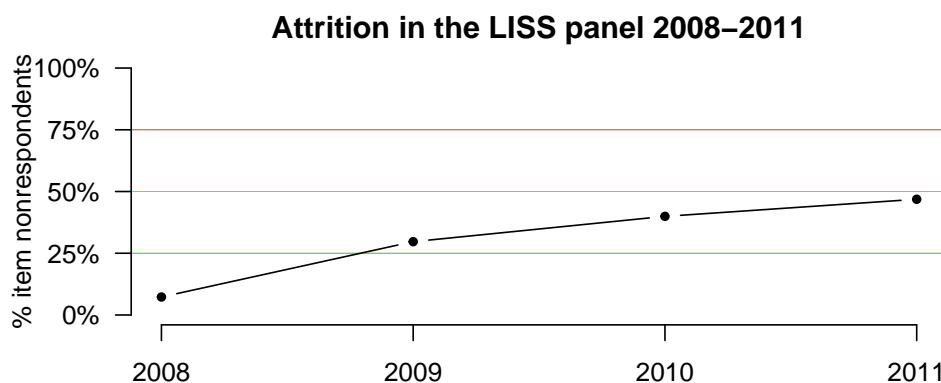
Attrition in the LISS panel 2008–2011

Figure 3: The percentage of item nonresponders to the internet use question in four consecutive waves of the LISS panel study.

mice (Buuren and Groothuis-Oudshoorn 2011), **mi** (Su, Gelman, Hill, and Yajima 2011), and **Amelia** (Honaker, King, and Blackwell 2011) can be used, but the user can also create multiply imputed data sets with an external program such as **WinBUGS** (Spiegelhalter, Thomas, Best, and Lunn 2003; Lunn, Thomas, Best, and Spiegelhalter 2000). This example uses the **mice** package to impute the dropouts 100 times (not run):

```
R> library("mice")
R> liss.imp <- mice(liss, m = 100, method = "norm", maxit = 100)
```

The **lavaan.survey** package follows the **survey** package's design in employing the **mitools** (Lumley 2012a) package to analyze multiply imputed data sets. This provides full flexibility by allowing the multiply imputed data sets to come from any source. After imputation using `mice`, an 'imputationList' object can be created by:

```
R> library("mitools")
R> liss.implist <- lapply(seq(liss.imp$m),
+    function(im) complete(liss.imp, im))
R> liss.implist <- imputationList(liss.implist)
```

The analysis can then proceed as before, using `liss.implist` as data; `lavaan.survey` will detect that multiply imputed data sets have been given as input and pool these in the estimation of the covariance and $\hat{\Gamma}$ matrices.

```
R> des.liss.imp <- svydesign(ids = ~nohouse_encr, prob = ~1,
+    data = liss.implist)
R> fit.liss.surv.mi <- lavaan.survey(fit.liss, des.liss.imp)
R> fit.liss.surv.mi

lavaan (0.5-10) converged normally after  26 iterations

  Number of observations                          7369
```

```
Estimator                                          ML      Robust
Minimum Function Chi-square                     10.611      8.659
Degrees of freedom                                  2           2
P-value                                         0.005       0.013
Scaling correction factor                                   1.225
   for the Satorra-Bentler correction
```

As can be seen in the output, with this method all 7369 available observations are used in the estimation – the uncertainty across imputations is taken into account via the $\hat{\Gamma}$ matrix calculated by `lavaan.survey`.

```
R> parameterEstimates(fit.liss.surv.mi)[24, ]
```

```
          lhs op                   rhs       label   est    se    z pvalue
1 reliab.ratio := vart08/(vart08+vare) reliab.ratio 0.612 0.011 56.1      0
  ci.lower ci.upper
1    0.591    0.634
```

Using the multiply imputed data set, the reliability estimate for the first time point is slightly lower than that when using the default listwise deletion. The confidence interval, in spite of the added uncertainty due to the multiple imputations, is narrower, indicating an overall increase in information used relative to listwise deletion.

4.4. Species diversity and O_2 productivity of algae in streams

The features of **lavaan.survey** can not only be applied to surveys, but more generally to any situation in which the observations are not *iid*. To demonstrate a non-survey analysis with dependent observations, we reproduce an analysis of an experiment on patches of algae in Californian streams. Cardinale, Bennett, Nelson, and Gross (2009) chose 20 streams in the Sierra Nevada. In each stream, they placed 5 or 10 PVC elbows containing different levels of nutrients and a small patch of agar on which algae could grow. They then returned to the streams about 42 days later and measured 1) species diversity in the stream, 2) species diversity in each patch, 3) biomass of the algae, and 4) rate of oxygen production on each patch. Their SEM explicates the indirect relationship between patch diversity and oxygen production and the role played by the experimentally manipulated nutrient supply. Data on 127 patches in 20 streams are available from:

```
R> data("cardinale")
```

Cardinale *et al.* (2009)'s path model relating log(Nutrients) and $\log{(\text{Nutrients})}^2$ to species diversity, biomass and oxygen production can be formulated as

```
R> model.card <- '
+   PatchDiversity ~ logNutrient + logNutrient2 + StreamDiversity
+   Biomass ~ PatchDiversity + logNutrient
+   O2Production ~ logNutrient + Biomass
+   logNutrient ~~ logNutrient2'
```

This model can be fitted to the patches of algae using

```
R> fit.card <- sem(model.card, data = cardinale, fixed.x = FALSE,
+    estimator = "MLM")
```

The Satorra-Bentler chi-square is 6.18 on 7 degrees of freedom with scaling factor 0.95 and thus appears to be in line with the observed covariances ($p = 0.519$).

The 127 patches were considered independent observations. In practice, however, patches are nested within streams. A way of taking this into account while still estimating the same target parameters is to use **lavaan.survey**, viewing the streams as clusters.

```
R> des.card <- svydesign(ids = ~Stream, probs = ~1, data = cardinale)
R> fit.card.survey <- lavaan.survey(fit.card, des.card, estimator = "MLM")
```

The corrected model yields a Satorra-Bentler chi-square of 3.88 on 7 degrees of freedom with scaling factor 1.52. The conclusion on model fit does not change ($p = 0.793$). The only qualitative difference between the non-*iid* analysis and the robust *iid* analysis is that the nonlinear effect of $\log(\text{Nutrient})^2$ on patch species diversity does not differ significantly from zero ($p = 0.051$) in the robust *iid* analysis, while it does ($p = 0.005$) in the non-*iid* analysis.

The Satorra-Bentler chi-square p value for the overall model fit statistic is derived from large-sample theory applied not only to the number of observations, but also to the number of clusters. Since there are only 20 clusters, a better finite-sample performance might be expected from a p value obtained from an F reference distribution with 19 denominator degrees of freedom. It can be obtained using

```
R> pval.pFsum(fit.card.survey, survey.design = des.card)
```

The p value from the F reference distribution (0.610) differs considerably from that obtained using the chi-square reference distribution (0.793), although neither leads to a rejection of the null hypothesis.

5. Summary

Structural equation modeling is frequently applied to samples that are not *iid*: **lavaan.survey** is designed to deal with this case. This article introduced the **lavaan.survey** package and demonstrated its usage and some of its features by application to four examples motivated by the literature. Because the package joins together the **lavaan** and **survey** packages, both very flexible implementations of respectively structural equation modeling and complex survey analysis, the number of combinations of SEM analyses and sampling designs is countless, and not all of these possibilities could be demonstrated. Instead the goal has been to demonstrate, on the one hand, the manner in which SEM analyses using **lavaan** might be adapted to incorporate the issue of non-*iid* samples, and, on the other, the application of the SEM analysis framework to common problems in complex survey analysis. By joining these two worlds in the open-source R environment, **lavaan.survey** hopes to stimulate progress in the various application areas of SEM, as well as provide a flexible framework for the analysis of methodological issues in survey methodology. An important limitation of **lavaan.survey** at

the time of writing is that categorical data cannot be incorporated, a feature that is planned for future releases of the package.

Acknowledgments

References

Alwin DF (2011). "Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach." In J Madans, K Miller, A Maitland, G Willis (eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality*, pp. 263–293. John Wiley & Sons, New York.

Arbuckle JL (2011). *IBM SPSS AMOS 20 User's Guide*. IBM Corporation, Armonk.

Asparouhov T (2005). "Sampling Weights in Latent Variable Modeling." *Structural Equation Modeling: A Multidisciplinary Journal*, **12**(3), 411–434.

Asparouhov T (2006). "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics – Theory and Methods*, **35**(3), 439–460.

Asparouhov T, Muthén B (2005). "Multivariate Statistical Modeling with Survey Data." In *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*. URL http://www.fcsm.gov/events/papers05.html.

Asparouhov T, Muthén BO (2010). "Simple Second Order Chi-Square Correction." *Technical report*, Statmodel.

Bentler PM (2008). *EQS 6 Structural Equations Program Book*. Multivariate Software, Inc., Encino, CA.

Bentler PM, Yuan KH (1999). "Structural Equation Modeling with Small Samples: Test Statistics." *Multivariate Behavioral Research*, **34**(2), 181–197.

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, others (2011). "**OpenMx**: An Open Source Extended Structural Equation Modeling Framework." *Psychometrika*, **76**(2), 306–317.

Bollen K, Tueller S, Oberski DL (2013). "Issues in the Structural Equation Modeling of Complex Survey Data." In *Proceedings of the 59th World Statistics Congress*. Hong Kong.

Bollen KA (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

Brick JM, Morganstein D, Valliant R (2000). "Analysis of Complex Sample Data Using Replication." *Technical report*, Westat. URL http://www.westat.com/wesvar/techpapers.

Browne MW (1984). "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology*, **37**(1), 62–83.

Buuren S, Groothuis-Oudshoorn K (2011). "**mice**: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67. URL http://www.jstatsoft.org/v45/i03/.

Byrnes JE, Reed DC, Cardinale BJ, Cavanaugh KC, Holbrook SJ, Schmitt RJ (2011). "Climate-Driven Increases in Storm Frequency Simplify Kelp Forest Food Webs." *Global Change Biology*, **17**(8), 2513–2524.

Cardinale BJ, Bennett DM, Nelson CE, Gross K (2009). "Does Productivity Drive Diversity or Vice Versa? A Test of the Multivariate Productivity-Diversity Hypothesis in Streams." *Ecology*, **90**(5), 1227–1241.

Chou CP, Bentler PM, Satorra A (2011). "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*, **44**(2), 347–357.

Cochran WG (1977). *Sampling Techniques*. 3rd edition. John Wiley & Sons, New York.

Duncan OD (1975). *Introduction to Structural Equation Models*. Academic Press.

Fay RE (1989). "Theory and Application of Replicate Weighting for Variance Calculations." In *Proceedings of the Section on Survey Research Methods*, pp. 212–217. American Statistical Association.

Ferla J, Valcke M, Cai Y (2009). "Academic Self-Efficacy and Academic Self-Concept: Reconsidering Structural Relationships." *Learning and Individual Differences*, **19**(4), 499–505.

Fox J (2006). "Teacher's Corner: Structural Equation Modeling with the **sem** Package in R." *Structural Equation Modeling: A Multidisciplinary Journal*, **13**(3), 465–486.

Fox J, Nie Z, Byrnes J (2012). *sem: Structural Equation Models*. R package version 3.0-0, URL http://CRAN.R-project.org/package=sem.

Fuller WA (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Fuller WA (2009). *Sampling Statistics*. John Wiley & Sons, New York.

Grace JB (2006). *Structural Equation Modeling and Natural Systems*. Cambridge University Press.

Graham JW, Hofer SM (2000). *Multiple Imputation in Multivariate Research*. Lawrence Erlbaum Associates.

Hahs-Vaughn DL, Lomax RG (2006). "Utilization of Sample Weights in Single-Level Structural Equation Modeling." *Journal of Experimental Education*, **74**(2), 163–190.

Honaker J, King G, Blackwell M (2011). "**Amelia** II: A Program for Missing Data." *Journal of Statistical Software*, **45**(7), 1–47. URL http://www.jstatsoft.org/v45/i07/.

Hu L, Bentler PM, Kano Y (1992). "Can Test Statistics in Covariance Structure Analysis Be Trusted?" *Psychological Bulletin*, **112**(2), 351.

Jöreskog KG (1970). "Estimation and Testing of Simplex Models." *British Journal of Mathematical and Statistical Psychology*, **23**(2), 121–145.

Jöreskog KG, Sörbom D (2006). ***LISREL** 8.8 for Windows*. Scientific Software International, Inc.

Kaplan D (2008). *Structural Equation Modeling: Foundations and Extensions*. Sage Publications.

Kaplan D, Ferguson AJ (1999). "On The Utilization of Sample Weights in Latent Variable Models." *Structural Equation Modeling: A Multidisciplinary Journal*, **6**(4), 305–321.

Kim JK, Brick JM, Fuller WA, Kalton G (2006). "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Society B*, **68**(3), 509–521.

Kline RB (2011). *Principles and Practice of Structural Equation Modeling*. 3rd edition. The Guilford Press, New York.

Kott P (1995). "A Paradox of Multiple Imputation." In *Proceedings of the Section on Survey Research Methods*, pp. 384–389. American Statistical Association.

Little RJA, Rubin DB (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Lumley T (2004). "Analysis of Complex Survey Samples." *Journal of Statistical Software*, **9**(8), 1–19. URL http://www.jstatsoft.org/v09/i08/.

Lumley T (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, New York.

Lumley T (2012a). ***mitools**: Tools for Multiple Imputation of Missing Data*. R package version 2.2, URL http://CRAN.R-project.org/package=mitools.

Lumley T (2012b). ***survey**: Analysis of Complex Survey Samples*. R package version 3.28-2, URL http://CRAN.R-project.org/package=survey.

Lunn D, Thomas A, Best NG, Spiegelhalter DJ (2000). "**WinBUGS** – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing*, **10**, 325–337.

Magnus JR, Neudecker H (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. 3rd edition. John Wiley & Sons, New York.

Marsh HW, Hau KT (2004). "Explaining Paradoxical Relations Between Academic Self-Concepts and Achievements: Cross-Cultural Generalizability of the Internal/External Frame of Reference Predictions Across 26 Countries." *Journal of Educational Psychology*, **96**(1), 56.

Mclntosh A, Gonzalez-Lima F (1994). "Structural Equation Modeling and its Application to Network Analysis in Functional Brain Imaging." *Human Brain Mapping*, **2**(1-2), 2–22.

Muthén B, Satorra A (1995). "Complex Sample Data in Structural Equation Modeling." *Sociological Methodology*, **25**, 267–316.

Muthén LK, Muthén B (2012). *Mplus User's Guide, 7th Edition.* Muthén & Muthén, Los Angeles, CA.

Neudecker H, Satorra A (1991). "Linear Structural Relations: Gradient and Hessian of the Fitting Function." *Statistics and Probability Letters*, **11**(1), 57–61.

Oberski D (2011). *Measurement Error in Comparative Surveys.* Tilburg University. URL http://arno.uvt.nl/show.cgi?fid=114255.

Oberski D (2013a). *lavaan.survey: Complex Survey Structural Equation Modeling (SEM).* R package version 1.0, URL http://CRAN.R-project.org/package=lavaan.survey.

Oberski DL (2013b). "Conditional Design Effects for Structural Equation Model estimates." In *Proceedings of the 59th World Statistics Congress.* Hong Kong. URL http://daob.nl/wp-content/uploads/2013/04/hk-oberski.pdf.

OECD (2009). *PISA Data Analysis Manual: SPSS and SAS.* 2nd edition. OECD.

Rao JNK, Scott AJ (1984). "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics*, **12**(1), 46–60.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Roelstraete B, Rosseel Y (2011). "**FIAR**: An R Package for Analyzing Functional Integration in the Brain." *Journal of Statistical Software*, **44**(13), 1–32. URL http://www.jstatsoft.org/v44/i13/.

Roosma F, Gelissen J, van Oorschot W (2013). "The Multidimensionality of Welfare State Attitudes: A European Cross-National Study." *Social Indicators Research*, **113**(1), 235–255.

Rosseel Y (2012). "**lavaan**: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. URL http://www.jstatsoft.org/v48/i02/.

Rubin DB (2004). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York.

Saris WE, Stronkhorst LH (1984). *Causal Modelling in Nonexperimental Research: An Introduction to the LISREL Approach*, volume 3. Sociometric Research Foundation Amsterdam.

Satorra A (1989). "Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach." *Psychometrika*, **54**(1), 131–151.

Satorra A (1992). "Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures." *Sociological Methodology*, **22**, 249–278.

Satorra A, Bentler PM (1994). "Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis." In A von Eye, CC Clogg (eds.), *Latent Variables Analysis: Applications to Developmental Research*. Sage, Thousand Oaks, CA.

Satterthwaite FE (1941). "Synthesis of Variance." *Psychometrika*, **6**(5), 309–316.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC.

Scherpenzeel AC (2011). "Data Collection in a Probability-Based Internet Panel: How the LISS Panel was Built and How it Can be Used." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, **109**(1), 56–61.

Skinner CJ, de Toledo Vieira M (2007). "Variance Estimation in the Analysis of Clustered Longitudinal Survey Data." *Survey Methodology*, **33**(1), 3–12.

Skinner CJ, Holt D, Smith TMF (1989). *Analysis of Complex Surveys*. John Wiley & Sons, New York.

Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2003). ***WinBUGS*** *Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

Stapleton LM (2002). "The Incorporation of Sample Weights into Multilevel Structural Equation Models." *Structural Equation Modeling: A Multidisciplinary Journal*, **9**(4), 475–502.

Stapleton LM (2006). "An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data." *Structural Equation Modeling: A Multidisciplinary Journal*, **13**(1), 28–58.

Stapleton LM (2008). "Variance Estimation Using Replication Methods in Structural Equation Modeling with Complex Sample Data." *Structural Equation Modeling: A Multidisciplinary Journal*, **15**(2), 183–210.

StataCorp (2011a). *Stata Data Analysis Statistical Software: Release 12*. StataCorp LP, College Station, TX. URL http://www.stata.com/.

StataCorp (2011b). *Stata Structural Equation Modeling Reference Manual*. Stata Press, College Station.

Sterba SK (2009). "Alternative Model-Based and Design-Based Frameworks for Inference from Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research*, **44**(6), 711–740.

Su YS, Gelman A, Hill J, Yajima M (2011). "Multiple Imputation with Diagnostics (**mi**) in R: Opening Windows into the Black Box." *Journal of Statistical Software*, **45**(2), 1–31. URL http://www.jstatsoft.org/v45/i02/.

Thomas DR, Rao JNK (1987). "Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling." *Journal of the American Statistical Association*, **82**(398), 630–636.

Ullman JB, Bentler PM (2003). "Structural Equation Modeling." In *Handbook of Psychology*. John Wiley & Sons.

Vieira MDT, Skinner CJ (2008). "Estimating Models for Panel Survey Data Under Complex Sampling." *Journal of Official Statistics*, **24**(3), 343–364.

Wickham H (2011). "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software*, **40**(1), 1–29. URL http://www.jstatsoft.org/v40/i01/.

Wolter K (2007). *Introduction to Variance Estimation.* 2nd edition. Springer-Verlag, New York.

Yuan KH, Bentler PM (1998). "Normal Theory Based Test Statistics in Structural Equation Modelling." *British Journal of Mathematical and Statistical Psychology*, **51**(2), 289–309.

Affiliation:

Daniel Oberski
Department of Methods and Statistics
Faculty of Social Sciences
Tilburg University
Tilburg, The Netherlands
E-mail: doberski@tilburguniversity.edu
URL: http://daob.org/