Model Fitting Tests in the Analysis of Panel Data

Marcel de Toledo Vieira *
Federal University of Juiz de Fora, Juiz de Fora, Brazil marcel.vieira@ice.ufjf.br

Abstract

The interest for fitting models to panel data has grown in the last few decades. Modelling methods have to consider the variation in the response variable across the population as well as across time in this context. We review model fitting statistics under the classical (simple random sampling) approach, while we propose new developments on fitting measures when working under the complex sampling approach. We modify the Wald goodness of fit test in the context of models for covariance structures, which is shown to be equivalent to modifying the scaled test statistics. We also propose a modification for the Wald significance test for nested hypothesis. Goodness of fit indices are also modified in order be utilised in the complex survey data context.

Keywords: covariance structure, goodness of fit, longitudinal surveys, multistage sampling.

1. Introduction

The interest for fitting models to longitudinal complex survey data has grown in the last few decades. Longitudinal data modelling methods have to consider the variation across the population as well as across time. A wide class of 'regression-type' models has found a broad range of useful applications with panel survey data (e.g. Diggle *et al*., 2002). Furthermore, a number of methods have been developed in the survey sampling literature to take account of complex sampling schemes in the regression analysis of cross section survey data. See Skinner, Holt and Smith (1989), and Chambers and Skinner (2003), for references. Moreover, some previous work on estimation for panel data models under complex designs has been undertaken, for example, by Sutradhar and Kovacevic (2000), Skinner and Holmes (2003), Skinner and Vieira (2007), and Vieira and Skinner (2008). See Vieira (2009), for an overview.

In this paper we extend this broad approach to model fitting statistics in the longitudinal data modelling context, allowing for complex sampling designs. We review model fitting statistics under the classical (simple random sampling - srs) approach, while we propose new developments on fitting measures when working under the complex sampling approach. The models we consider are presented in Section 2. The paper proceeds in Section 3 to consider model testing statistics in both classic and complex survey contexts, and brief remarks are presented in Section 4.

2. Models

Let the finite population be denoted by $\mathcal{U}$, which is treated as fixed on occasions $1,\ldots,T$. Let N represent the size of $\mathcal{U}$ and $N_o = N \cdot T$. Let $\underline{Y}_i = (Y_{i1},\ldots,Y_{iT})'$ be a random vector containing T repeated observations on the study variable for unit $i = 1,2,\ldots,N$ over the T waves of the survey. Moreover, $E(\underline{Y}_i) = \underline{\mu}_i(\underline{\beta})$ is a $T \times 1$ vector, where $\underline{\mu}_i(\underline{\beta}) = [\mu(\underline{x}_{i1},\underline{\beta}),\ldots,\mu(\underline{x}_{iT},\underline{\beta})]'$ and $\underline{x}_{it}$ is a vector of values of the covariates.

A covariance structure model is a model for the $T \times T$ symmetric population variance-covariance matrix of $\underline{Y}_i$, which is

$$\Sigma = \mathrm{COV}(\underline{Y}_i) = E\{[\underline{Y}_i - \underline{\mu}_i][\underline{Y}_i - \underline{\mu}_i]'\} = \Sigma(\underline{\theta}), \qquad (1)$$

where COV(.) denote population covariance. We assume (1) is the same for each unit i, and that $k = T(T+1)/2$ distinct elements of the variance-covariance matrix $\Sigma(\underline{\theta})$ are constrained to be functions of the $b \times 1$ parameter vector $\underline{\theta}$, with $b < k$.

If we consider a uniform correlation model (Model A in Skinner and Holmes, 2003), $\Sigma(\underline{\theta})$ has diagonal values $\sigma_u^2 + \sigma_v^2$ and off-diagonal values σ_u^2, where σ_v^2 is the

variation on the same individual, σ_u^2 and is the variance across individuals (Lindsey, 1994), and $\underline{\theta} = (\sigma_u^2, \sigma_v^2)'$, with $b = 2$. If we consider a transitory random effects model with a first-order autoregressive process (Model B in Skinner and Holmes, 2003), then $\underline{\theta} = (\sigma_u^2, \sigma_v^2, \gamma)'$, where γ is a regression parameter and $-1 < \gamma < 1$.

3. Model testing

We assume that: (*i*) the observations are equally spaced in time; (*ii*) the sample size is 'large' relative to the number of repeated observations; (*iii*) the sample is selected on one occasion and then the same sample units are returned to on each of the $T-1$ subsequent waves; and (*iv*) there is no nonresponse. Estimation procedures for $\underline{\beta}$ are discussed in details by Skinner and Vieira (2007), Vieira and Skinner (2008), and Vieira (2009), where methods on inference about the covariance matrix Σ, for the variance estimation of $\hat{\Sigma}$, and for estimation of the parameter $\underline{\theta}$ are also developed under the complex surveys approach, including generalized least squares (GLS) and pseudo maximum likelihood (PML) methods.

Note that (1) is our covariance structure hypothesis. Model fit measures are used to assist in the evaluation of whether (1) is valid or not, and if not, such measures could assist to calculate the deviation of Σ from $\Sigma(\underline{\theta})$. In this section, let $\hat{\underline{\theta}}$ denote an estimator of $\underline{\theta}$ which minimizes either $F(\underline{\theta})_{ML}$ or $F(\underline{\theta})_{GLS}$, which are maximum likelihood (ML) and GLS fitting functions, respectively, and $F(\underline{\theta})_{PML}$ or $F(\underline{\theta})_{GLSC}$, which are PML and GLS fitting functions, under the complex sampling context.

3.1 Model testing in the classical context

In this sub-section we work under the assumptions of independent and identically distributed observations. Both Σ and $\Sigma(\underline{\theta})$ are unknown population parameters. Thus for calculating model fit measures we would in fact need to consider their estimators S and $\Sigma(\hat{\underline{\theta}})$, respectively, where $\Sigma(\hat{\underline{\theta}})$ is the covariance matrix evaluated at $\hat{\underline{\theta}}$.

In order to perform a goodness of fit test, we may initially define a null and a generic alternative hypothesis as $H_0 : \Sigma = \Sigma(\underline{\theta})$ against $H_1 : \Sigma$, which is an unrestricted covariance matrix (any $T \times T$ positive definite matrix). Let the population residual covariance matrix be denoted by $\mathrm{E_p}$, so that

$$\mathrm{E_p} = [\Sigma - \Sigma(\underline{\theta})]. \qquad (2)$$

When H_0 is true, $\mathrm{E_p}$ is a zero matrix. The sample residual covariance matrix $\hat{\mathrm{E}}$, defined as $\hat{\mathrm{E}} = [\mathrm{S} - \Sigma(\hat{\underline{\theta}})]$, is the simplest model fit measure. Let $[S_{tt'} - \Sigma(\hat{\underline{\theta}})_{tt'}]$ be individual sample residual covariances, where $S_{tt'}$ and $\Sigma(\hat{\underline{\theta}})_{tt'}$ are the tt'th elements in S and $\Sigma(\underline{\theta})$ respectively.

Furthermore, Jöreskog and Sörbom (1989) proposed the following statistic for summarising the residuals,

$$\mathrm{RMR} = \sqrt{2 \cdot \sum_{t=1}^{T} \sum_{t'=1}^{t} \frac{[S_{tt'} - \Sigma(\hat{\underline{\theta}})_{tt'}]^2}{T(T+1)}},$$

where RMR stands for root mean-square residual. This measure may be adopted to compare the fit of two different models for the same data.

Sample residuals are not only affected by differences between Σ and $\Sigma(\underline{\theta})$, but also by the scales of $\underline{Y}_i$ and by sampling fluctuations (errors). A direct solution for the scales issue may be to calculate correlation residuals as (Bollen, 1989) $r_{tt'} - \hat{r}_{tt'}$, where $r_{tt'}$ is the sample correlation between $\underline{Y}_{it}$ and $\underline{Y}_{it'}$, and $\hat{r}_{tt'}$ denotes the model predicted correlation, so that

$$\hat{r}_{tt'} = \frac{\Sigma(\hat{\underline{\theta}})_{tt'}}{\sqrt{\Sigma(\hat{\underline{\theta}})_{tt} \cdot \Sigma(\hat{\underline{\theta}})_{t't'}}} .$$

Although this correlation residual is allowed to range from -2 to $+2$, we should expect values rather close to zero for models with a reasonably good fit.

Regarding sampling errors, even when H_0 is true the expected amplitude of the individual sample residual covariances depends on *n*, and $\lim_{n\to\infty} \hat{\mathrm{E}} = \mathrm{E_p}$. Therefore, we present below a simultaneous significance test, based on sample residuals. Notice that the H_0 discussed above could be tested via a chi-square test.

Under multivariate normality of $\underline{Y}_i$ and no covariates, if H_0 holds, if $\underline{\theta}$ is identified, and if

$$\Delta = \frac{\partial\{vech[\Sigma(\underline{\theta})]\}}{\partial \underline{\theta}}, \qquad (3)$$

is of full rank *b*, where (3) is a $k \times b$ matrix of partial derivatives of elements of $vech[\Sigma(\underline{\theta})]$ with respect to the elements of $\underline{\theta}$, it holds that $n \cdot F(\hat{\underline{\theta}})_{ML} \sim \chi^2_{k-b}$ under H_0, when $F(\underline{\theta})_{ML}$ is evaluated at the final estimates and the model is true. When $\underline{Y}_i$ is multivariate normally distributed, $n \cdot F(\hat{\underline{\theta}})_{GLS}$ has the same property as $n \cdot F(\hat{\underline{\theta}})_{ML}$. Note that $n \cdot F(\hat{\underline{\theta}})_{GLS}$ is a Wald goodness of fit test statistic, which may be obtained as the minimum value of $F(\underline{\theta})_{GLS}$, when evaluated at $\hat{\underline{\theta}}_{GLS}$.

Satorra (1989), propose asymptotically equivalent significance tests for the difference in chi-square statistics for nested models: (*i*) likelihood ratio test (LRT) or chi-square difference test); (*ii*) Lagrangian multiplier test (LMT) or efficient score test; and (*iii*) Wald test (WT); which all assume that $F(\underline{\theta})$ is asymptotic optimal, i.e. leads to efficient estimators and chi-square statistics. In general, these types of test aim to compare an 'initial' model with a restricted model, which has a sub-vector of parameters that is set to be equal to zero. There are also alternative approaches for performing model selection that have been proposed in the literature. Jöreskog and Sörbom (1989), for example, have proposed a goodness of fit index (GFI) and an adjusted index (AGFI), which penalises the models with more parameters.

4.2 Model testing under complex sampling

In this sub-section, we consider some further developments on covariance structure model fitting statistics when assuming the sample is selected under the complex survey design approach. According to Skinner, Holt and Smith (1989), ignoring the characteristics of the complex samples can lead to invalid statistical tests.

For calculating model fit measures in the present context we adopt S_w (Vieira and Skinner, 2008), the survey weighted sample covariance matrix, as an estimator of Σ, considering that $E(E_p(\mathrm{S}_w)) \doteq E(\mathrm{S}_N) = \Sigma$, as shown by Vieira (2009). We thus consider model fit measures which are functions of S_w and Σ.

For examining (2), i.e. for identifying components of the variance-covariance matrix that are not well fit, we may adopt the sample weighted residual covariance matrix $\hat{\mathrm{E}}_c = \mathrm{S}_w - \Sigma(\hat{\underline{\theta}})$, where the subscript *c* denotes 'complex'. Moreover, let $[S_{w,tt'} - \Sigma(\hat{\underline{\theta}})_{tt'}]$ be the individual weighted sample residual covariances, where $S_{w,tt'}$ is the tt'th element in S_w and $\Sigma(\hat{\underline{\theta}})_{tt'}$. The RMR measure may be adapted to

$$\mathrm{RMR}_c = \sqrt{2 \cdot \sum_{t=1}^{T} \sum_{t'=1}^{t} \frac{[S_{w,tt'} - \Sigma(\hat{\underline{\theta}})_{tt'}]^2}{T(T+1)}} ,$$

which has as special case the RMR measure, when the sampling weights are constant.

Theory developed in the categorical complex survey data analysis and modelling literature (Rao and Scott, 1979) suggests that, under complex sampling, $n \cdot F(\underline{\theta})_{PML}$, would not be asymptotically chi-squared distributed. Simple corrections have been proposed, which may are adapted here to be applied to $n \cdot F(\underline{\theta})_{PML}$ in order to make it approximately chi-squared distributed. We follow an approach proposed by Skinner (1989, Section 3.4).

Remark 1: We initially consider the case of a GLSC type estimator, obtained by minimizing the $F(\underline{\theta})_{GLSC}$ fitting function, with matrix U given by

$$\mathrm{U} = 2 \cdot \mathrm{K}'(\mathrm{W} \otimes \mathrm{W})\mathrm{K}, \qquad (4)$$

where W is any consistent estimator of Σ. Under this situation, a Wald goodness of fit test statistic is given by (Skinner, 1989)

$$X^2_{\mathrm{W},srs} = n \cdot \{vech[\mathrm{S}_w] - vech[\Sigma(\theta)]\}' \mathrm{U}^{-1}\{vech[\mathrm{S}_w] - vech[\Sigma(\theta)]\},$$

which implies,

$$X^2_{\mathrm{W},srs} = n \cdot \left(\frac{1}{2}\right) \cdot tr\{[I - \Sigma(\underline{\theta})\mathrm{S}_w^{-1}]^2\},$$

when S_w is considered as a choice for W in (4). Note, nevertheless, that according to Skinner (1989), the test statistic $X^2_{\mathrm{W},srs}$ is no longer asymptotically chi-squared distributed, but in fact asymptotically

$$X^2_{\mathrm{W},srs} \sim \sum_{d=1}^{k-b} \lambda_d \chi^2_1$$

under H_0, where χ^2_1 are independent chi-squared distributed random variables, and λ_d are non-zero eigenvalues of

$$\mathrm{H} = \mathrm{U}^{-1}\mathrm{C}_c - \mathrm{U}^{-1}\mathrm{C}_c\mathrm{U}^{-1} \cdot \Delta(\Delta'\mathrm{U}^{-1}\Delta)^{-1}\Delta',$$

with C_c defined as the asymptotic covariance matrix of $vech[\mathrm{S}_w]$. As a one moment approximation (Skinner, 1989)

$$\frac{(k-b) \cdot X^2_{\mathrm{W},srs}}{tr(\mathrm{H})},$$

is asymptotically distributed as a χ^2_{k-b}, and may be adopted for testing the goodness of fit of a covariance structure model in a complex sampling context. We also consider substituting matrix U by $\hat{\mathrm{C}}_c$, i.e.

$$X^2_{\mathrm{W}} = n \cdot \{vech[\mathrm{S}_w] - vech[\Sigma(\theta)]\}' \hat{\mathrm{C}}_c\{vech[\mathrm{S}_w] - vech[\Sigma(\theta)]\},$$

where $\hat{\mathrm{C}}_c$, given by

$$a\mathrm{COV}(\hat{\underline{\theta}}) = n^{-1}\left(\Delta'\mathrm{U}^{-1}\Delta\right)^{-1}\Delta'\mathrm{U}^{-1}\mathrm{C}_c\,\mathrm{U}^{-1}\Delta\left(\Delta'\mathrm{U}^{-1}\Delta\right)^{-1},$$

is a consistent estimator for C_c (Skinner and Holmes, 2003), with $\mathrm{C}_c = \mathrm{VAR}\{\sqrt{n} \cdot vech[\mathrm{S}_w]\}$. In this context, X^2_{W} is approximately distributed as χ^2_{k-b} under H_0 (Skinner, 1989). Note that when U is consistent for C_c,

$$X^2_{\mathrm{W}} = X^2_{\mathrm{W},srs} = \frac{(k-b) \cdot X^2_{\mathrm{W},srs}}{tr(\mathrm{H})}, \text{ as } tr(\mathrm{H}) = k - b \text{ in that situation.} \qquad \blacksquare$$

We now consider a Wald significance test for nested hypothesis for situations where the PML fitting function is adopted. Note that we shall follow an approach proposed by Skinner (1989, Section 3.4).

Remark 2: We assume that the asymptotic covariance matrix of $vech[\mathrm{S}_w]$ and the information matrix of the model are non-singular. Thus, let $\underline{\hat{\theta}}_{PML,r}$ be the PML estimator for the restrictive (nested) model, and let $\underline{\hat{\theta}}_{PML,u}$ be the PML estimator for the covariance structure model without constraints. Let $\underline{\theta}^R$ be a $b^* \times 1$ vector when $\underline{\theta}^R = \underline{0}$ corresponds to the restraints imposed to the unrestricted model, where $b^* < b$. We propose a modification to the Wald test, so that

$$\mathrm{WT}_c = \underline{\hat{\theta}}^R_{PML,u}{}' \cdot \left\{ \left[\frac{\partial \underline{\hat{\theta}}^R_{PML,u}}{\partial \underline{\hat{\theta}}_{PML,u}} \right]' \cdot \left[a\,\mathrm{cov}(\underline{\hat{\theta}}_{PML,u}) \right] \cdot \left[\frac{\partial \underline{\hat{\theta}}^R_{PML,u}}{\partial \underline{\hat{\theta}}_{PML,u}} \right]' \right\}^{-1} \cdot \underline{\hat{\theta}}^R_{PML,u},$$

where we adopt the approach of Binder (1983), adapted to the covariance structure models context, for calculating $a\,\mathrm{cov}(\underline{\hat{\theta}}_{PML,u})$. When $\sqrt{n^{-1}} \cdot (\underline{\hat{\theta}}_{PML} - \underline{\theta})$ is asymptotically normal distributed, then the modified WT_c statistic introduced above is asymptotically $\chi^2_{b^*}$ distributed (Skinner, 1989) under H_0. ■

We also propose modifying the overall model fit descriptive measures, such as the Jöreskog and Sörbom (1989) goodness of fit indices (GFI). Therefore, let

$$\mathrm{GFI}_{c,PML} = 1 - \left\{ \frac{tr[(\Sigma(\underline{\hat{\theta}})^{-1}\mathrm{S}_w - I)^2]}{tr[(\Sigma(\underline{\hat{\theta}})^{-1}\mathrm{S}_w)^2]} \right\},$$

be a modified version GFI for complex survey data, when considering $F(\underline{\theta})_{PML}$. We also modify Tanaka and Huba (1985) GLS version of goodness of fit indices, so that

$$\mathrm{GFI}^1_{c,GLSC} = 1 - \left\{ \frac{tr[(I - \Sigma(\underline{\hat{\theta}})\mathrm{S}_w{}^{-1})^2]}{T} \right\}, \text{ or}$$

$$\mathrm{GFI}^2_{c,GLSC} = 1 - \left\{ \frac{\{vech[\mathrm{S}_w] - vech[\Sigma(\underline{\hat{\theta}})]\}' \mathrm{U}^{-1} \{vech[\mathrm{S}_w] - vech[\Sigma(\underline{\hat{\theta}})]\}}{vech[\mathrm{S}_w]' \mathrm{U}^{-1} vech[\mathrm{S}_w]} \right\}.$$

Moreover, modified adjusted fit indices (AGFI_c), could be calculated as

$$\mathrm{AGFI}_c = 1 - \left(\frac{k}{\mathrm{df}} \right) \cdot (1 - \mathrm{GFI}_c).$$

5 Concluding remarks

Model testing is an important step in any model fit procedure. According to Menard (1991), in longitudinal models there is an increase to problems comparably to cross-sectional models in this regard. Eltinge (1999) acknowledged the need to improve and to develop new techniques for model assessment and diagnostic in the complex survey data context. Classic measures that are often used in model testing are appropriate for situations where data is obtained from a srs design.

We propose some new developments on model fitting statistics when working with longitudinal data in a complex survey design framework. Following Rao and Scott's conceptions (Rao and Scott, 1979), we propose modifying the Wald goodness of fit test in the context of models for covariance structures. Furthermore, we also propose a modification for the Wald significance test for nested hypothesis, following an approach suggested by Skinner (1989, Section 3.4). Goodness of fit indices proposed by Jöreskog and Sörbom (1989) are also modified in order be utilised in the complex survey data context.

Further research could involve evaluating the consequences of using standard model fit test techniques without consideration to the complexity of the sample, by extending the simulation study performed in Vieira and Skinner (2008).

Acknowledgement: The author thanks FAPEMIG for grant number CEX-APQ-02071-12 (Universal), and for supporting the attendance at the 59th WSC of the ISI.

References

Binder, D. A. (1983) On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-292.

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

Chambers, R. L. and Skinner, C. J. eds. (2003) *Analysis of Survey Data*. Chichester, John Wiley & Sons.

Diggle, P. J., Heagerty, P., Liang, K. & Zeger, S. L. (2002) *Analysis of Longitudinal Data*. 2nd ed. Oxford, Oxford University Press.

Eltinge, J. L. (1999) *Assessment of Information Capacity and Sensitivity in the Analysis of Complex Surveys*. Bulletin of the International Statistical Institute. 52nd Session Proceedings.

Jöreskog, K. G. and Sörbom, D. (1989) *Lisrel 7: A Guide to the Program and Applications*. Chicago, SPSS publications.

Lindsey, J. K. (1994) *Models for Repeated Measurements*. Oxford, Oxford University Press.

Rao, J. N. K. and Scott (1979) Chi-squared Tests for Analysis of Categorical Data from Complex Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Satorra, A. (1989) Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach. *Psychometrika*, Vol. 54, N. 1, p. 131-151.

Skinner, C. J. (1989) Domain Means, Regression and Multivariate Analysis. In Skinner, C. J., Holt, D. and Smith, T. M. F. eds. *Analysis of Complex Surveys*. Chichester, John Wiley & Sons.

Skinner, C. J. and Holmes, D. (2003) Random Effects Models for Longitudinal Survey Data. In Chambers, R. L. and Skinner, C. J. eds. *Analysis of Survey Data*. Chichester, John Wiley & Sons.

Skinner, C. J., Holt, D. and Smith, T. M. F. eds. (1989) *Analysis of Complex Surveys*. Chichester, John Wiley & Sons.

Skinner, C. J. and Vieira, M. D. T. (2007) Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*, Vol. 33, No. 1, pp. 3-12.

Tanaka, J. S., and Huba, G. J. (1985) A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, Vol. 38, 197-201.

Vieira, M. D. T. and Skinner, C. J. (2008) Estimating Models for Panel Survey Data under Complex Sampling. *Journal of Official Statistics*, 24, 343-364.

Vieira, M. D. T. (2009) *Analysis of Longitudinal Survey Data*. 1. ed. Saarbrücken: VDM Verlag Dr. Müller.