A parametric test to discriminate between a linear regression model and a linear latent growth model

Marco Barnabani
Department of Statistics, Informatics, Applications
V.le Morgagni, 59
50134 Florence, Italy.
e-mail: barnaban@disia.unifi.it

Abstract

In longitudinal studies with subjects measured repeatedly across time, an important problem is how to select a model generating data by choosing between a linear regression model and a linear latent growth model. Approaches based both on information criteria and asymptotic hypothesis tests of the variances of "random" components are widely used but not completely satisfactory. We propose a test statistic based on the trace of the product of an estimate of a variance covariance matrix defined when data come from a linear regression model and a sample variance covariance matrix. We studied the sampling distribution of the test statistic giving a representation in terms of an infinite series of generalized F-distributions. Knowledge about this distribution allows us to make inference within a classical hypothesis testing framework. The test statistic can be used by itself to discriminate between the two models and/or, if duly modified, it can be used to test randomness on single components. Moreover, in conjunction with some AIC indicators it gives additional information which can help in choosing the model.

keywords: Generalized F-distribution; Hypothesis testing; Linear Mixed Models; Longitudinal data.

1 Introduction

It is common practice in many applications to collect multiple measurements on subjects across time by focusing on the process of change when, typically, both data dependency and differential growth for different individuals can occur. If we assume that the subjects constitute a sample from the population of interest and we wish to draw conclusions about typical patterns in the population and the subject-to-subject variability of these patterns, we are fitting linear latent growth models. In this paper these models are analyzed by using a mixed-modeling framework (Laird and Ware, 1982). Linear mixed models can be viewed as extensions of linear regression models and attempt to account for within-subject dependency in the multiple measurements by including one or more subject-specific latent variables in the regression model. Typically, an additional random effect is included for each regression coefficient that is expected to vary among subjects. An important practical problem is how to discriminate between a linear regression model and a linear mixed

model and how to choose the random effect components. In order to address the issue of which model is more suitable, one might use standard model selection measures based on information criteria such as the widely used Akaike Information Criteria (AIC; Akaike (1973)), the Bayesian Information Criteria (BIC; Schwarz (1978)) the conditional Akaike Information Criterion ($cAIC$, Vaida and Blanchard (2005)). These approaches are based on the choice of models that minimize an estimate of a specific criterion which usually involves a trade-off between the closeness of the fit to the data and the complexity of the model. We refer to the paper of Muller et al. (2013) for a review of these approaches and other methods such as shrinkage methods like the LASSO (Tibshirani, 1996), Fence methods (Jiang et al., 2008) and Bayesian methods.

The validity of all the methods proposed depends on the underlying assumptions. The review paper of Muller et al. (2013) gives an overview of the limits and most important findings of above approaches, extracting information from some published simulation results. As is known, one of the major drawbacks of these approaches is that they fail to give any measure of the degree of uncertainty of the model chosen. The value they produce does not mean anything by itself.

Alternatively, because model selection is closely related to hypothesis testing, the choice between a linear regression model (LRM) and a linear latent growth model ($LLGM$) and the evaluation of its uncertainty could be conducted considering a formal hypothesis test on the variances of "random" components. Noting that models are nested, it is natural to consider the likelihood ratio test. However, the difficulty with this is that it makes the usual approach of comparing the likelihood ratio test statistic with the chi-square distribution inappropriate. The question of whether the variance of a component is zero depends on whether said variance takes its value on the boundary of the parameter space. This situation is known as "non-standard" in relation to the other uses of the likelihood ratio test. The major consequence is that in large samples -2 times the logarithm of the likelihood ratio cannot be treated as a chi-square distribution, but rather, as a mixture of chi-square distributions. Determining the weights of this mixture distribution is difficult especially for testing multiple variance components or a subset of them. For more details see, for example, Self and Liang (1987), Stram and Lee (1994), Verbeke and Molenberghs (2003), Giampaoli and Singer (2009) . Comparing the likelihood ratio statistic with the critical value from a chi-square sampling distribution tends not to reject the null as often as it should. Other tests not based on the likelihood function can be implemented (Silvapulle and Sen, 2005) but their validity should be carefully detected when applied to linear mixed models. Moreover, all these tests are only valid asymptotically. Finite sample distributions of the likelihood ratio test require simulations and are only reported in particular cases. For example, Crainiceanu and Ruppert (2004) introduced an efficient simulation algorithm based on the spectral representations of the likelihood ratio test and the restricted likelihood ratio test statistics for models with a single variance component.

When we extend the analysis to multiple variance components, the complexity and difficulties increase. In these cases we have to consider variance covariance matrices and the problem of testing the equality of two positive definite matrices. Hypothesis testing approaches based on the equality of two positive definite matrices have a distinguished history in multivariate statistics. In most cases the likelihood ratio approach is used and the resulting test statistics involve the ratio of the determinant of the sample covariance matrix under the null hypothesis and the alternative hypothesis. Other researchers have studied tests based on the trace of two covariance matrices. Roy (1953), Pillai (1955), Pillai and Jayachandran (1968) and Nagao (1973) develop trace-based tests and compare their performance to that of determinant-based tests. The trace test proposed by Pillai

for testing the equality of two variance covariance matrices appears to be useful in discriminating between an LRM and an $LLGM$ by appropriately defining the two matrices involved.

Let's denote with V the variance covariance matrix of the ordinary least square estimators when data come from an LRM and let $V + \Omega$ be the variance covariance matrix of the same estimators when data come from an $LLGM$ where Ω denotes the covariance matrix of the random effects. The Pillai test statistic proposed in the paper is based on an estimate of $\frac{1}{k} tr\ V^{-1}(V + \Omega)$ with Ω that plays a a crucial role in discriminating between the two models. If Ω is a positive semi definite matrix, $\Omega \succeq 0$, $\frac{1}{k} tr\ V^{-1}(V + \Omega) > 1$. In this case we can state that data come from an $LLGM$. If $\Omega = 0$, $\frac{1}{k} tr\ V^{-1}(V + \Omega) = 1$ and data come from an LRM. In section 2 the test statistic is defined after introducing several notations. In section 3 we analyze the sampling distribution. When data come from an LRM it has a "standard" F-distribution, when data come from an $LLGM$ the sampling distribution is more complex. It is a linear combination of standard F-distributions. The exact form of a linear combination of standard F-distributions is studied. Following the work of Kourouklis and Moschopoulos (1985) a unified sampling distribution involving a generalized F-distribution is proposed which is based on an infinite series representation and is relatively easy to implement. In section 4 we discuss the test statistic to make inference. In section 5, we analyze a slight modification of the test so that inference on randomness of single components of the model is possible. Finally two applications are investigated. In section 6 we applied the test to a data set on tourism. Said data set is sufficiently "regular" to allow a clear-cut answer regarding the choice of the model. The answers produced by the test are not in conflict with those given by AIC indicators. The advantage derived from the hypothesis testing approach is that we can attach a measure of the degree of uncertainty to the choice of the model. In section 7 the test is applied to a Cadralazine data set previously analyzed by Vaida and Blanchard (2005). In this case different AIC indicators applied to the data set do not give clear-cut indications about the model. There is a substantial indeterminacy which also remains when using the test proposed in this paper but we can still provide additional information by computing an estimate of the probability of accepting the "wrong" model.

2 Notation and test statistic

Let us suppose that t observations on the i-th of n units are described by the model $y_i = X\beta_i + u_i$, $i = 1, \ldots, n$, where X is a $t \times k$ matrix containing a column of ones and a column of constant time values, β_i is a $k \times 1$ vector of coefficients unique to the i-th experimental unit, and u_i is a $t \times 1$ vector whose component is the measurement error at a time point for individual i.

Let us also suppose that each experimental unit and its response curve is considered to be selected from a larger population of response curves; thus the regression coefficient vectors β_i may be viewed as random drawings from some k-variate population: $\beta_i = \theta + v_i$, $i = 1, \ldots, n$, where v_i is an unobserved random variable that configures individual growth.

In this paper we discuss testing under the following assumptions: (a) $u_i \sim N\left(0, \sigma^2 I_t\right)$, (b) $v_i \sim N(0, \Omega)$, Ω is a positive semi definite matrix, (c) $u_i \perp v_i$, where the symbol $\perp$ indicates independence of random variables (d) $\beta_i \perp u_i$. We refer to this model as a linear latent growth model.

If $\Omega = 0$, then the regression coefficients are fixed. We refer to this model as a linear regression model. The normality assumptions are introduced for testing purposes.

By replacing the random component in the model we have

$$y_i = X\theta + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, X\Omega X' + \sigma^2 I_t\right)$$

Let $b_i = (X'X)^{-1}X'y_i$ be the ordinary least square estimators of θ computed for each individual unit. Note that the b_i's are independent and normally distributed with mean θ and variance-covariance matrix $\sigma^2(X'X)^{-1} + \Omega$. Let $S_b = (n-1)^{-1}\sum_{i=1}^{n}\left(b_i - \overline{b}\right)\left(b_i - \overline{b}\right)'$ be the sample variance covariance matrix of b_i with $\overline{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$. When data come from an $LLGM$, S_b is an unbiased consistent estimate of $\sigma^2(X'X)^{-1} + \Omega$ (Gumpertz and Pantula, 1989) when data come from an LRM S_b is an unbiased consistent estimate of $\sigma^2(X'X)^{-1}$.

To discriminate between an LRM or an $LLGM$ we propose the following test statistic

$$T = \frac{1}{k}\, tr\, \frac{(X'X)S_b}{s^2} \tag{1}$$

where $s^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$, with $s_i^2 = \frac{(y_i - Xb_i)'(y_i - Xb_i)}{T-k}$ (Swamy, 1970).

When data come from a linear regression model ($\Omega = 0$), $(1/s^2)(X'X)$ is "close" to S_b and we expect the test statistic T to be approximately equal to one. When data come from an $LLGM$ we expect that $T > 1$. The greater T the stronger is the evidence against an LRM.

The sampling distribution of T is analyzed in the next section.

Observe that the inverse of $\frac{(X'X)S_b}{s^2}$ can be seen as an estimate of $s^2(X'X)^{-1}\left[s^2(X'X)^{-1} + \Omega\right]^{-1}$ the trace of which (divided by k) has been proposed by Theil (1963) to measure the shares of prior and sample information in the posterior precision in the mixed regression estimation (Barnabani, 2014).

3 Sampling distribution of test statistic

When data come from an LRM, $\Omega = 0$ and $(n-1)S_b/\sigma^2 \sim W_k\left((X'X)^{-1}, n-1\right)$ (W_k is for Wishart distribution). Then, $(n-1)(X'X)^{1/2}\frac{S_b}{\sigma^2}(X'X)^{1/2} \sim W_k\left(I, n-1\right)$ where $(X'X)^{1/2}$ is the square root of $(X'X)$. We have the following results

(i) $(n-1)s_{ii}/\sigma^2 \sim \chi^2_{n-1}$ where s_{ii}, $i = 1, \ldots, k$ is the $i-th$ diagonal element of $(X'X)^{1/2}S_b(X'X)^{1/2}$. Replacing σ^2 by s^2 we have

$$\frac{(n-1)s_{ii}}{s^2} = \frac{(n-1)s_{ii}/\sigma^2}{\frac{n(t-k)s^2}{n(t-k)\sigma^2}} \sim \frac{\chi^2_{n-1}}{\chi^2_{n(t-k)}/n(t-k)} \tag{2}$$

and

$$\frac{s_{ii}}{s^2} \sim F_{n-1,n(t-k)} \tag{3}$$

(ii) By independence $\sum_{i=1}^{k}(n-1)s_{ii}/\sigma^2 \sim \chi^2_{k(n-1)}$ and

$$\frac{1}{k}\, tr\, \frac{(X'X)S_b}{\sigma^2} \sim \frac{\chi^2_{k(n-1)}}{k(n-1)}$$

because $tr\left((X'X)^{1/2}\frac{S_b}{\sigma^2}(X'X)^{1/2}\right) = tr\,(X'X)\,\frac{S_b}{\sigma^2}$.
With the following equality

$$\frac{1}{k}\,tr\,\frac{(X'X)S_b}{s^2} = \frac{1}{k}tr\,\frac{(X'X)S_b}{s^2}\,\frac{\sigma^2}{\sigma^2}\,\frac{n(t-k)}{n(t-k)}$$

we derive the sampling distribution of T

$$T \sim F_{k(n-1),n(t-k)} \tag{4}$$

When data come from an $LLGM$ $(n-1)S_b/\sigma^2 \sim W_k\left[(X'X)^{-1} + \Omega/\sigma^2, n-1\right]$. Therefore, a non singular matrix Q exists such that $\frac{n-1}{\sigma^2}Q'S_bQ \sim W_k\left(I + \frac{D}{\sigma^2}, n-1\right)$ where D is a diagonal matrix of eigenvalues $\eta_i \geq 0$ of the matrix $(X'X)^{1/2}\Omega(X'X)^{1/2}$ and
$trQ'S_bQ = tr\left((X'X)^{1/2}S_b(X'X)^{1/2}\right) = tr\,(X'X)\,S_b$. We have the following results:

(i) $(n-1)\,{}_os_{ii}/\sigma^2 \sim \left(1+\eta_i/\sigma^2\right)\chi^2_{n-1}$ where ${}_os_{ii}$ denotes the $i-th$ diagonal element of $Q'S_bQ$ and

$$\frac{{}_os_{ii}}{s^2} \sim \left(1+\eta_i/\sigma^2\right)F_{n-1,n(t-k)} \tag{5}$$

(ii) As regards the distribution of T, it must be observed that

$$\sum_{i=1}^{k}(n-1)_os_{ii}/\sigma^2 = \sum_{i=1}^{k}(n-1)s_{ii}/\sigma^2 \sim \sum_{i=1}^{k}\left(1+\frac{\eta_i}{\sigma^2}\right)\chi^2_{n-1} \tag{6}$$

When we replace σ^2 by s^2, we have

$$T = \frac{1}{k}\sum_{i=1}^{k}\frac{s_{ii}}{s^2} \sim \frac{1}{k}\sum_{i=1}^{k}\left(1+\frac{\eta_i}{\sigma^2}\right)F_{(n-1),n(t-k)} \tag{7}$$

The above sampling distributions are now reproposed in terms of Generalized Fisher-distribution (GF-distribution). This is necessary because (7) is difficult to implement in practice and it does not allow for computing the power of the test.

Let us consider (2). The statistic can be seen as the ratio of two independent gamma random variables where the numerator is distributed as $G\left(\alpha = \frac{n-1}{2}, \lambda_1 = 2n(t-k)\right)$ and the denominator is distributed as $G\left(\gamma = \frac{n(t-k)}{2}, \lambda_2 = 2\right)$ where $G(.,.)$ is for gamma distribution, α and γ are shape parameters, λ_1 and λ_2 scale parameters. The distribution of the ratio, Z, is called GF-distribution and has pdf (Malik, 1967)

$$f(z) = \frac{\delta^\gamma}{B(\alpha,\gamma)}\,(z+\delta)^{-(\alpha+\gamma)}\,z^{\alpha-1} \tag{8}$$

where $B(\alpha,\gamma)$ is the Beta function, $\delta = \lambda_1/\lambda_2$. Expression (8) is also known as Compound Gamma Distribution (Dubey, 1970). Therefore, we have

$$\frac{(n-1)s_{ii}}{s^2} \sim GF\,(\delta,\alpha,\gamma) \tag{9}$$

The standard F-distribution (4) can be seen as a GF-distribution with $\delta = n(t-k)/k(n-1)$, $\alpha = k(n-1)/2$, $\gamma = n(t-k)/2$.

The distribution given by (5) is a scalar multiple of a F variate which is a GF-distribution with $\delta = n(t-k)\left(1+\eta_i/\sigma^2\right)/(n-1)$, $\alpha = (n-1)/2$ and $\gamma = n(t-k)/2$.

The result given by (7) is a linear combination of independent F variates whose distribution does not admit a closed and simple form. However, because it can be seen as a linear combination of ratios of independent gamma variates, the gamma-series representation proposed by Kourouklis and Moschopoulos (1985) and Moschopoulos (1985) is particularly useful for defining the distribution of (7). Following these papers we have

$$\sum_{i=1}^{k} \frac{(n-1)s_{ii}}{\sigma^2} \sim \sum_{l=0}^{\infty} w_l \; G\left(\rho + l, 2\,\eta\right)$$

where $0 < \eta < \infty$ is arbitrary.

In the expression of the series, $\rho = \sum_{i=1}^{k} \alpha_i = (n-1)k/2$, $w_l = C d_l$, $l = 0, 1, 2, \ldots$, $d_0 = 1$, $C = \prod_{i=1}^{k} \left(\eta/(1+\frac{\eta_i}{\sigma^2})\right)^{\alpha_i}$, $d_l = (1/l)\sum_{i=1}^{l} i\, g_i\, d_{l-i}$ with $g_i = (1/i)\sum_{j=1}^{k} \alpha_j \left(1-\eta/(1+\frac{\eta_j}{\sigma^2})\right)^i$.

When we replace σ^2 by s^2, we have

$$\sum_{i=1}^{k} \frac{(n-1)s_{ii}}{s^2} = \frac{\sum_{i=1}^{k}(n-1)s_{ii}/\sigma^2}{\frac{n(t-k)s^2}{n(t-k)\sigma^2}} \sim \sum_{l=0}^{\infty} w_l \; \frac{G\left(\rho + l, 2\,\eta\, n(t-k)\right)}{G\left(n(t-k)/2, 2\right)} \tag{10}$$

Finally, from (10) we have the distribution of the trace,

$$T \sim \sum_{l=0}^{\infty} w_l \; GF\left(\delta, \alpha, \gamma\right) \tag{11}$$

with $\delta = \frac{n(t-k)}{k(n-1)}\,\eta$.

The series representation of the GF-distribution is not difficult to implement in practice and in most statistical software there is a function that computes the generalized F-distribution. In this paper computations are made with R (R Core Team, 2014) where a library (GB2) (or flexsurv) allows us to compute density, distribution function, quantile function and random generation for the GF-distribution.

The weights of the series representation can be troublesome to implement. Moreover, their computation can become too CPU-time consuming. In these cases η may be adjusted to make the convergence of the series faster (Kourouklis and Moschopoulos, 1985).

When the variability of the scale parameters is large and/or the shape parameters are small the convergence of the weights is extremely slow. This fact can discourage a large-scale simulation and application of the expression proposed and an approximation of the weights is needed. For $\eta \leq \min\{\eta_j : j = 1, \ldots, k\}$ the weights, w_l, define probabilities of an infinite discrete distribution (Vellaisamy and Upadhye, 2009) and they can be approximated by a theoretical discrete distribution. For more than two random variables, Barnabani (2015) proposed to approximate these probabilities with the generalized negative binomial distribution of Jain and Consul (1971) resulting in a fast and "excellent" approximation. For two linear independent random variables,

simple algebra shows that the weights are described exactly by a negative binomial distribution (Barnabani, 2015). The infinite discrete distribution $(l, w_l)_{0,1,2,...}$ must be truncated after the desired accuracy.

4 Inference on the model

When data come from an $LLGM$, the sampling distribution of T depends on $\eta_i \geq 0$, the eigenvalues of the matrix $(X'X)^{1/2}\,\Omega\,(X'X)^{1/2}$. The expected value of T is given by $E(T) = \frac{n(t-k)}{n(t-k)-2}\overline{\eta}$ where $\overline{\eta} = (1/k)\sum_{i=1}^{k}(1+\eta_i/\sigma^2)$. We can observe that $\overline{\eta} = 1 \Leftrightarrow \Omega = 0$ that is, if and only if data come from an LRM; $\overline{\eta} > 1 \Leftrightarrow \Omega \succeq 0$ if and only if data come from an $LLGM$. In the first case the estimator T has a GF-distribution (F-distribution), in the second case T has an infinite series representation of GF-distributions. $\overline{\eta} > 1$ occurs when at least one eigenvalue is greater than zero. The term $\frac{\eta_i}{\sigma^2}$ can be seen as the extra factor due to the $i-th$ random effect. It is zero when the random effect does not occur.

"Natural" estimators of η_i's are $\widehat{\eta}_i$'s $i = 1, \ldots, k$, the eigenvalues of $(X'X)^{1/2}\,\widehat{\Omega}\,(X'X)^{1/2}$ where $\widehat{\Omega}$ is an estimate of Ω. $\widehat{\Omega}$ can be estimated in several ways. Following Swamy (1970) we define $\widehat{\Omega} = S_b - s^2(X'X)^{-1}$ as a difference of two matrices. This definition can yield negative estimates for variances of some of the coefficients and/or may not be a positive definite matrix. In this case we could have negative eigenvalues. Although negative $\widehat{\eta}_i$ could appear to be misleading, the definition of $\widehat{\Omega}$ is coherent with the above sampling distributions and allows for obtaining the equality $T = \frac{1}{k}\,tr\,\frac{(X'X)S_b}{s^2} = (1/k)\sum_{i=1}^{k}(1+\widehat{\eta}_i/s^2)$ which shows that T can be seen as an estimate of $\overline{\eta}$.

The models describing T are different for the two data sources. The series representation of GF-distribution used to describe an $LLGM$ contains the other as a special case constraining the parameter $\overline{\eta}$ to one. We call the more general model the alternative hypothesis and the restricted model the null hypothesis. We can make inference by defining the null hypothesis $H_0 : \overline{\eta} = 1$ ($H_0 : \overline{\eta} \leq 1$) against the alternative $H_1 : \overline{\eta} > 1$. Thus, H_0 is rejected if T is "much" greater than one.

The knowledge of η_i/σ^2 is necessary to compute the probability of making a Type II error and/or to compute the probability of rejecting a false null hypothesis. Usually this knowledge is not available and only an estimate of these probabilities is possible by replacing σ^2 with s^2 and η_i with $\widehat{\eta}_i$. When n is large then the probabilities are accurate.

5 Inference on a single component

If T is greater than a critical value or the $p-value$ is small, then likely data come from an $LLGM$ and it is important to investigate which component is random.

When data come from an $LLGM$ the sampling distribution of T depends on $\left(1+\eta_i/\sigma^2\right)$ with η_i/σ^2 that can be seen as the extra factor due to the random effect. An estimate of this parameter replacing η_i with $\widehat{\eta}_i$ and σ^2 with s^2 can help to identify the number of random components but not which component is random. Therefore, we propose to modify the extra factor, η_i/σ^2, replacing η_i with ω_{ii} and σ^2 with $\sigma^2 x^{ii}$ where ω_{ii} is the entry (i,i) of the matrix Ω and x^{ii} the entry (i,i) of the matrix $(X'X)^{-1}$. The "new" parameter, $\phi_i = \left(1+\frac{\omega_{ii}}{\sigma^2 x^{ii}}\right)$, expresses the extent of "total"

variability of the $i-th$ coefficient ($\sigma^2 x^{ii}+\omega_{ii}$) in relation to the "residual" variance $\sigma^2 x^{ii}$. Given a finite $\sigma^2 > 0$ and varying ω_{ii}, ϕ_i is greater than one and it measures how far we move from a situation of zero variance. The greater the value of ϕ_i the stronger this evidence. When $\omega_{ii} = 0$ the parameter ϕ_i is equal to one and the $i-th$ component is zero variance. Given that $\omega_{ii} > 0$ and increasing σ^2, ϕ_i approaches one.

The reciprocal of ϕ_i, $\phi_i^{-1} = \frac{\sigma^2 x^{ii}}{\sigma^2 x^{ii}+\omega_{ii}}$, can be seen as the share of "residual" variance on the "total" variability. It ranges between zero and one. When $\omega_{ii} > 0$, $\phi_i^{-1} < 1$ and we face a randomness on the $i-th$ component. When $\omega_{ii} = 0$, $\phi_i^{-1} = 1$ and the $i-th$ component is zero variance. Observe that ϕ_i^{-1} can be seen as a scalar form of the matrix product $\sigma^2(X'X)^{-1}\left[\sigma^2(X'X)^{-1}+\Omega\right]^{-1}$ the trace of which (divided by k) has been proposed by Theil (1963) to measure the shares of prior and sample information in the posterior precision in the mixed regression estimation.

A "natural" estimator of ϕ_i is $\widehat{\phi}_i = 1 + \frac{\widehat{\omega}_{ii}}{s^2 x^{ii}}$ where $\widehat{\omega}_{ii}$ is the entry (i,i) of the matrix $\widehat{\Omega}$. The sampling distribution of $\widehat{\phi}_i$ is immediate. Because of the equality $\widehat{\phi}_i = 1 + \frac{\widehat{\omega}_{ii}}{s^2 x^{ii}} = \frac{\widehat{s}_{ii}}{s^2 x^{ii}}$ where $\widehat{s}_{ii}$ is the (i,i) entry of the matrix S_b, we have

$$\widehat{\phi}_i \sim \left(1 + \frac{\omega_{ii}}{\sigma^2 x^{ii}}\right) F_{(n-1),n(t-k)} \tag{12}$$

which is a scale multiple of an F variate and can be seen as an GF-distribution with
$\delta = n(t-k)\left(1+\frac{\omega_{ii}}{\sigma^2 x^{ii}}\right)/(n-1)$, $\alpha = (n-1)/2$ and $\gamma = n(t-k)/2$.

The sampling distribution (12) is obtained by observing that $(n-1)S_b/\sigma^2 \sim W_k\left((X'X)^{-1}+\Omega/\sigma^2, n-1\right)$ and $(n-1)\widehat{s}_{ii}/\sigma^2 \sim \left(x^{ii}+\frac{\omega_{ii}}{\sigma^2}\right)\chi^2_{n-1}$. This implies that $\frac{\widehat{s}_{ii}}{\sigma^2 x^{ii}} \sim \left(1+\frac{\omega_{ii}}{\sigma^2 x^{ii}}\right)\frac{\chi^2_{n-1}}{n-1}$, replacing σ^2 with s^2 we get (12).

When data come from an LRM, $\omega_{ii} = 0$ and $\phi_i = 1$. We define $H_0 : \phi_i = 1$ ($H_0 : \phi_i \leq 1$) the null hypothesis. In this case the estimate $\widehat{\omega}_{ii}$ can assume values that are greater or less than zero with $\widehat{\phi}_i$ ranging around one according to an $F-$distribution. Actually, $\widehat{\omega}_{ii} \leq 0$ if and only if $\widehat{\phi}_i \leq 1$ and the probability $P(\widehat{\omega}_{ii} \leq 0)$ can be computed with the $F-$distribution. If data come from an $LLGM$, $\omega_{ii} > 0$ and $\phi_i > 1$. We call $H_1 : \phi_i > 1$ the alternative hypothesis. In this case the estimate $\widehat{\omega}_{ii}$ can still assume values that are greater or less than zero but the negative values become increasingly less frequent the stronger the evidence against the null hypothesis. Thus, the null hypothesis H_0 is rejected if $\widehat{\phi}_i$ is "much" greater than one. Of course a $p-value$ can also be computed.

A "confounding" situation can appear when the "residual" variance $\sigma^2 x^{ii}$ is large compared with the elements of Ω. In this case ϕ_i could be close to one and the test statistic $\widehat{\phi}_i$ has a GF-distribution close to an $F-$distribution. In this case there is a large probability of failing to reject the null hypothesis in favor of the alternative. This problem is clearly explained, for example, in the work of Gumpertz and Pantula (1989).

Observe that $\frac{\widehat{s}_{ii}}{s^2 x^{ii}}$ is a pivotal quantity and a confidence interval for ϕ_i can be computed when data come from an $LLGM$. Fixing α we can determine two percentiles of F-distribution such that

$$P\left(F_{(n-1),n(t-k),1-\alpha/2}\ \frac{s^2 x^{ii}}{\widehat{s}_{ii}} \leq \phi_i^{-1} \leq F_{n-1,n(t-k),\alpha/2}\ \frac{s^2 x^{ii}}{\widehat{s}_{ii}}\right) = 1-\alpha \tag{13}$$

Thus, if data come from an $i-th$ random component, we can compute a confidence interval for the share. This result can give further information about the choice of random components. If we automatically compute the confidence interval for each component we could face two situations: (a) an interval contained in $(0, 1)$, in which case the component is presumably random, and (b) an interval around one, in which case a substantial indeterminacy occurs. We could have a zero variance component or a random component with σ^2 which dominates the variance of the component, thereby confounding the choice.

6 An application: Tourism data

A data set on Tourism in Tuscany (Italy) consists of the index number (base year 2002) of accommodations (the response variable) on 260 Municipalities from 2003 to 2009. These data were first processed in order to obtain homogeneous groups of units. In the paper we work with 98 "homogeneous" Municipalities. Trajectories of index numbers of this group are plotted in the left panel of Fig.: 1. By observing the tourism data, each unit appears to have its own trajectory

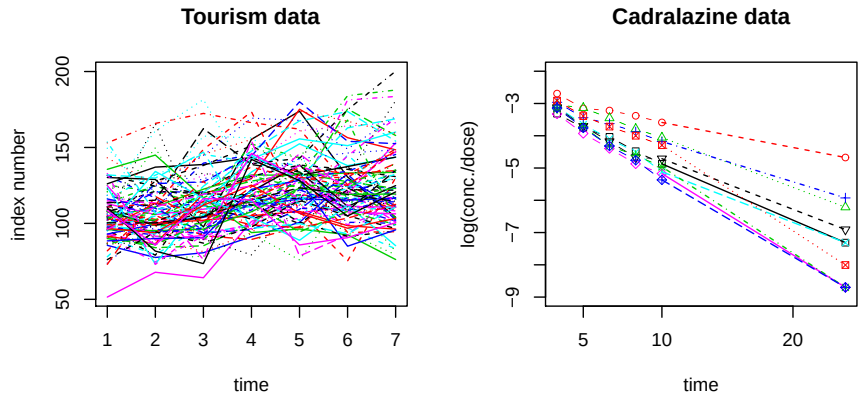

Figure 1: Spaghetti plots for Tourism data and Cadralazine data

approximated by linear functions with specific intercepts and slopes.

The trajectories are "high" or "low" suggesting two hypotheses from an economic point of view. One is that the growth of tourism in each Municipality at time t could be determined solely by an overall regional political economy. Statistically this is modeled with a vector of fixed population parameters which capture the regional political economy plus an overall random deviation from the same.

On the other hand, data show different steepness across Municipalities, suggesting that the unit-specific intercepts and slopes could not be fixed but vary across units with a growth in tourism influenced not only by the regional political economy, but also by specific characteristics of each Municipality. This suggests that data could be modeled by adding a random component to the parameter vector in order to distinguish the various trajectories.

Statistically we ask whether it is more appropriate to model data with a linear regression model or a linear latent growth model.

By applying the hypothesis testing approach proposed in this paper we can make the following comments:

- We found a value of the test statistic $T = 4.76$, which, when compared with the critical value $F_{194,490,0.95} = 1.212$, falls into the rejection region. Consequently, we reject the hypothesis that data come from an LRM. The computation of the probability of Type II error requires Ω which is unknown. Replacing the variance covariance matrix with $\widehat{\Omega}$ we observe an estimated (conditional) probability close to zero.
- The $p - value$ is close to zero confirming strong evidence against the null hypothesis.
- $\widehat{\phi}_1 = 3.245$ and $\widehat{\phi}_2 = 3.7313$ compared with $F_{97,490,0.95} = 1.279$ confirm that both components are random.
- The confidence intervals of the shares are: $0.21719 \leq \phi_1^{-1} \leq 0.40331$ and $0.19351 \leq \phi_2^{-1} \leq 0.3593$. It is likely that the "true" shares belong to the interval $(0, 1)$, thus confirming randomness on both coefficients.

The above results are also compared with several indicators normally used in model selection. These indicators are computed with the package $lme4$ (Bates et al., 2014) of R (R Core Team, 2014). The results are shown in table 1 All the above indicators confirm the choice of a linear

	AIC	BIC
LRM	6122.482	6136.075
$LLGM$	5925.687	5952.872

Table 1: AIC indicators for the linear regression model and linear latent growth model for Tourism data.

latent growth model to describe data.

We also computed the conditional AIC proposed by Vaida and Blanchard (2005), defined for linear mixed models only and not comparable with other indicators. The value it produces, $cAIC = 5776.097$, does not mean anything by itself and the fact that it is less than the others, does not mean that the $LLGM$ must be chosen. However, with the test proposed in this paper we can give an estimate of the degree on uncertainty to accept the mixed model. This will be done for the next application.

7 An application: Cadralazine data

In the previous section we discussed a data set which made it possible to give clear and evident answer about the choice of the model. To illustrate some difficulties that could arise when discriminating between a linear regression model and a latent growth model let us consider the case study of a pharmacokinetics dataset, the Cadralazine data, analyzed in the paper of Vaida and Blanchard (2005) to which we refer for further explanations of data. The dataset consists of plasma drug concentrations from 10 cardiac failure patients who were given a single intravenous dose of 30 mg of cadralazine, an anti-hypertensive drug. Each subject has the plasma drug concentration, in mg/l, measured at 2, 4, 6, 8, 10 and 24 hours, for a total of 6 observations per subject. The plot of

the response versus time is given in the right panel of Fig.: 1. The data for each patient are well described by a straight line, but the slopes and intercepts of the ten regression lines differ from subject to subject. Two models are proposed, a linear regression model with fixed intercepts and slopes, and a mixed effects model with random intercepts and slopes.

The choice between the two models is first conducted through several AIC type indicators. If

	AIC	BIC
LRM	161.717	168.0
$LLGM$	157.923	170.5

Table 2: AIC indicators for the linear regression model and linear latent growth model for Cadralazine data.

we compare the AIC and BIC indicators in table 2, we can see that there is substantial indeterminacy. They produce conflicting results with the AIC indicating that we should choose an $LLGM$, while the BIC value gives a different interpretation, thus reversing the choice. Moreover, observe the differences between the indicators computed, $\Delta AIC = 161.72(LRM) - 157.92(LLGM) = 3.794$, $\Delta BIC = 170.5(LLGM) - 168.0(LRM) = 2.5$ This values appear to be "low" even though they do not mean anything.

The indeterminacy emerging in this example is not removed with the test proposed in this paper, however it may provide additional information to help choose the model:

- We found a value of the test statistic $T = 1.7829$, which when compared with $F_{18,40,0.95} = 1.8682$ falls into the acceptance region. Therefore, we fail to reject an LRM. The closeness of the observed value to the critical value suggests caution in choosing the model. Indeed, we found a $p - value = 0.0639$ that confirms our caution. These results reflect the indeterminacy of both AIC and BIC indicators.
- In this application the probability of a Type II error is important for quantifying the uncertainty of the model chosen, however the computation requires knowledge about Ω. Unless some information is available, the best we can do is to replace the "true" variance covariance matrix with $\widehat{\Omega}$ estimated by the data. This allows for estimating the GF-distribution under the alternative hypothesis. The result is a conditional probability, $P\left(T \leq F_{18,40,0.95} | \Omega = \widehat{\Omega}\right) = 0.58$ that could be taken as an estimate of the probability of the Type II error. Therefore, while the BIC indicator suggests the choice of an LRM, we must also point out that there is a "large" estimated probability of failing to reject the false model on the basis of information contained in the data set. See Fig.: 2 (a). We can also say that 0.58 is the degree of uncertainty associated to the choice of an LRM.
- The AIC indicator guides the choice towards an $LLGM$ (see Tab.: 2). We also computed the conditional AIC showing a lower, though not comparable, value than the other AIC indicators. Therefore, given a substantial indeterminacy in the choice of the model, with the test statistic proposed in the paper we can give an estimate of the degree of uncertainty of accepting an $LLGM$ instead of an LRM. We suggest proceeding as follows:
    1. Estimate the variance covariance matrix $\widehat{\Omega}$. Here we proceed with an estimate produced by the package $lme4$ of R even though another estimate is plausible,

$$\widehat{\Omega} = \begin{bmatrix} 0.00054686 & 0.003727 \\ 0.003727 & 0.025400 \end{bmatrix}.$$

2. We assume data come from an $LLGM$ and we compute a critical value through a GF-distribution at a significant level of 0.05 conditionally to $\Omega = \widehat{\Omega}$. The resulting critical value is 0.971.
3. Compute $P\left(T > 0.971 | \Omega = 0\right) = 0.478$ through the F-distribution. This estimated probability is taken as a degree of uncertainty associated with the choice of an $LLGM$. See Fig.: 2 (b).

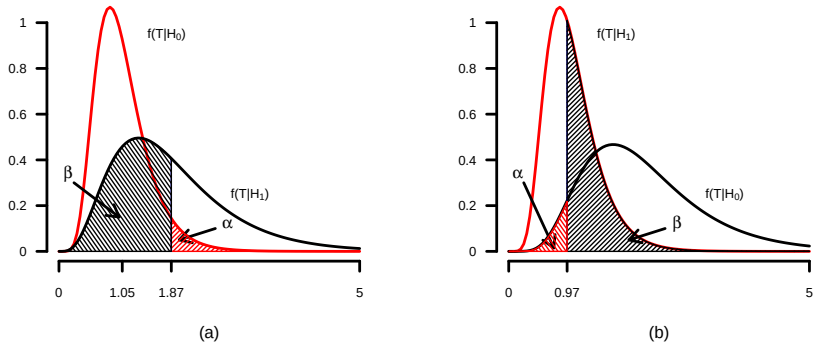

Figure 2: Hypothesis testing with statistic T on Cadralazine data. $f(T|H_0)$ is the density of T when H_0 is true; $f(T|H_1)$ is the density of T when H_1 is true; $\alpha = 0.05$ is the probability of a Type I error; β is the probability of a Type II error; the numbers 1.87 and 0.97 are critical values.

8 Conclusions

We propose a finite sample parametric test to discriminate between a linear regression model and a linear latent growth model. The test statistic is based on the trace of the product of an estimate of a variance covariance matrix defined when data come from a linear regression model and a sample variance covariance matrix based on ordinary least squares estimators. The sampling distribution of the test statistic depends on the model generating the data and can have a "standard" F-distribution or a linear combination of F-distributions. In this paper a unifying sampling distribution based on an infinite series representation of generalized F-distributions is given. This result allows us to frame the choice of the model in a classical hypothesis testing approach. By appropriately modifying the test statistic it is also possible to test hypotheses on randomness of single elements of the linear latent growth model, thus avoiding the boundary problem of the likelihood ratio statistic. The test statistic proposed in this paper has been applied to two data sets. With the Tourism data it is used by itself to discriminate between the two models, with the Cadralazine data it is used in conjunction with several indicators based on information criteria that give an estimate of the probability of accepting or rejecting the model chosen.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (eds. H. Akaike, B. N. Petrov and F. Csaki), 267–281. Akadèmiai Kiadò.

Barnabani, M. (2014) Some proposals of Theil used to discriminate between a linear latent growth model and a linear regression model. *Far East Journal of Theoretical Statistics*, **47(1)**, 19–40.

— (2015) An approximation to the convolution of gamma distributions. *Communications in Statistics, Simulation and Computation*, **doi:10.1080/03610918.2014.963612**.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014) lme4: Linear mixed-effects models using eigen and s4. *ArXiv e-print; submitted to Journal of Statistical Software*. URL`http://arxiv.org/abs/1406.5823`.

Crainiceanu, C. and Ruppert, D. (2004) Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc. B*, **66 (1)**, 165–185.

Dubey, S. D. (1970) Compound Gamma, Beta and F distributions. *Metrika*, **16**, 27–31.

Giampaoli, V. and Singer, J. (2009) Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, **139**, 1435–1448.

Gumpertz, M. and Pantula, S. (1989) A simple approach to inference in random coefficient model. *The American Statistician*, **43, No. 4**, 203–210.

Jain, G. C. and Consul, P. (1971) A generalized negative binomial distribution. *SIAM J. Appl. Math.*, **21, No. 4**, 501–513.

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008) Fence methods for mixed model selection. *Ann. Statist.*, **36**, 1669–1692.

Kourouklis, S. and Moschopoulos, P. G. (1985) On the distribution of the trace of a noncentral-Wishart matrix. *Metron*, **XLIII - N. 1-2**, 85–92.

Laird, N. M. and Ware, J. K. (1982) Random effect models for longitudinal data. *Biometrics*, **38**, 963–974.

Malik, H. (1967) The exact distribution of the quotient of independent generalized gamma variables. *Canadian Mathematical Bulletin*, **Vol. 10**, 463–465.

Moschopoulos, P. G. (1985) The distribution of the sum of independent gamma random variables. *Ann. Inst. Statist. Math.*, **37, Part A**, 541–544.

Muller, S., Scealy, J. L. and Welsh, A. H. (2013) Model selection in linear mixed models. *Statistical Science*, **28, No. 2**, 135–167.

Nagao, H. (1973) On some test criteria for covariance matrix. *Annals of Statistics*, **1**, 700–709.

Pillai, K. C. S. (1955) Some new test criteria in multivariate analysis. *Ann. Mathem. Stat.*, **No. 1, 26**, 117–121.

Pillai, K. C. S. and Jayachandran, K. (1968) Power comparison of tests of equality of two covariance matrices based on four criteria. *Biometrika*, **55**, 335–342.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL`http://www.R-project.org/`.

Roy, S. (1953) On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, **24**, 220–238.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Self, S. G. and Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators

and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Silvapulle, J. M. and Sen, P, K. (2005) *Constrained Statistical Inference: Inequality, order and shape restrictions*. Hoboken, New Jersey: John Wiley.

Stram, D. and Lee, J. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50, No. 4**, 1171–1177.

Swamy, P. (1970) Efficient Inference in a Random Coefficient Regression Model. *Econometrica*, **38**, 311–323.

Theil, H. (1963) On the Use of Incomplete Prior Information in Regression Analysis. *Journal of America Statistical Association*, **Vol. 58**, 401–414.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of Roy. Statist. Soc. Ser. B*, **58**, 267–288.

Vaida, F. and Blanchard, S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika*, **92 (2)**, 351–370.

Vellaisamy, P. and Upadhye, N. S. (2009) On the sums of compound negative binomial and gamma random variables. *Journal of Applied Probability*, **46**, 272–283.

Verbeke, G. and Molenberghs, G. (2003) The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.