UNIVERSITÀ
DEGLI STUDI
FIRENZE

DISIA
DIPARTIMENTO DI STATISTICA,
INFORMATICA, APPLICAZIONI
“GIUSEPPE PARENTI”

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

Marco Barnabani

DISIA WORKING PAPER
2015/04

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 choosing between a linear regression model and a linear latent growth model. Approaches based both on information criteria and on asymptotic hypothesis test on the variances of "random" components are largely used but not completely satisfactory. In the paper we propose a finite sample parametric test based on the trace of the product of estimates of two variance covariance matrices, one defined when data come from a linear regression model, the other defined when data come from a linear latent growth model. The sampling distribution of the test statistic so defined depends on the model generating data. It can be a "standard" F-distribution or a linear combination of F-distributions. In the paper a unified sampling distribution based on a generalized F-distribution is proposed. The knowledge of this distribution allows us to make inference in a classical hypothesis testing framework. The test statistic can be used by itself to discriminate between the two models and/or, duly modified, it can be used to test randomness on single components of the linear latent growth model avoinding the boundary problem of the likelihood ratio test statistic. Moreover, it can be used in conjunction with some indicators based on information criteria giving estimates of probability of accepting or rejecting the model chosen.

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

1 Introduction

It is common practice in many applications to collect multiple measurements on subjects across time focusing interest 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 the 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. To address the issue of which model is more suitable, one might use standard model selection measures of 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 choosing models that minimize an estimate of a specific criterion that 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 the above approaches extracting information from some published simulation results. As known one of the major drawback of these approaches is that they do not give any measure on the degree of uncertainty of the model chosen. The value they produces 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, there is difficulty with this that makes the usual approach of comparing the likelihood ratio test statistic to the chi-square distribution inappropriate. Asking whether the variance of a component is zero corresponds to whether this variance takes its value on the boundary of the parameter space. This situation is known as ”non-standard” relative to the other uses of the likelihood ratio test. The major consequence is that in large sample -2 times the logarithm of the likelihood ratio cannot be treated as a chi-square distribution but instead 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 to the critical value from a chi-square sampling distribution tend to not reject the null as often as it should. Other test not based on the likelihood function can be implemented (Silvapulle and Sen, 2005) but their validity is to be detected carefully when applied to linear mixed models. Moreover, all these tests are valid only asymptotically. Finite sample distributions of the likelihood ratio test require simulation and are known only for 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, 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 has a distinguished history in multivariate statistics. In most cases it is used the likelihood ratio approach and the resulting test statistics involve the ratio of the determinant of the sample covariance matrix under the null hypothesis and under 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 to discriminate between a LRM and a $LLGM$ defining appropriately the two matrices involved.

Let's denote with V the variance covariance matrix of the ordinary least square estimators when data come from a LRM and let $V+\Omega$ be the variance covariance matrix of the same estimators when data come from a $LLGM$ where Ω denotes the covariance matrix of the random effects. The Pillai's type test statistic proposed in the paper is based on an estimate of $\frac{1}{k}tr\ V^{-1}(V+\Omega)$ with Ω that has a crucial role to discriminate 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 a $LLGM$. If $\Omega = 0$, $\frac{1}{k}tr\ V^{-1}(V+\Omega) = 1$ and data come from a LRM. In section 2 after introducing some notation, the test statistic is defined. In section 3 we analyze the sampling distribution. When data come from a LRM it has a ”standard” F-distribution, when data come from a $LLGM$ the sampling distribution is more complex. It is a linear combination of standard F-distributions whose exact form is studied. Following the work of Kourouklis and Moschopoulos (1985) a unified sampling distribution involving a generalized F-distribution is proposed. This distribution is based on a 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. This data set is sufficiently ”regular” to allow a clear-cut answer on the choice of the model. The answers produced by the test are not conflicting with those given by AIC's indicators. The advantage coming from a 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's indicators applied to the data set does not give clear-cut indications about the model. There is a substantial indeterminacy which remains also using the test proposed in the paper but still again we can give some more information computing an estimate of the probability to accept the ”wrong” model.

2 Notation and test statistic

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, u_i is a $t \times 1$ vector whose component is the measurement error at a time point for individual i.

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) $v_i \sim N\left(0, \Omega\right)$, 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 linear latent growth model.

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

By replacing the random component into 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 $S_b = (n-1)^{-1}\sum_{i=1}^{n}\left(b_i - \bar{b}\right)\left(b_i - \bar{b}\right)'$ be the sample variance covariance matrix of b_i with $\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$. When data come from a $LLGM$, S_b is an unbiased consistent estimate of $\sigma^2(X'X)^{-1} + \Omega$ (Gumpertz and Pantula, 1989) when data come from a LRM S_b is an unbiased consistent estimate of $\sigma^2(X'X)^{-1}$.

To discriminate between a LRM or a $LLGM$ 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 that the test statistic T is approximately equal to one. When data come from a $LLGM$ we expect that $T > 1$. The greater T the stronger is the evidence against a 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 a 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). In this case $(n-1)(X'X)^{1/2}\frac{S_b}{\sigma^2}(X'X)^{1/2} \sim W_k\,(I, n-1)$ 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 $tr\left((X'X)^{1/2}\frac{S_b}{\sigma^2}(X'X)^{1/2}\right) = tr\,(X'X)\,\frac{S_b}{\sigma^2}$.
By 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 have the sampling distribution of T

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

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

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

(ii) As to the distribution of T, observe 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

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

We summarize the results in Table 1.

Table 1: Summary Table

Data come from:	LRM	$LLGM$
$T \sim$	$F_{k(n-1),n(t-k)}$	$\frac{1}{k}\sum_{i=1}^{k}\left(1 + \frac{\eta_i}{\sigma^2}\right)F_{(n-1),n(t-k)}$

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 to compute 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

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

The standard F-distribution (3) can be seen as a GF-distribution with $\delta = n(t-k)/(n-1)$, $\alpha = (n-1)/2, \gamma = n(t-k)/2$.

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 (6) is a linear combination of independent chi-square variates whose distribution does not admit a closed and simple form. However, the gamma-series representation proposed by Kourouklis and Moschopoulos (1985) and Moschopoulos (1985) is particularly useful for our purposes. 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, by (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 $\delta = \frac{n(t-k)}{k(n-1)}\,\eta$. We summarize the results in Table 2.

The series representation of GF-distribution is not complex to implement in practice and in most statistical softwares there is a function that compute the generalized F-distribution. In this paper computations are made with R (R Core Team, 2014) where a library (GB2) (or flexsurv)

Table 2: Summary Table

Data come from	LRM	$LLGM$
$T \sim$	$GF\,(\delta, \alpha, \gamma)$ $\delta = \frac{n(t-k)}{k(n-1)}$ $\alpha = \frac{k(n-1)}{2}$ $\gamma = \frac{n(t-k)}{2}$	$\sum_{l=0}^{\infty} w_l\ GF\,(\delta, \alpha, \gamma)$ $\delta = \frac{n(t-k)}{k(n-1)}\,\eta$ $\alpha = \rho + l$ $\gamma = \frac{n(t-k)}{2}$

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 result too much 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 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 $(l, w_l)_{0,1,2,\ldots}$ must be truncated after a desired accuracy.

4 Inference on the model

By table (2) we can see that the sampling distribution of T depends on η_i, the eigenvalues of the matrix $(X'X)^{1/2}\,\Omega\,(X'X)^{1/2}$. ”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 $\widehat{\Omega} = S_b - s^2(X'X)^{-1}$. $\widehat{\Omega}$ is the difference of two matrices and may yield negative estimates for variances of some of the coefficients and/or could not be a positive definite matrix. In this case we could have negative eigenvalues. Although negative $\widehat{\eta}_i$ could appear to be misleading this definition of $\widehat{\Omega}$ is coherent with above sampling distributions. Actually, observe that $E(T) = \frac{n(t-k)}{n(t-k)-2}\overline{\eta}$ where $\overline{\eta} = (1/k)\sum_{i=1}^{k}(1+\eta_i/\sigma^2)$. $\widehat{\Omega}$ so defined allows us to show that the test statistic T defined in (1) is equal to $(1/k)\sum_{i=1}^{k}(1 + \widehat{\eta}_i/s^2)$. Therefore, T

can be seen as an estimate of $\overline{\eta}$. Moreover, $\overline{\eta} = 1 \Leftrightarrow \Omega = 0$ that is, if and only if data come from a LRM. In this case the estimator T has a GF-distribution (F-distribution). $\overline{\eta} > 1 \Leftrightarrow \Omega \succeq 0$. In this case data come from a $LLGM$ and the distribution of 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. Therefore, the models describing T are different for the two sources of data. Under $LLGM$ the model contains the other as a special case. More specifically, constraining the parameter $\overline{\eta}$ to one we have the model under LRM. We call alternative hypothesis the more general model and null hypothesis the restricted model.

We tackle the hypothesis testing problem defining the null hypothesis $H_0 : \overline{\eta} \leq 1$ against the alternative $H_1 : \overline{\eta} > 1$ taking T as estimator of $\overline{\eta}$. The comparison of the two hypotheses can be reduced to a $p - value$, that is, the probability of seeing T as large as (as small as) we did, or even larger (smaller), when, in fact, H_0 is adequate. When the $p - value$ is small (close to zero) we "reject H_0 in favor of H_1". On the other hand, when the $p - value$ is not small we "fail to reject H_0", there is a non-negligible probability that T could reasonably be the result of random chance and presumably data come from LRM.

The comparison between H_0 and H_1 can also be conducted following a classical decision approach. Under H_0 we can compute a critical value and then rejecting or accepting the null Hypothesis if the observed statistic, T, is greater or less than the critical value. 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. Unfortunately, this knowledge is not available and only an estimate of the probability is possible replacing σ^2 with s^2 and η_i with $\widehat{\eta}_i$. When n is large the estimates of the probabilities are accurate.

5 Inference on single component

If T is greater than a critical value or the $p - value$ is small likely data come from a $LLGM$. In this case it can be useful to investigate which component is random. Table 1 and Table 2 show the role of the parameter $\left(1 + \eta_i/\sigma^2\right)$ in defining the sampling distributions when data come from a $LLGM$ 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 us to pick out the number of random components but not which of them are random. To this matter 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 $(X'X)^{-1}$. This "new" parameter, $\phi_i = \left(1 + \frac{\omega_{ii}}{\sigma^2 x^{ii}}\right)$, can be seen as expressing the extent of "total" variability of $i - th$ coefficient $(\sigma^2 x^{ii} + \omega_{ii})$ in relation to the "residual" variance $\sigma^2 x^{ii}$. The reciprocal of this parameter, $\phi_i^{-1} = \frac{\sigma^2 x^{ii}}{\sigma^2 x^{ii} + \omega_{ii}}$, can be seen as the share of "residual" variance on "total" variability. It ranges between zero and one. If $\omega_{ii} > 0$ then $\phi_i^{-1} < 1$ 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 the 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.

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 of the $i-th$ component. The greater the value of ϕ_i the stronger is this evidence. When $\omega_{ii}=0$ the parameter ϕ_i is equal to one and the $i-th$ component is zero variance. Given $\omega_{ii}>0$ and increasing σ^2, ϕ_i tends towards one.

A "natural" estimator of ϕ_i is $\widehat{\phi}_i = 1+\frac{\widehat{\omega}_{ii}}{s^2x^{ii}}$ where $\widehat{\omega}_{ii}$ is the entry (i,i) of the matrix $\widehat{\Omega}$.

The sampling distribution of $\widehat{\phi}_i$ is immediate. Because $(n-1)S_b/\sigma^2 \sim W_k\left((X'X)^{-1}+\Omega/\sigma^2, n-1\right)$, $(n-1)\widehat{s}_{ii}/\sigma^2 \sim \left(x^{ii}+\frac{\omega_{ii}}{\sigma^2}\right)\chi^2_{n-1}$ where $\widehat{s}_{ii}$ is the (i,i) entry of the matrix S_b. This implies that

$$\frac{\widehat{s}_{ii}}{\sigma^2x^{ii}} \sim \left(1+\frac{\omega_{ii}}{\sigma^2x^{ii}}\right)\frac{\chi^2_{n-1}}{n-1}$$

replacing σ^2 with s^2 we get

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

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

Because of the definition of $\widehat{\Omega}$, simple algebra allows to show that $\widehat{\phi}_i = 1+\frac{\widehat{\omega}_{ii}}{s^2x^{ii}} = \frac{\widehat{s}_{ii}}{s^2x^{ii}}$.

When data come from a LRM, $\omega_{ii}=0$ and $\phi_i=1$. We call $H_0:\phi_i=1$ the null hypothesis. In this case the estimate $\widehat{\omega}_{ii}$ can assume values grater or less than zero and $\widehat{\phi}_i$ ranges 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 $P\left(\widehat{\omega}_{ii}\leq 0\right)$ can be computed with the $F-$distribution. If data come from a $LLGM$, $\omega_{ii}>0$ 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 grater or less than zero but the negative values are becoming less and less frequent the stronger is the evidence against the null hypothesis, that is, the higher is ϕ_i. The test statistic $\widehat{\phi}_i$ assumes values greater than one and if it is greater than a critical value computed with a $F-$distribution, we reject the null hypothesis (in favor of a $LLGM$). Of course a $p-value$ can also be computed.

A "confounding" situation can appear when the "residual" variance σ^2x^{ii} is large compared with the elements of Ω. In this case ϕ_i is close to one and the test statistic $\widehat{\phi}_i$ has a GF-distribution close to an $F-$distribution. This situation is well known in a classical statistical hypothesis testing and there is a large probability to fail to reject the null hypothesis in favor of the alternative.

By (12), $\frac{\widehat{s}_{ii}}{\phi_i s^2x^{ii}}$ is a pivotal quantity which allows to construct a confidence interval for ϕ_i. Fixing α we can determine two percentiles of F-distribution such that

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

Thus, if data come from a $i-th$ 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 compute automatically the confidence interval for each component we could face two situations: (a) an interval contained in $(0,1)$. In this case presumably the component is random, (b) an interval around one. In this case a substantial indeterminacy occurs. We could have a zero variance component or a random component but σ^2 dominates the variance of the component confounding the choice.

6 An application: Tourism data

A data set on Tourism in Tuscany (Italy) consist of the Index number (base year 2002) of accommodations (the response variable) on 260 Municipalities from 2003 to 2009. These data have been firstly processed so that to obtain homogeneous groups of units. In the paper we work with 98 ”homogeneous” Municipalities, see the left panel of Fig.: 1. Looking at the tourism data each unit appears to have its own trajectory approximated by linear functions with specific intercept and slope that determine the trend. Moreover, the trajectories are ”high” or ”low” suggesting two hypotheses from an economic point of view. One is that the growth of the tourism of 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 it.

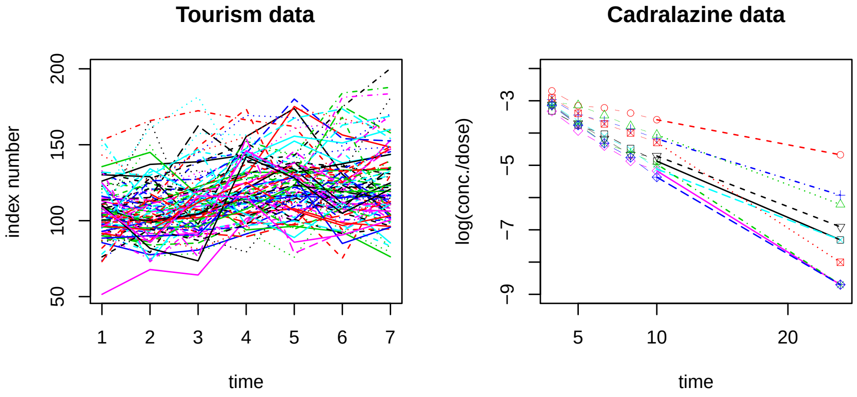

Figure 1: Spaghetti plots for Tourism data and Cadralazine data

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 of tourism influenced not solely by the regional political economy but also by specific characteristics of each municipality. This suggests that data could be modeled adding a random component to the parameter vector so that to distinguish the various trajectories. On the basis of our data we ask whether the specificity of the municipalities contributes to the growth of the tourism other than the regional political economy. Statistically we ask whether it is more appropriate modeling data with a linear regression model or a linear latent growth model.

By applying the hypothesis testing approach proposed in the paper we found:

- A value of the test statistic $T = 4.76$ which compared with the critical value $F_{194,490,0.95} = 1.212$ falls into the rejection region. We reject the hypothesis that data come from a LRM with a probability of Type II error close to zero.
- We observe a $p - value \simeq 0$. Confirming a strong evidence against the null hypothesis.
- $\widehat{\phi}_1 = 3.245$ and $\widehat{\phi}_2 = 3.7313$ compared with $F_{97,490,0.95} = 1.279$ indicate that both components are random.

- The confidence intervals of the shares: $0.21719 \leq \phi_1^{-1} \leq 0.40331$ and $0.19351 \leq \phi_2^{-1} \leq 0.3593$ confirm that data come from a $LLGM$.

The above results are compared with some usual indicators used in model selection. These indicators are computed with R (R Core Team, 2014) and the package $lme4$ (Bates *et al.*, 2014). The results are shown in table 3

	AIC	BIC	$cAIC$
LRM	6122.482	6136.075	
$LLGM$	5925.687	5952.872	5776.097

Table 3: Comparison of AIC's for the linear regression model and linear latent growth model for Tourism data.

where $cAIC$ is the conditional AIC proposed by Vaida and Blanchard (2005).

All the above indicators confirm the choice of a linear latent growth model to describe data.

7 An application: Cadralazine data

In the previous section we discussed a data set which allowed to give clear and evident information on the choice of the model. To illustrate some difficulties we could face to discriminate 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 an anti-hypertensive drug, the cadralazine. 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 intercepts and slopes fixed, and a mixed effects model where intercepts and slopes are considered random.

The choice between the two models is firstly conducted through AIC's type indicators. From table 4 we can see that there is a substantial indeterminacy by comparing AIC and BIC indicators. They produce conflicting results, AIC addresses us to choose a $LLGM$, the BIC value gives a different interpretation reversing the choice. Moreover, how to evaluate the differences of the values produced? While no rigorous theory is available, Burnham and Anderson (2002),

	AIC	BIC	$cAIC$
LRM	161.717	168.0	
$LLGM$	157.923	170.5	143.016

Table 4: Comparison of AIC's for the linear regression model and linear latent growth model for Cadralazine data.

suggest that a difference of at most 2 in AIC is not reliable for ranking two models, whereas a difference of 10 is overwhelmingly in favor of the model with the smaller AIC. Of course the values 2 and 10 don't mean anything.

The conditional AIC defined only for linear mixed models shows a value inferior to the others in favor of a $LLGM$. However it is not comparable with the other AIC's indicators and the value it produces does not mean anything by itself. Moreover, given this substantial indeterminacy choosing the model, what is the degree of uncertainty to accept a $LLGM$ instead of a LRM?

The indeterminacy emerging in this example is not removed with the test proposed in this paper but it can give some further information useful to accompany the choice the model:

- We found a value of the test statistic $T = 1.7829$ that compared to $F_{18,40,0.95} = 1.8682$ falls into the acceptance region. Then, we fail to reject a LRM. The closeness of the observed value to the critical value suggests us a certain caution about the choice of the model. Actually, we found a $p - value = 0.0639$ that confirms our caution. These results reflect the indeterminacy of AIC and BIC indicators.
- The probability of Type II error is important to quantify the uncertainty about the model chosen. Its computation requires the knowledge of 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 allow to estimate the GF-distribution under the alternative hypothesis. The result is the 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, if BIC indicator suggests the choice of a LRM we adjoin that there is an estimated large probability to accept it on the basis of information contained in the data set. See Fig.: 2 (a).
- The AIC and in particular the $cAIC$ indicator addresses the choice towards a $LLGM$. What can we say about the probability to accept this model when it is "wrong"? We proceed as follows:
  1. Estimate the variance covariance matrix with the package $lme4$ of R taking this estimate as a hypothesis on Ω, $\widehat{\Omega} = \begin{bmatrix} 0.00054686 & 0.003727 \\ 0.003727 & 0.025400 \end{bmatrix}$.
  2. Conditionally to $\Omega = \widehat{\Omega}$ we assume data come from a $LLGM$. We compute a critical value through a GF-distribution at a significant level of 0.05. The 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 to the choice of a $LLGM$. See Fig.: 2 (b).

8 Conclusions

In the paper 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 estimates of two variance covariance matrices, one defined when data come from a linear regression model, the other defined when data come from a linear latent growth model. The sampling distribution of the test statistic depends on the model generating data and can have a "standard"

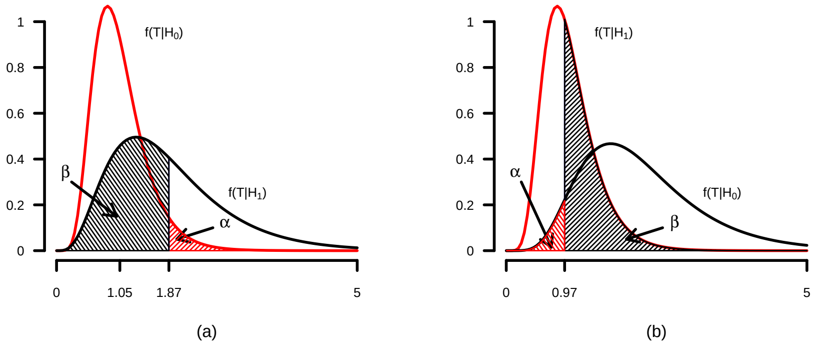

Figure 2: Hypothesis testing with statistic T on Cadralazine data. $f(T|H_0)$ is the density of T when H_0 is true; $f(T|H_1)$ is the density of T when H_1 is true; $\alpha = 0.05$ is the probability of Type I error; β is the probability of Type II error; the numbers 1.87 and 0.97 are critical values.

F-distribution or a linear combination of F-distributions. In the paper a unifying sampling distribution based on GF-distribution has been proposed. This result allows us to frame the choice of the model in a classical hypothesis testing approach. By modifying appropriately the test statistic it is also possible to test hypotheses on randomness of single elements of the linear latent growth model avoinding the boundary problem of the likelihood ratio statistic. The test statistic proposed in the paper has been applied to two data set. With Tourism data it is used by itself to discriminate between the two models, with Cadralazine data it is used in conjunction with some indicators based on information criteria giving an estimate of the probability of accepting or rejecting the model chosen.

References

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

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 Theorethical Statistics*, 47(1), 19–40.

Barnabani M. (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*.

Burnham K. and Anderson D. (2002) *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*, 2nd ed. New York: Springer.

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.

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 condition, *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.