serie ad

WP-AD 2010-18

Things that make us different: analysis of variance in the use of time

Jorge González Chapela

Ivie

Working papers

Working papers

ng papers

Edita / Published by: Instituto Valenciano de Investigaciones Económicas, S.A.

Depósito Legal / Legal Deposit no.: V-1699-2010

Impreso en España (abril 2010) / Printed in Spain (April 2010)

WP-AD 2010-18

Things that make us different: analysis of variance in the use of time[*]

Jorge González Chapela[**]

Abstract

The bounded character of time-use data poses a challenge to the analysis of variance based on classical linear models. This paper investigates a computationally simple variance decomposition technique suitable for these data. As a by-product of the analysis, a measure of fit for systems of time-demand equations that possesses several useful properties is proposed.

Keywords: Time allocation, multivariate regression, deviance.

JEL Classification: C52, J22.

[*] A previous version of this paper, circulated as "Things that make us different: Analyzing individual-level heterogeneity in the use of time," benefited from the comments of Lola Collado, Daniel Hamermesh, Ángel López, Andrés Romeu, and seminar participants at Universidad de Alicante and Universidad Politécnica de Cartagena. I am also grateful to participants in the 31st IATUR meeting at Lueneburg, particularly to Jay Stewart, and to a referee. Financial support from the Instituto Valenciano de Investigaciones Económicas (Ivie) and the Spanish Ministry of Education (ECO2008-05721/ECON) is gratefully acknowledged.

[**] Jorge González Chapela: University of Alicante. E-mail: jorge@merlin.fae.ua.es.

1 Motivation

The analysis of variance (ANOVA) is a collection of statistical techniques utilized in a wide variety of disciplines for different, although interrelated, purposes. As commonly applied, the decomposition of variance is based on a linear regression model which, although generally appropriate for data with Normal errors, is considered inadequate when the dependent variable is bounded. McCullagh and Nelder (1989, p. 35), for example, point out that the linear model inadequacy is reflected in that ANOVA sums of squares are no longer appropriate measures of the contribution of a factor to the total discrepancy observed in the data.

The bounded nature of time-use data, which represent proportions of a given total time, poses therefore a challenge to classical ANOVA. Nevertheless, ANOVA has been the technique generally utilized for screening the effects of explanatory factors over the allocation of time,[1] perhaps due to the technical and computational simplicity allowed by the linear model. In this paper, we investigate a computationally simple variance decomposition technique suitable for time-use data. The many uses of ANOVA, including exploratory data analysis, testing means across groups of observations, and testing nested sequences of models, and the increasing availability and usage of time-use data,[2] calls for an adequate variance decomposition technique for this kind of data.

We start off in Section 2 by reviewing the literature on specification and estimation of systems of time-demand equations. Because of its robustness to distributional failure and computational simplicity, we advocate the multinomial logit specification and quasi-

[1] See, e.g., Gershuny (2000, Ch. 6) and Freeman and Schettkat (2005).

[2] An up-to-date description of data and developments in time-use analysis can be found at the University of Oxford's Centre for Time Use Research, http://www.timeuse.org/.

likelihood estimation method proposed in Mullahy and Robert (2008) as an attractive statistical approach for time-use data. Then, in Section 3, we set out the statistical theory needed for performing an analysis of variance on a sample of time-use data based on Mullahy and Robert's statistical framework. As a generalized measure of discrepancy between a sample of time-use data and a set of fitted values derived from a model we use the concept of deviance. Our main result, which is a particular case of a more general theory presented for example in McCullagh and Nelder (1989, Ch. 9), is that if we are willing to specify the mean and variance function of a sample of time-use data as those of a one-trial multinomial distribution, this distribution could be then utilized to assess the deviance as if it had truly generated the data. Thus, the scaled value of the resulting quasi-likelihood would allow us to assess the contribution of a factor to the total deviance observed in the data.

To our knowledge, there is no goodness-of-fit measure for systems of time-demand equations. In Section 4, and as a by-product of the previous analysis, we propose an R^2 measure for systems of time-demand equations that possesses several important properties. Our R^2 is an extension of Hauser's (1978) pseudo-R^2 for multinomial regression models that can be computed using quasi-likelihood statistics, instead of maximum likelihood statistics as in Hauser's original formulation. The new measure may be for example interpreted in a similar way to R^2 in the linear regression context, namely as the fraction of variation of the dependent variable accounted for by the explanatory variables.

Section 5 illustrates the previous methods on a sample of time-use data taken from the Spanish Time Use Survey. The results of sequential and partial analysis of deviance reveal that employment status is the major contributor to deviance in the allocation of time between noon and 1 pm on a representative weekday, whereas the proposed R^2 suggests that the logarithm of age fits better the data than age itself. Section 6 offers some conclusions.

2 Specification and Estimation of a System of Time-Demand Equations

In multivariate analyses of the allocation of time, the total time analyzed (T) is generally classified into M mutually exclusive and exhaustive-of-T activities. Then, letting t_m denote the amount of time devoted to activity m and $\mathbf{x} \equiv (1, x_2, \ldots, x_K)$ represent a $1 \times K$ vector of explanatory variables, a linear form for the regression of t_m on $\mathbf{x}$ is commonly specified,

$$E(t_m | \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_m, \quad m = 1, \ldots, M, \tag{1}$$

where $\boldsymbol{\beta}_m$ is a $K \times 1$ vector of parameters. As it seems natural for an analysis of variance exercise, we assume that the same set of explanatory variables shows up in each of these M linear regressions, though this assumption is not necessary and could be relaxed.

In such a framework, the equation-by-equation Ordinary Least Squares (OLS) is a logical estimator for the parameters of (1), for it is easy to apply, it has good econometric properties, and it automatically takes into account the natural requirement that predicted times must add up to T. Nevertheless, the OLS estimator can not preserve that predicted times lie in the interval $[0, T]$, and, given the bounded character of the dependent variables, the linear model implication of constant partial effects can not be literally true unless the range of $\mathbf{x}$ is severely restricted. In practice, both drawbacks seem to be counterbalanced by the technical and computational simplicity allowed by the linear model.

For systems of equations in which the components of the multivariate dependent variable are non-negative, may take on certain values with positive probability, and add up to a constant, Wales and Woodland (1983) developed two alternative econometric models estimated by the maximum likelihood principle. Although both approaches yield parameter estimates with good econometric properties, we think that the technical and computational complexities involved in Wales and Woodland's approaches make them more appealing for final data analysis[3] than for exploratory decomposition of variance exercises.

In an article published in 1996, Papke and Wooldridge developed an attractive specification as well as a simple quasi-likelihood estimation method for a regression model of a dependent variable bounded between 0 and 1. More recently, Mullahy and Robert (2008) have generalized the Papke and Wooldridge approach to the context of a complete system of time-demand equations where the total time analyzed is normalized to 1. The population regression considered in Mullahy and Robert (2008) is of the multinomial logit form,

$$E\left(y_m|\mathbf{x}\right)=\frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)}{\sum_{k=1}^{M}\exp\left(\mathbf{x}\boldsymbol{\beta}_k\right)},\quad m=1,\ldots,M, \tag{2}$$

where $y_m=t_m/T,\ m=1,\ldots,M$. This nonlinear specification ensures for example that the predicted value of y_m lies in the interval $\left(0,1\right)$, that the sum from 1 to M of the predicted y_ms adds up to 1, and that the partial effect of x_j on $E\left(y_m|\mathbf{x}\right)$ is not constant, but depending on $\mathbf{x}$. Another important feature is that equation (2) is well defined even if every y_m can take on 0 or 1 with positive probability. The normalization $\boldsymbol{\beta}_1=\mathbf{0}$ is generally imposed for identification purposes.

A particular quasi-likelihood method is advocated in Mullahy and Robert (2008) to estimate the parameters of (2). The multinomial logit log-likelihood function

$$l\left(\mathbf{b}\right)\equiv\sum_{m=1}^{M}y_m\left(\mathbf{x}\mathbf{b}_m-\ln\left(\sum_{k=1}^{M}\exp\left(\mathbf{x}\mathbf{b}_k\right)\right)\right), \tag{3}$$

where $\mathbf{b}\equiv\left(\mathbf{0}',\mathbf{b}_2',\ldots,\mathbf{b}_M'\right)'$ is a generic element of the parameter space, is an objective function of the class associated with linear exponential probability distributions. Given the availability

[3] Performed, for example, in Dong, Gould, and Kaiser (2004) and Prowse (2009).

of a sample of N independent observations $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, 2, \ldots, N\}$, where $\mathbf{y}_i \equiv (y_{i1}, \ldots, y_{iM})'$, the quasi-maximum likelihood estimator (QMLE) of $\boldsymbol{\beta} \equiv (\mathbf{0}', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_M')'$ obtained from the maximization problem

$$\max_{\mathbf{b}} \sum\nolimits_{i=1}^{N} l_i(\mathbf{b}), \tag{4}$$

is consistent for $\boldsymbol{\beta}$ and asymptotically normal provided that equation (2) holds.[4] In other words, although the conditional-on-$\mathbf{x}$ probability distribution of the random vector $\mathbf{y}$ is not multinomial, if the conditional mean is correctly specified the fact that the assumed probability distribution is linear exponential makes the QMLE $\hat{\boldsymbol{\beta}}$ to have satisfying econometric properties regardless of the true distribution of $\mathbf{y}$ given $\mathbf{x}$. Although technically more complex than the equation-by-equation OLS estimator, this QMLE is not much more difficult to compute, for, as Mullahy and Robert (2008) point out, it can be implemented using minor modifications of ordinary multinomial logit estimation algorithms.

The asymptotic covariance matrix of the multinomial logit QMLE shares the general shape of the QMLE variance matrix given for example in Gourieroux et al. (1984). In the particular case that

$$V(\mathbf{y}_i | \mathbf{x}_i) = \sigma^2 \mathbf{V}_i, \tag{5}$$

where σ^2 denotes a dispersion parameter, $\mathbf{V}_i$ represents a variance function with *mk*th element $p_{im}(\delta_{imk} - p_{ik})$,

$$p_{im} \equiv \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_m)}{\sum_{k=1}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_k)}, \tag{6}$$

[4] A general exposition of the properties of quasi-maximum likelihood estimators is provided in Gourieroux et al. (1984), for example.

and δ_{imk} is an indicator variable equal to 1 if $m=k$ and equal to 0 if $m\neq k$, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ could be simply estimated as

$$\hat{\sigma}^2\hat{\mathbf{A}}^{-1}. \tag{7}$$

Following McCullagh and Nelder (1989),

$$\hat{\sigma}^2=\left(NM-N-MK\right)^{-1}\sum_{im}\left(y_{im}-\hat{p}_{im}\right)^2\Big/\hat{p}_{im}\left(1-\hat{p}_{im}\right), \tag{8}$$

while

$$\hat{\mathbf{A}}\equiv-\sum\nolimits_{i=1}^{N}\left(\hat{\mathbf{V}}_i\otimes\mathbf{x}_i'\mathbf{x}_i\right), \tag{9}$$

the symbol $\otimes$ denoting the Kronecker product.

3 Variance Decomposition in a System of Time-Demand Equations

The analysis of variance is a useful statistical method for screening the effects of explanatory factors and their interactions on a possibly multidimensional dependent variable. Its fundamental technique is a partitioning of the total sum of squares of a dependent variable into a component related to the factors included in the model and a residual component:

$$\sum_i\left(y_i-\bar{y}\right)^2=\sum_i\left(y_i-\hat{y}_i\right)^2+\sum_i\left(\hat{y}_i-\bar{y}\right)^2, \tag{10}$$

where y_i denotes a data value, $\hat{y}_i$ is a value predicted by the model being fitted, and $\bar{y}$ is a sample average. As it is well known, the partitioning in (10) holds in the ordinary regression analysis of an unbounded response variable, but it does not generally apply when the response variable is bounded and a nonlinear regression model for $E\left(y|\mathbf{x}\right)$ has to be estimated. Nevertheless, the literature on generalized linear models (see for example the monograph by McCullagh and Nelder, 1989) has generalized the analysis of variance to certain non-linear contexts based on the concept of deviance. In the remainder of this section, we present the concept of deviance in connection with the Kullback-Leibler distance function, and we show how the deviance can be used to assess the discrepancy to data of a non-linear model estimated by maximum likelihood. Then, we show how to compute the deviance in a sample of time-use data when there is insufficient information to construct a likelihood function. The concepts presented next are established results that are re-stated here for readability.

Let $f_{\mathbf{y}_i}$ and $f_{\mathbf{p}}$ denote two absolutely continuous probability distributions associated to the $M \times 1$ random vector $\mathbf{y}$ that differ only in terms of their means: $f_{\mathbf{y}_i}$ is centred at a realization of $\mathbf{y}$ (denoted $\mathbf{y}_i$), whereas $f_{\mathbf{p}}$ is centred at $E[\mathbf{y}] = \mathbf{p}$. A standard measure of discrepancy between $f_{\mathbf{y}_i}$ and $f_{\mathbf{p}}$ is the Kullback-Leibler (KL) divergence,

$$K(\mathbf{y}_i, \mathbf{p}) \equiv 2E_{\mathbf{y}_i} \ln\left(f_{\mathbf{y}_i}(\mathbf{y}) / f_{\mathbf{p}}(\mathbf{y}) \right), \qquad (11)$$

where $E_{\mathbf{y}_i}$ refers to expectation with respect to $f_{\mathbf{y}_i}$ and the factor 2 has been added for convenience. The KL divergence averages a measure of discrepancy between the two probability distributions over their support, giving more weight in this average to values of higher probability, as determined by $f_{\mathbf{y}_i}$. Thus, although the argument is not defined when $f_{\mathbf{y}_i}(\mathbf{y}) = f_{\mathbf{p}}(\mathbf{y}) = 0$, such values of $\mathbf{y}$ have no effect when computing the average. The fact that $K(\mathbf{y}_i, \mathbf{p}) \geq 0$,[5] with equality if and only if $f_{\mathbf{y}_i} \equiv f_{\mathbf{p}}$, may lead us to think of $K(\mathbf{y}_i, \mathbf{p})$ as representing a distance between $\mathbf{y}_i$ and $\mathbf{p}$. But since K is asymmetric in its arguments, the term deviation or divergence is generally preferred.

Efron (1978) showed that when f belongs to the linear exponential family (LEF) of probability distributions, the expectation in (11) drops out and $K(\mathbf{y}_i, \mathbf{p})$ is simply given by

$$K(\mathbf{y}_i, \mathbf{p}) \equiv 2\ln\left(f_{\mathbf{y}_i}(\mathbf{y}_i) / f_{\mathbf{p}}(\mathbf{y}_i) \right). \qquad (12)$$

[5] $K(\mathbf{y}_i, \mathbf{p}) = \infty$ is allowed, which occurs when $f_{\mathbf{y}_i} > 0$ and $f_{\mathbf{p}} = 0$ for some $\mathbf{y}$.

Expression (12), which measures the discrepancy between the probability distributions $f_{\mathbf{y}_i}$ and $f_{\mathbf{p}}$ just at the point $\mathbf{y}_i$, is called the deviance (or component of deviance). The deviance is generally regarded a goodness of fit measure, for if a parametric model for $\mathbf{p}$ were specified, the deviance would quantify the prediction error in using $\mathbf{p}$ to predict $\mathbf{y}_i$.[6] As an example, if f represented a one-parameter Gaussian density with $\sigma^2 = 1$, the deviance would be equivalent to $\left(y_i - p\right)^2$, a loss criterion whose minimization is the basis of several estimators, including OLS.

Given a sample of N independent observations and a particular data generating process for the data, denoted $f_{\mathbf{p}}$, the estimated deviance between the observations $\mathbf{Y} = \left(\mathbf{y}_1, \ldots, \mathbf{y}_N\right)$ and the fitted values $\hat{\mathbf{P}} = \left(\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_N\right)$ would be given by

$$K\left(\mathbf{Y}, \hat{\mathbf{P}}\right) \equiv 2\sum\nolimits_{i=1}^{N}\left(\ln f_{\mathbf{y}_i}\left(\mathbf{y}_i\right) - \ln f_{\hat{\mathbf{p}}_i}\left(\mathbf{y}_i\right)\right). \qquad (13)$$

The term $\sum\nolimits_{i=1}^{N} \ln f_{\hat{\mathbf{p}}_i}\left(\mathbf{y}_i\right)$ is the estimated log likelihood based on $f_{\mathbf{p}}$. In the simplest case, $f_{\mathbf{p}}$ has an $M \times 1$ vector of parameters (the null model), giving rise to a common vector of fitted values for all the $\mathbf{y}_i\mathrm{s}$. The resulting estimated deviance might be called the data total deviance, $K\left(\mathbf{Y}, \hat{\mathbf{P}}_{\mathbf{0}}\right)$, where the sub-index $\mathbf{0}$ refers to fitted values obtained from the null model. As an example, if f represented a one-parameter Gaussian density with $\sigma^2 = 1$, and estimation were done by minimizing squared error loss, the total deviance would be the total sum of squares, $\sum\nolimits_{i=1}^{N}\left(y_i - \bar{y}\right)^2$. At the other extreme, $f_{\mathbf{p}}$ contains as many $M \times 1$ vectors of parameters as observations (the full model), and the fitted values derived from it match the data exactly, leaving no room for deviance. In this case, $f_{\mathbf{p}}$ achieves its maximum log-

[6] If a dispersion parameter, σ^2, further characterized f, the discrepancy of the model to data would be scaled by σ^2.

likelihood, denoted $\sum_{i=1}^{N} \ln f_{\mathbf{y}_i}(\mathbf{y}_i)$. The difference $K(\mathbf{Y}, \hat{\mathbf{P}}_{\mathbf{0}}) - K(\mathbf{Y}, \hat{\mathbf{P}})$ is a measure of the reduction in deviance achieved by the fitted model, i.e., due to the inclusion of explanatory variables.

The foregoing discussion assumed the process by which the data were generated was known, but often this process is unknown. We may, however, be able to estimate the deviance if we are willing to specify certain features of the data. Let the mean and variance function of $\mathbf{y}$ be those of a linear exponential probability distribution, denoted f in (15). Then, as shown for example in McCullagh and Nelder (1989, Ch. 9), the deviance between the observations $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ and the fitted values $\hat{\mathbf{P}} = (\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_N)$ can be computed as

$$-2Q(\hat{\mathbf{P}}; \mathbf{Y}), \tag{14}$$

where

$$Q(\hat{\mathbf{P}}; \mathbf{Y}) = \sum_{i=1}^{N} \left(\ln f_{\hat{\mathbf{p}}_i}(\mathbf{y}_i) - \ln f_{\mathbf{y}_i}(\mathbf{y}_i) \right) \tag{15}$$

is the (estimated) quasi-likelihood for $f_{\mathbf{p}}$ based on data $\mathbf{y}$, which generally differs from the log-likelihood due to the presence of the constant $\sum_{i=1}^{N} \ln f_{\mathbf{y}_i}(\mathbf{y}_i)$. Note that there is no support restrictions in calculating the quasi-likelihood, so the $\mathbf{y}_i$ need not belong to the support of f. In the case of a sample of time-use data, we may assume that the vector $\mathbf{y}$ has conditional mean $\mathbf{p}$ with mth element as given in (2) and covariance matrix $\mathbf{V}(\mathbf{y}|\mathbf{x})$ as specified in (9), which, except for the presence of σ^2, are the mean and variance of a one-trial multinomial distribution. The quasi-likelihood would be then given by

$$Q(\mathbf{P}; \mathbf{Y}) = \sum_{i=1}^{N} \left(\sum_{m=1}^{M} y_{im} \left(\mathbf{x}_i \mathbf{b}_m - \ln \left(\sum_{k=1}^{M} \exp(\mathbf{x}_i \mathbf{b}_k) \right) \right) - \sum_{m=1}^{M} y_{im} \ln y_{im} \right), \tag{16}$$

the data total deviance could be computed as $-2Q(\hat{\mathbf{P}}_{\mathbf{0}}; \mathbf{Y})$, and the reduction in deviance achieved by the inclusion of explanatory variables would be $-2\left(Q(\hat{\mathbf{P}}_{\mathbf{0}}; \mathbf{Y}) - Q(\hat{\mathbf{P}}; \mathbf{Y})\right)$.

An interesting property of LEF models that use the canonical link is that the KL divergence exhibits the Pythagorean property (see Hastie, 1987, pp. 19-20; Simon, 1973):

$$K\left(\mathbf{Y},\hat{\mathbf{P}}_{\mathbf{0}}\right)=K\left(\mathbf{Y},\hat{\mathbf{P}}\right)+K\left(\hat{\mathbf{P}},\hat{\mathbf{P}}_{\mathbf{0}}\right). \qquad (17)$$

In this case, the difference $K\left(\mathbf{Y},\hat{\mathbf{P}}_{\mathbf{0}}\right)-K\left(\mathbf{Y},\hat{\mathbf{P}}\right)$ can be interpreted not only as the reduction in deviance due to inclusion of explanatory variables, but also as the deviance explained by the regression model, $K\left(\hat{\mathbf{P}},\hat{\mathbf{P}}_{\mathbf{0}}\right)$. Since

$$\mathbf{x}\boldsymbol{\beta}_m=\ln\left[\frac{p_m}{p_1}\right],\ m=1,\ldots,M\ , \qquad (18)$$

the mean function specified in (2) corresponds to the canonical link of the multinomial distribution. Hence, if the mean and variance function of a sample of time-use data are those of a multinomial distribution, it turns out that

$$-2Q\left(\hat{\mathbf{P}}_{\mathbf{0}};\mathbf{Y}\right)=-2\left(Q\left(\hat{\mathbf{P}};\mathbf{Y}\right)+Q\left(\hat{\mathbf{P}}_{\mathbf{0}};\hat{\mathbf{P}}\right)\right), \qquad (19)$$

and the difference $-2\left(Q\left(\hat{\mathbf{P}}_{\mathbf{0}};\mathbf{Y}\right)-Q\left(\hat{\mathbf{P}};\mathbf{Y}\right)\right)$ would equal the deviance explained by the regression model.

4 An *R*-Squared Measure of Goodness of Fit for Systems of Time-Demand Equations

A commonly reported goodness-of-fit statistic in the standard linear regression model is the coefficient of determination, or R^2, which, among other possible interpretations, generally conveys the intuitive meaning of fraction of variation of the dependent variable explained by the explanatory variables. It is well known, however, that the direct application of this statistic to nonlinear contexts is troublesome, for it can lie outside the $[0,1]$ interval and decrease as explanatory variables are added. For this reason, alternative R^2-type goodness-of-fit statistics (generally called pseudo-R^2s) have been constructed for particular nonlinear models using a variety of methods. For multinomial regression models, Hauser (1978) proposed a pseudo-R^2 derived from information theory and calculated using maximum likelihood statistics which, among other satisfying properties, lies between 0 and 1 and is non-decreasing as explanatory variables are added. Later on, Cameron and Windmeijer (1997) proposed a pseudo-R^2 measure based on the KL divergence for exponential family regression models estimated by Maximum Likelihood, of which Hauser's goodness-of-fit statistic is a particular case.

In this section, we draw upon the exposition in Cameron and Windmeijer (1997) to extend Hauser's pseudo-R^2 measure to be computed using QML statistics. We also reinterpret Hauser's pseudo-R^2 in the light of the deviance measure of discrepancy defined above. A possible extension of Cameron and Windmeijer's pseudo-R^2 measure to be computed using QML statistics is left for future research.

Under the conditions that let $-2Q$ to be a measure of deviance, a measure of the proportionate reduction in total deviance achieved by the fitted regression model can be calculated as:

$$R_Q^2 = 1 - Q\left(\hat{\mathbf{P}};\mathbf{Y}\right)\Big/Q\left(\hat{\mathbf{P}}_{\mathbf{0}};\mathbf{Y}\right). \tag{20}$$

The R_Q^2 has the following properties:

1. R_Q^2 is non-decreasing as explanatory variables are added. Proof: The QMLE maximizes $Q(\mathbf{P};\mathbf{Y})$, which will therefore not decrease as explanatory variables are added, i.e., as constraints on the coefficients are removed.
2. $0 \le R_Q^2 \le 1$. Proof: The lower bound of 0 occurs if inclusion of explanatory variables leads to no change in the fitted values, and the upper bound occurs when the model fit is perfect.

3. R_Q^2 is a scalar multiple of the quasi-likelihood ratio (QLR) test statistic for the hypothesis that the coefficients of all the explanatory variables, save the constants, are 0. Proof: Re-expressing R_Q^2 as $\left(\sum_{i=1}^{N}\ln f_{\hat{\mathbf{p}}_0}(\mathbf{y}_i)-\sum_{i=1}^{N}\ln f_{\hat{\mathbf{p}}_i}(\mathbf{y}_i)\right)\Big/Q\left(\hat{\mathbf{P}}_0;\mathbf{Y}\right)$, where $\hat{\mathbf{p}}_0$ is the vector of fitted values derived from the null model, it turns out that $R_Q^2=\frac{\hat{\sigma}_u^2}{-2Q\left(\hat{\mathbf{P}}_0;\mathbf{Y}\right)}QLR$, where $\hat{\sigma}_u^2$ is a consistent estimate of σ^2 obtained from unrestricted estimation.

4. R_Q^2 can be equivalently expressed as

$$R_Q^2=\frac{Q\left(\hat{\mathbf{P}}_0;\hat{\mathbf{P}}\right)}{Q\left(\hat{\mathbf{P}}_0;\mathbf{Y}\right)}, \tag{21}$$

where $Q\left(\hat{\mathbf{P}}_0;\hat{\mathbf{P}}\right)$ is (up to the factor -2) the estimated deviance between the null model and the fitted model. Hence, R_Q^2 could be equivalently interpreted as the fraction of deviance explained by the fitted model. Proof: See the discussion surrounding expression (19).

Properties 1 and 2 are standard properties often desired for $R-\text{squared}$ measures. Property 3 generalizes a similar result for the linear regression model under normality (Anderson, 1958), and has the practical intent of avoiding conflicting signals between the ranking of models generated by R_Q^2 and the related statistical test. Property 4 is also desirable for it allows R_Q^2 to be interpreted similarly as the usual $R-\text{squared}$ in the linear regression model: either as the proportionate reduction in the deviance due to inclusion of explanatory variables, or as the fraction of deviance explained by the regression model.

5 Application

As an illustration of the previous methods, we perform an analysis of deviance on a sample of time-use data as well as a test of non-nested hypotheses using R_Q^2. Data are from the Spanish Time Use Survey (STUS), which between October 1, 2002, and September 30, 2003, interviewed a representative sample of the non-institutionalized population. Among other information, all sample members aged 10 years old or older were requested to list their activities in every 10-minutes slot of a particular 24-hours cycle.[7]

Our sample contains 2,341 persons, who were selected following these criteria: persons must be 10 years old or older, they must live in Galicia—a north-western region of Spain, the time diary must pertain to a weekday, and the potential explanatory variables (detailed below) must code valid and non-missing answers. Following the suggestions of Ås (1978), these persons' allocation of time to main activities was classified into four aggregate time uses: necessary time (made up mainly of eating, sleeping, and cleansing), contracted time (working in the market, searching for job, studying), committed time (including housework and caring for children), and free time (volunteering, exercising, etc.) Time travelling was classified according to the declared purpose of the trip.

[7] The STUS sample is a two-stage sample. At a first stage, selection was from a list of *secciones censales*, which are clusters of housing units generally comprising between 500 and 2000 inhabitants. The sample of *secciones censales* was uniformly distributed along the 52 weeks of the sample period, with half of the housing units in each *sección censal* being assigned a weekday (Monday to Thursday) and half a weekend day. The housing units themselves, in which all residents aged 10 years old or older were asked to fill in a time diary, were selected at the second stage. The STUS is designed to produce reliable estimates at the region and country level.

The variables whose inclusion in $\mathbf{x}$ is to be assessed are employment status, sex, child status, household income, health status, completed schooling, and trimester of interview. Each of these factors is represented by a set of dummy variables whose cardinality is generally determined by the number of answer alternatives in the corresponding survey question. In two cases, however, the answers were simplified: employment status represents only two outcomes: employed (working full- or part-time or receiving education) and non-employed (rest of situations); child status, by which we mean the number and age of co-resident children, has three possible values: being part of a family with no children, being part of a family whose youngest child is 5 years or younger, or being part of a family whose youngest child is between 6 and 17 years. Of course, many other factors could be relevant for explaining the allocation of time on weekdays in Galicia, but in order to simplify the exposition, only these seven factors are considered. (Below, in a non-nested model selection exercise, we shall discuss the introduction of age as an additional explanatory variable.) For table layout simplicity, possible overlaps among these factors are ignored.

Although STUS respondents do provide information on the allocation of time for a complete 24-hours cycle, we focus on the allocation of time just during one hour of the assigned diary day, specifically between noon and 1 pm. We do this to take into account the intraday variability in the timing of activities, which, if the day were selected as the unit of analysis, would be obscured (Winston, 1982). Indeed, the seven potential explanatory factors listed above are not equally significant for explaining the allocation of time in each of the 24 parts that the day is divided into. In our sample, the hour between noon and 1 pm is the part in which these factors are capable of explaining the largest fraction of deviance, the lowest fraction explained being between 10 pm and 11 pm.

Tables 1 and 2 record, respectively, the results of sequential and partial analyses of deviance on our time-use data. The sequential analysis of deviance table illustrates a method of organizing a series of model comparisons tests which is relevant when explanatory factors are added one at a time, the order generally based on judgment and/or convention. In this case, the order, which is given in the *Source* column of Table 1, is based on our own presumptions about causal priority. The sequential deviance for a factor is the gain in prediction from a model including that factor plus those preceding it in column 1 of Table 1, over a model including the preceding factors only. Thus, for example, the sequential deviance for *Child status* is the gain in prediction from a model including *Employment status*, *Sex,* and *Child Status* over a model including *Employment status* and *Sex* only. By construction, the total deviance explained by all seven factors, called hereafter the model deviance, equals the sum of their sequential deviances.

On the other hand, the partial analysis of deviance table examines the contribution of each factor over and above the joint contribution of the remaining factors. Hence, partial deviances are calculated as the model deviance minus the deviance in the sub-model in which only the factor of interest is eliminated. Thus, for example, the partial deviance for *Child status* is the gain in prediction from a model including all seven factors over a model excluding *Child status*. The sum of partial deviances does not generally add up to the model deviance because the explanatory terms tend to be correlated.

The columns of both tables also list the degrees of freedom (df), the values of the quasi-likelihood ratio statistic (QLR) for testing the statistical significance of each explanatory factor or of the overall model, and the *p-values* associated to the values of this statistic. The QLR statistic is computed based on the difference in the quasi-likelihood function with and without the restrictions imposed, that is,

$$QLR = \frac{-2\left(\sum_{i=1}^{N} \ln f_{\hat{\mathbf{p}}_r}\left(\mathbf{y}_i\right) - \sum_{i=1}^{N} \ln f_{\hat{\mathbf{p}}_u}\left(\mathbf{y}_i\right)\right)}{\hat{\sigma}_u^2}, \tag{22}$$

where $\hat{\sigma}_u^2$ is calculated as indicated in expression (8) using results from the unrestricted estimation (see Wooldridge, 2002, p. 370). The QLR statistic has a χ_2 limiting distribution under H_0, with degrees of freedom given by the number of restrictions being tested. Since each variable may have associated non-zero coefficients in three out of the four uses of time being explained, the degrees of freedom are given by three times the number of dummy variables representing the factor(s) whose statistical significance is evaluated.

In our sample of 2,341 observations, the total deviance amounts to 5,183.6. The model is able to explain 1,807.6 when the seven explanatory factors are jointly included in $\mathbf{x}$. Hence, the value of R_Q^2 amounts to 0.3487, clearly conveying the message that this model provides a good fit for these data. Employment status is the major contributor to deviance in the allocation of time between noon and 1 pm on a representative weekday in Galicia: its partial deviance represents about 16 per cent of the total deviance observed in the data, whereas, by itself, it represents about 28 per cent of the total deviance. Both analyses reveal also sizeable sex, education, and health status effects, whereas they show modest effects associated to child status, household income, and trimester of interview. As shown in Table 2, sex, education, or health status are adding to the ability of the model to predict the allocation of time even when all other six factors are included in $\mathbf{x}$. On the other hand, child status is not a significant predictor of the allocation of time in a model containing employment status and sex (Table 1), or when the model includes all other six factors (Table 2). Likewise, household income is not a significant predictor in a model with employment status, sex, and child status (Table 1), or when all other six factors are included in $\mathbf{x}$ (Table 2). The figures for trimester of interview in both tables do coincide because the contribution to deviance of this factor in Table 1 is evaluated once all other six factors are included in $\mathbf{x}$, the conclusion being that trimester of interview does not serve as a significant predictor of the allocation of time.

Table 1. Sequential analysis of deviance in time use

	No. of observations: 2,341				$R_Q^2 = .3487$
Source	Sequential deviance	df	$\hat{\sigma}^2$	QLR	Prob > QLR
Model deviance	1,807.6	81	1.106	1,634	.0000
Employment status	1,451.7	3	1.115	1,302	.0000
Sex	219.7	3	1.100	199.7	.0000
Child status	5.34	6	1.100	4.85	.5631
Household income	21.50	21	1.108	19.41	.5588
Health	38.71	12	1.090	35.51	.0004
Education	58.52	27	1.095	53.44	.0018
Trimester of interview	12.15	9	1.106	10.99	.2766
Residual deviance	3,376.0	2,257			
Total deviance	5,183.6	2,338			

Notes: Author's calculations based on a sample from the STUS. QLR: quasi-likelihood ratio statistic, calculated as $\textit{sequential deviance}/\hat{\sigma}^2$. $\hat{\sigma}^2$ is computed as shown in expression (8) using results from the corresponding unrestricted model.

Table 2. Partial analysis of deviance in time use

	No. of observations: 2,341			$R_Q^2 = .3487$
Source	Partial deviance	df	QLR	Prob > QLR
Model deviance	1,807.6	81	1,634	.0000
Employment status	847.8	3	766.5	.0000
Sex	217.4	3	196.6	.0000
Child status	9.15	6	8.27	.2189
Household income	15.95	21	14.42	.8508
Health	31.03	12	28.06	.0054
Education	59.16	27	53.49	.0017
Trimester of interview	12.15	9	10.99	.2766
Residual deviance	3,376.0	2,257		
Total deviance	5,183.6	2,338		

Notes: Author's calculations based on a sample from the STUS. QLR: quasi-likelihood ratio statistic, calculated as $\textit{partial deviance}/\hat{\sigma}^2$. $\hat{\sigma}^2$, computed as shown in expression (8) using results from the model with all seven factors included, equals 1.106.

Besides quantifying the proportion of total deviance explained by the fitted model, the R_Q^2 proposed in this paper may also be useful to select among alternative non-nested models, provided that the models contain the same number of parameters. Suppose for example that we decided to add each person's age to the set of seven explanatory factors considered so far, but we wonder whether it is better (in predictive ability terms) to add age in levels or its natural logarithm. It turns out that both age and its natural logarithm are statistically significant predictors for the allocation of time, but when age is added to $\mathbf{x}$, R_Q^2 increases to 0.3510, whereas R_Q^2 equals to 0.3514 when it is its natural logarithm the variable included in $\mathbf{x}$. Although for a small margin, the logarithm function of age seems to fit better the data.

6. Conclusion

The multinomial logit specification and quasi-likelihood estimation method proposed in Mullahy and Robert (2008) make up an attractive statistical approach for time-use data. Additionally, if we are willing to specify the variance function of the data as that of a one-trial multinomial distribution, a variance decomposition exercise can be performed from ordinary quasi-likelihood statistics. From this output it is also possible to construct a measure of fit statistic that has the right interpretation at the limits of the unit interval, as well as an intuitively appealing interpretation between these limits. An empirical application to Spanish time-use data illustrates the usefulness of these methods: employment status is the major contributor to deviance in the allocation of time between noon and 1 pm on a representative weekday, whereas the logarithm of age seems to fit better the data than age itself.

References

Anderson, T.W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York. Wiley.

Ås, D. 1978. Studies of Time-Use: Problems and Prospects. *Acta Sociologica* 21(2):125-141.

Cameron, A.C. and F.A.G. Windmeijer. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics* 77:329-342.

Dong, D., B.W. Gould, and H.M. Kaiser. 2004. Food demand in Mexico: an application of the Amemiya-Tobin approach to the estimation of a censored food system. *American Journal of Agricultural Economics* 86(4): 1094-1107.

Efron, B. 1978. The geometry of exponential families. The *Annals of Statistics* 6(2):362-376.

Freeman, R.B. and R. Schettkat. 2005. Marketization of household production and the EU-US gap in work. *Economic Policy* 20(41):5-50.

Gershuny, Jonathan. 2000. *Changing times. Work and leisure in postindustrial society*. New York, Oxford University Press.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo Maximum Likelihood methods: theory. *Econometrica* 52(3): 681-700.

Hastie, T. 1987. A closer look at the deviance. *The American Statistician* 41(1):16-20.

Hauser, J.R. 1978. Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. *Operations Research* 26(3):406-421.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. Second edition. Boca Raton (Florida), Chapman & Hall/CRC.

Mullahy, J. and S.A. Robert. 2008. No time to lose? Time constraints and physical activity. NBER Working Paper 14513. Cambridge (MA), National Bureau of Economic Research.

Papke, L.E. and J.M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401 (K) plan participation rates. *Journal of Applied Econometrics* 11(6): 619-632.

Prowse, V. 2009. Estimating labour supply elasticities under rationing: a structural model of time allocation behaviour. *Canadian Journal of Economics* 42(1): 90-112.

Simon, G. 1973. Additivity of information in exponential family probability laws. *Journal of the American Statistical Association* 68(342):478-482.

Wales, T.J. and A.D. Woodland. 1983. Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics* 21: 263-285.

Winston, G.C. 1982. *The timing of economic activities – Firms, households, and markets in time-specific analysis*. Cambridge (UK), New York, and Melbourne. Cambridge University Press.

Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge (Massachusetts) and London (England). The MIT Press.

PUBLISHED ISSUES

WP-AD 2010-01	“Scaling methods for categorical self-assessed health measures” P. Cubí-Mollá. January 2010.
WP-AD 2010-02	“Strong ties in a small world” M.J. van der Leij, S. Goyal. January 2010.
WP-AD 2010-03	“Timing of protectionism” A. Gómez-Galvarriato, C.L. Guerrero-Luchtenberg. January 2010.
WP-AD 2010-04	“Some game-theoretic grounds for meeting people half-way” P.Gadea-Blanco, J.M. Jiménez-Gómez, M.C. Marco-Gil. February 2010.
WP-AD 2010-05	“Sequential city growth: empirical evidence” A. Cuberes. February 2010.
WP-AD 2010-06	“Preferences, comparative advantage, and compensating wage differentials for job routinization”. C. Quintana-Domeque. February 2010.
WP-AD 2010-07	“The diffusion of Internet: a cross-country analysis” L. Andrés, D. Cuberes, M.A. Diouf, T. Serebrisky. February 2010.
WP-AD 2010-08	“How endogenous is money? Evidence from a new microeconomic estimate” C. Cuberes, W.R. Dougan. February 2010.
WP-AD 2010-09	“Trade liberalization in vertically related markets” R. Moner-Colonques, J.J. Sempere-Monerris, A. Urbano. February 2010.
WP-AD 2010-10	“Tax evasion as a global game (TEGG) in the laboratory” M. Sánchez-Villalba. February 2010.
WP-AD 2010-11	“The effects of the tax system on education decisions and welfare” L.A. Viianto. March 2010.
WP-AD 2010-12	“The pecuniary and non-pecuniary costs of job displacement. The risky job of getting back to work” R. Leombruni, T. Razzolini, F. Serti. March 2010.
WP-AD 2010-13	“Self-interest and justice principles” I. Rodríguez-Lara, L. Moreno-Garrido. March 2010.

[*] Please contact Ivie's Publications Department to obtain a list of publications previous to 2010.

WP-AD 2010-14 “On spatial equilibria in a social interaction model”
P. Mossay, P.M. Picard. March 2010.

WP-AD 2010-15 “Noncooperative justifications for old bankruptcy rules”
J.M. Jiménez-Gómez. March 2010.

WP-AD 2010-16 “Anthropometry and socioeconomics in the couple: evidence from the PSID”
S. Oreffice, C. Quintana-Domeque. April 2010.

WP-AD 2010-17 “Differentiated social interactions in the US schooling race gap”
L.J. Hall. April 2010.

WP-AD 2010-18 “Things that make us different: analysis of variance in the use of time”
J. González Chapela. April 2010.

serie ad

Ivie

Guardia Civil, 22 - Esc. 2, 1º
46020 Valencia - Spain
Phone: +34 963 190 050
Fax: +34 963 190 055

Department of Economics
University of Alicante

Campus San Vicente del Raspeig
03071 Alicante - Spain
Phone: +34 965 903 563
Fax: +34 965 903 898

Website: http://www.ivie.es
E-mail: publicaciones@ivie.es