

REVISTA ESTADISTICA

Vol. 63 - NRO. 180 – junio 2011

CONTENIDO – CONTENTS

THE CONCEPT OF VARIABILITY IN TIME SERIES ANALYSIS

ESTELA BEE DAGUM
University of Bologna, Italy
estela.beedagum@unibo.it

ABSTRACT

The concept of variability is at the essence of the development of statistics pragmatic epistemology. In the particular case of time series to be discussed in this study, one basic assumption is that variability due to time can be present in the signal and/or the noise of the data generating process.

In recent years, the variability of time series models for series observed daily, such as those in finance, biostatistics, weather, and so on, has shown non-linear dependence of the level on previous data points. When dealing with nonlinearities, one should make the distinction between: (1) Linear time series where shocks are assumed to be uncorrelated but not necessarily identically independent distributed (iid), and (2) Nonlinear time series where shocks are assumed to be iid, but there is a nonlinear function relating the observed time series and the underlying shocks.

Among the most applied non-linear time series models are those representing changes of variance along time (heteroskedasticity). These models are called autoregressive conditional heteroskedasticity (ARCH) and the collection comprises a wide variety of representations (GARCH, TARCH, EGARCH, FIGARCH, CGARCH, etc). Here changes in variability are predicted by recent past values of the observed series. This is in contrast to other possible representations of locally varying variability, where the variability might be modelled as being driven by a separate time-varying process, as in a doubly stochastic model.

The main goal of this study is to discuss the impact of time variability in the data generating process of a given phenomenon. This is done by introducing different assumptions concerning the impact of time.

Key words

Time series models of deterministic variability, time series models of linear and nonlinear stochastic variability.

RESUMEN

El concepto de variabilidad está en la esencia del desarrollo epistemológico pragmático de la Estadística. En el caso particular de las series temporales, sobre las que trata este estudio, un supuesto básico es que la variabilidad debida al tiempo puede estar presente en la señal y/o en el ruido del proceso generador de datos.

En años recientes, la variabilidad de los modelos de series de tiempo para series diarias, tales como las de finanzas, bioestadística, clima y otras, ha demostrado una dependencia no-lineal del nivel de los datos respecto a valores previos. Cuando se trata de procesos no-lineales se debe distinguir entre: (1) Series de tiempo lineales donde los shocks se suponen no-correlacionados pero no necesariamente independientes idénticamente distribuidos (iid) , y (2) Series de tiempo no lineales donde los shocks se suponen iid, pero existe una función no lineal que relaciona la serie de tiempo observada y los shocks subyacentes.

Entre los modelos de series de tiempo más utilizados están aquellos que representan cambios en la varianza a través del tiempo (heteroscedasticidad). Estos modelos se llaman heteroscedasticidad condicional autoregresiva (ARCH) y el conjunto comprende una amplia variedad de representaciones (GARCH, TARCH, EGARCH, FIGARCH, CGARCH, etc). En estos, los cambios en la variabilidad se predicen por los valores del pasado reciente de la serie observada. Esto contrasta con otras posibles representaciones de la variabilidad local, donde la misma debería ser modelada utilizando otro proceso que varíe en el tiempo, como en un modelo doblemente estocástico.

El principal objetivo de este estudio es discutir el impacto de la variabilidad temporal en el proceso generador de los datos de un fenómeno dado. Se plantearán diferentes supuestos respecto al impacto del tiempo.

Palabras clave

Modelos temporales de variabilidad determinística, modelos temporales de variabilidad estocástica lineal y no lineal.

PRODUCING ICT INDICATORS IN DEVELOPING COUNTRIES: CHALLENGES AND INITIATIVES

SUSAN TELTSCHER
International Telecommunication Union (ITU)
Susan.Teltscher@itu.int

JOSÉ CERVERA-FERRI
DevStat
jcervera@devstat.com

ABSTRACT

The follow-up to the World Summit on the Information Society (WSIS) requires, as any other international agreement, a monitoring and evaluation system consisting of a set of agreed indicators. A core list of Information and Communication Technology (ICT) indicators was proposed by the international community and endorsed by the UN Statistical Commission, including 48 indicators on ICT infrastructure, ICT access and use by households and businesses, the ICT producing sector, international trade in ICT goods and use of ICT in education. For National Statistical Systems, the production of statistics for the calculation of these indicators represents a challenge, with organisational, technical and resource implications. OECD member economies have produced survey-based ICT indicators since the late 1990s, but the availability of ICT data is still low in developing countries. Global initiatives have been put in place for building capacity for the production and dissemination of ICT statistics, such as the activities of the *Partnership on Measuring ICT for Development*. This paper discusses first the main institutional and technical challenges that developing economies face in relation to the production of ICT indicators. In the second part, the paper makes suggestions on how these challenges could be addressed presenting relevant national and international activities undertaken to enhance the availability of comparable ICT statistics.

Keywords

ICT indicators, developing countries, statistical capacity building

RESUMEN

El seguimiento de la Cumbre Mundial sobre la Sociedad de la Información (CMSI) exige, al igual que otros acuerdos internacionales, un sistema de monitoreo y evaluación consistente en un conjunto aceptado de indicadores. En este sentido, una lista de indicadores fundamentales sobre Tecnología de Información y Comunicación (TIC) fue propuesta por la comunidad internacional y validada por la Comisión de Estadística de las Naciones Unidas, comprendiendo 48 indicadores sobre la infraestructura de TIC, el acceso a las TIC por hogares e individuos, el sector de producción de TIC, el comercio internacional de bienes TIC y el uso de TIC en la educación. La producción de estadísticas para el cálculo de estos indicadores representa un desafío para los Sistemas Estadísticos Nacionales, con implicaciones organizativas, técnicas y de recursos. Las economías de la OCDE llevan produciendo indicadores de TIC basados en encuestas desde el final de los años 90, pero la disponibilidad de datos sobre TIC es todavía escasa en los países en desarrollo. Algunas iniciativas globales han sido lanzadas para reforzar la capacidad para la producción y la difusión de estadísticas de TIC, tales como la Asociación para la Medida de las TIC para el Desarrollo. Este artículo presenta en la primera parte los principales retos institucionales y técnicos que deben afrontar las economías en desarrollo en

relación con la producción de indicadores sobre TIC. En la segunda parte, el artículo ofrece recomendaciones sobre cómo afrontar dichos desafíos, presentando algunas actividades nacionales e internacionales pertinentes que han sido llevadas a cabo para mejorar la disponibilidad de estadísticas comparables sobre las TIC.

Palabras clave

Indicadores TIC, países en desarrollo, construcción de capacidad estadística

**COMPONENTES PRINCIPALES A PARTIR DE DATOS
CON INFORMACIÓN FALTANTE.
EFICIENCIA DE ALGORITMOS ALTERNATIVOS**

MARTA BEATRIZ QUAGLINO

*Escuela de Estadística, Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario, Argentina
mquaglino@fcecon.unr.edu.ar*

MARÍA SUSANA VITELLESCHI

*Escuela de Estadística, Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario, Argentina
mvitelle@fcecon.unr.edu.ar*

RESUMEN

El Análisis de Componentes Principales (PCA) es uno de los métodos estadísticos más utilizados para el análisis de datos multivariados, situación en la que es frecuente que aparezca información faltante. En este trabajo se consideran dos algoritmos propuestos para construir modelos PCA en estas situaciones, que utilizan toda la información disponible: NIPALS y EM. Se comparan por simulación, diferentes aspectos del modelo PCA que podrían afectarse frente a la pérdida de información: bondad de ajuste y de predicción, ortogonalidad de la matriz de “scores”, reproducción de la variabilidad total, variancia explicada y coeficientes de las primeras componentes; abarcando una variedad de situaciones que combinan distintas dimensiones de la matriz de datos, mecanismos, porcentajes y localización de pérdidas. En estos escenarios se encuentra que con ambos métodos se afectan escasamente la variabilidad explicada y la estructura de las primeras componentes. Además, el algoritmo EM es, en general, más eficiente si las matrices tienen más individuos que variables, mientras que NIPALS es superior en cuanto a medidas predictivas cuando hay más variables que individuos.

Palabras Claves

Modelos PCA, EM, NIPALS, información faltante

ABSTRACT

Principal Component Analysis (PCA) is a very popular technique for analyzing multivariate data, situation in which missing information is frequent. Two algorithms are considered for building PCA models which use all the available information: NIPALS and EM. Different aspects of the PCA models that could be affected when there is missing information are

compared by simulation: the goodness of fit and prediction, the orthogonality of the score matrix, the reproduction of the total variability, the explained variance and coefficients of the first principal components, covering a variety of situations that combine different dimensions of the data matrix and mechanisms, percentages and location of missing information. In these situations, both methods were found to slightly affect the explained variability and the structure of the first components. Besides, the EM algorithm is generally more efficient if matrixes have more individuals than variables, while NIPALS is superior in predictive measures when there are more variables than individuals.

Key Words

PCA models, EM, NIPALS, missing information

**SELF-ASSESSED HEALTH AS A HIDDEN DETERMINANT
OF SMOKING STATUS. EVIDENCE FROM ARGENTINA**

MARIANA CONTE GRAND

Department of Economics, Universidad del CEMA, Buenos Aires, Argentina

VANESA V. D'ELIA

Department of Economics, Universidad del CEMA, Buenos Aires, Argentina

vvd04@cema.edu.ar

ABSTRACT

The aim of this study is to analyze self-assessed health importance in smoking status, accounting for unobservable variables that jointly determine lifestyles and for health endogeneity. Specifically, health behaviors (smoking, diet, exercise, alcohol) are functions of socio-economic variables and self-assessed health (SAH). But, SAH varies with socio-economic indicators and depends on health indicators that derive from past habits (overweight, blood pressure, diabetes, cholesterol levels). Our database is adults in Argentina's 2005 Risk Factors National Survey. We find that more education and marriage decreases the chance of smoking, while adverse socio-economic indicators increase it. We confirm links between tobacco and other risk factors. But, more importantly, evaluating own health as good or better increases the probability of being a smoker by 10 percentage points. The impact of SAH is underestimated when other lifestyles are not considered, but mostly when SAH is taken as an exogenous variable.

Keywords

Smoking status, self-assessed health, lifestyles, multivariate probit, Argentina

RESUMEN

El objetivo de este estudio es analizar la importancia de la auto percepción sobre la propia salud en el consumo de tabaco, considerando variables no observables que determinan conjuntamente los estilos de vida y la endogeneidad de la salud. En concreto, las conductas de salud (tabaco,

dieta, ejercicio, alcohol) son funciones de las variables socioeconómicas y de la salud auto percibida (SAP). Sin embargo, la SAP varía con los indicadores socioeconómicos y depende de los indicadores de salud que se derivan de hábitos pasados (exceso de peso, presión arterial, diabetes, niveles de colesterol). La base de datos son los adultos de la Encuesta Nacional de Factores de Riesgos de Argentina para el año 2005. Encontramos que una mayor educación y estar casados reduce la probabilidad de fumar, mientras que los indicadores socioeconómicos adversos aumentan dicha probabilidad. Confirmamos los vínculos entre el tabaco y otros factores de riesgo. Pero, más importante aún, la auto evaluación de la salud como buena o mejor aumenta la probabilidad de ser fumador en 10 puntos porcentuales. El impacto de la SAP se subestima cuando otros estilos de vida no son considerados, pero sobre todo cuando la SAP se toma como una variable exógena.

Palabras clave

Consumo de tabaco, salud auto percibida, estilos de vida, probit multivariado, Argentina