

Contract Number BRG5380004

Project Title: Statistical Methods of data Analysis in Biology Environment and Medicine:
Theory and Applications II

Principal Investigator: Assoc. Prof. Dr. Montip Tiensuwan

Address: Department of Mathematics Faculty of Science, Mahidol University,
Rama 6 Road, Bangkok, Thailand 10400

Telephone:02-2015538, 02-2015340, **Fax:** 02-2015343, **e-mail:** montip.tie@mahidol.ac.th

Abstract

Data analysis is an integral part of statistics, and statistical methods to analyze data arising from various sources abound in the literature. Nevertheless, the need for new methods of data analysis or application of conventional methods to new area often arises because of emerging new and difficult data collection methods and innovative statistical problems. In the context of biology environment and medicine, we focused on nutritional assessment based on prices, and eleven nutrient indicators, namely energy, carbohydrates, fats, proteins, calcium, iron, vitamin A, thiamin, riboflavin, vitamin C, and niacin from groups of vegetables, foods, fruits, and desserts. Nutrient indicators and prices are assessed by the Multiple Criteria Decision Making (MCDM) and Electre methods. The overall index average is ranked from the highest to the lowest based on these indicators. Hence, the MCDM and Electre methods are shown to be feasible measures to compare and rank the types of vegetables, foods, fruits, and desserts based on nutrient indicators and prices within each category. A time series analysis is adopted as a technique for forecasting of monthly rainfall and water level along the Chao Phraya River in Thailand. The Box-Jenkins technique is used for identifying the parameters of an Autoregressive Integrated Moving Average (ARIMA) model. The Akaike information criterion, the Schwarz's Bayesian criterion and the mean square error are used throughout to test for simplification of any particular model. The periodogram analysis is used to confirm the existence of a seasonal period in the ARIMA model. The ARIMA with seasonal model possibly predicts the monthly rainfall and water level along the Chao Phraya River one year ahead with acceptable accuracy. The nonlinear mixed effects model is also proposed for repeated measures data to describe longitudinal changes in Carcinoembryonic Antigen (CEA) levels of colorectal cancer patients over time. The CEA level of colorectal cancer patients is regarded as the marker choice for monitoring. Parameters of the proposed model are estimated by

using Lindstrom and Bates (LB) and the Stochastic Approximation version of the standard Expectation and Maximization step (SAEM) algorithms. The results show that the estimates of the recurrent time (ϕ_2) by using LB and SAEM are equal to 17.4 and 22.8 months, respectively. Further, the residual sum of squares of the proposed model by using the LB and SAEM algorithms are also compared. Dose response models are the mathematical functions that relate the dose to the measure of observed effect. Three models are considered, namely, the multistage Weibull, logistic and log-logistic models. The study focuses on estimating parameters and comparing estimates in these three models based on the method of maximum likelihood and Berkson's minimum chi-square method. The two methods are also applied to four real experimental data sets for each of the three models. Lastly, we focused on apply the concept of sufficiency and Fisher information for the comparison of experiments of different bivariate normal populations with arbitrarily specified means and variances and with different forms of the correlation parameter ρ and to apply the concept of sufficiency for the simple linear regression model $y_x = \alpha + \beta x + e_x, e_x \sim N(0, \sigma^2), x \in [-1, 1]$ non-stochastic regression and for the one-way random normal model.

Keywords: Akaike's information criteria, asymptotic comparison, Berkson minimum chi-square, bivariate/trivariate normal populations, Box-Jenkins models, carcinoembryonic antigen, electre method, Fisher information, fixed effects, information, linear models, logistic model, log-logistic model, maximum likelihood, most uniform allocation, multistage Weibull model, multiple criteria decision making method, nonlinear mixed effects model, nutrient indicators, one-way random effects model, optimal designs, prices, random effects, repeated measures, Schwartz's Bayesian criterion, sufficient experiments, sufficiency, variance of white noise

Contract Number BRG5380004

ชื่อโครงการ: ระเบียบวิธีเชิงสถิติสำหรับการวิเคราะห์ข้อมูลทางชีววิทยา สิ่งแวดล้อมและการแพทย์: ทฤษฎีและการประยุกต์ II

หัวหน้าโครงการ: รองศาสตราจารย์ ดร. มนต์ทิพย์ เทียนสุวรรณ

ที่อยู่: ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล ถนนพระรามที่ 6 กรุงเทพมหานคร 10400

โทรศัพท์: 02-2015538, 02-2015340, **Fax:** 02-2015343, **e-mail:** montip.tie@mahidol.ac.th

บทคัดย่อ

การวิเคราะห์ข้อมูลคือภาคจำนวนเต็มของสถิติและวิธีเชิงสถิติในการวิเคราะห์ข้อมูลที่เกิดขึ้นจากแหล่งแตกต่างกันมีมากมายในเอกสารข้อมูลหรือวรรณกรรม อย่างไรก็ตามความต้องการวิธีการใหม่ๆสำหรับการวิเคราะห์ข้อมูลหรือการประยุกต์วิธีธรรมดาที่นำไปสู่เนื้อหาใหม่จะเกิดขึ้นเสมอๆเพราะข้อมูลใหม่ที่ปรากฏและวิธีการเก็บรวบรวมข้อมูลที่ได้ยากและปัญหาเชิงสถิติใหม่ๆ ในส่วนข้อมูลทางชีววิทยา สิ่งแวดล้อมและการแพทย์ ผู้วิจัยได้ให้ความสนใจการประเมินตัวชี้วัดราคาและสารอาหารทั้งหมด 11 ชนิด ได้แก่ พลังงาน คาร์โบไฮเดรต ไขมัน โปรตีน แคลเซียม เหล็ก วิตามินเอ วิตามินบี 1 วิตามินบี 2 วิตามินซี และวิตามินบี 3 จากกลุ่มของผัก อาหาร ผลไม้ และขนมหวาน ตัวชี้วัดสารอาหารได้ถูกประเมินและเปรียบเทียบโดยใช้วิธีเกณฑ์การตัดสินใจพหุคูณและวิธี Electre ค่าเฉลี่ยดัชนีทั้งหมด ได้ถูกจัดอันดับจากค่าสูงสุดไปยังค่าต่ำสุดโดยอาศัยตัวชี้วัดสารอาหารและราคาทั้งหมด ดังนั้น จะเห็นได้ว่าวิธีเกณฑ์การตัดสินใจพหุคูณและวิธี Electre เหมาะสมในการเปรียบเทียบและการวัดจัดอันดับสำหรับผัก อาหาร ผลไม้ และขนมหวานแต่ละชนิดในแต่ละประเภท อนุกรมเวลาได้นำมาใช้สำหรับการพยากรณ์ปริมาณน้ำฝนและระดับน้ำรายเดือนของแม่น้ำเจ้าพระยาในประเทศไทย เทคนิคบอซซ์-เจนกินส์ได้ถูกนำมาใช้ในการระบุพารามิเตอร์ของตัวแบบการถดถอยในตัวเองรวมค่าเฉลี่ยเคลื่อนที่ (ARIMA) เกณฑ์ที่ใช้ในการพิจารณาความเหมาะสมของตัวแบบคือ เกณฑ์สารสนเทศของอาคาอิกะ (AIC) เกณฑ์ซวาร์ชเบเชียน (SBC) และค่าประมาณความแปรปรวนของส่วนรบกวน การวิเคราะห์แผนภาพเป็นคาบได้ถูกนำมาใช้เพื่อยืนยันการมีอยู่ของคาบฤดูกาลในตัวแบบ ARIMA ตัวแบบ ARIMA แบบฤดูกาลอาจจะใช้ทำนายปริมาณน้ำฝนรายเดือนและระดับน้ำรายเดือนของแม่น้ำเจ้าพระยา หนึ่งปีล่วงหน้าด้วยความแม่นยำที่ยอมรับได้ ผู้วิจัยได้นำเสนอตัวแบบอภิปหุคูณผสมไม่เชิงเส้นสำหรับข้อมูลที่มีการวัดซ้ำในช่วงเวลา เพื่ออธิบายการเปลี่ยนแปลงระดับ Carcinoembryonic Antigen (CEA) ของผู้ป่วยโรคมะเร็งลำไส้ใหญ่ ระดับ CEA ของผู้ป่วยโรคมะเร็งลำไส้ใหญ่ได้ใช้เป็นตัวบ่งชี้ตัวหนึ่งสำหรับการดูแลผู้ป่วยโรคมะเร็งลำไส้ใหญ่ กล่าวคือถ้าระดับ CEA ในผู้ป่วยโรคมะเร็งลำไส้ใหญ่ ภายหลังได้รับการผ่าตัด มีค่าสูงขึ้นเรื่อยๆ บ่งถึงว่าอาจมีการกลับมาเป็นซ้ำของโรคอีก ผู้วิจัยได้เสนอตัวแบบอภิปหุคูณผสมไม่เชิงเส้นของระดับ CEA ของผู้ป่วยมะเร็งลำไส้ใหญ่ โดยอภิปหุคูณผสมเกิดจากอภิปหุคูณที่และ

อิทธิพลสุ่ม พารามิเตอร์ ϕ_2 ในตัวแบบที่นำเสนอใช้แทนเวลาเมื่อเกิดการกลับมาเป็นซ้ำของโรค(จำนวนเดือนภายหลังการผ่าตัด) พารามิเตอร์ในตัวแบบประมาณค่าใช้ขั้นตอนวิธี Lindstrom and Bates (LB) และ Stochastic Approximation version of the standard Expectation and Maximization step (SAEM) ผลการศึกษาพบว่า ตัวประมาณค่าโดยใช้ขั้นตอนวิธีLB และ SAEM ให้ค่ากลับมาเป็นซ้ำเมื่อผู้ป่วยได้รับการผ่าตัดไปแล้ว 17.4 และ 22.8 เดือน ตามลำดับ ตัวแบบการตอบสนองขนาดยาคือฟังก์ชันทางคณิตศาสตร์ที่เกี่ยวข้องกับขนาดยาและการวัดผลกระทบล่าสังเกตุ ตัวแบบที่พิจารณา 3 ตัวแบบคือ ตัวแบบไวบูลล์หลายขั้นตอน ตัวแบบลอจิสติกและตัวแบบลอก-ลอจิสติก การศึกษาครั้งนี้ให้ความสนใจกับการประมาณค่าพารามิเตอร์และการเปรียบเทียบตัวประมาณค่าจากตัวแบบทั้ง 3 ตัวแบบดังกล่าวด้วยวิธีภาวะน่าจะเป็นสูงสุดและวิธีโคกกำลังสองต่ำสุดของเบอร์คสัน วิธีประมาณค่าพารามิเตอร์ทั้งสองวิธีได้ประยุกต์ใช้กับข้อมูล 4 ชุดการทดลองสำหรับแต่ละตัวแบบ สุดท้ายผู้วิจัยให้ความสนใจเชิงแนวคิดของความเพียงพอและสารสนเทศของฟิชเชอร์สำหรับการเปรียบเทียบการทดลองที่แตกต่างกันของประชากรแบบทวิปรกติ เมื่อกำหนดค่าเฉลี่ยและค่าแปรปรวนในรูปแบบต่างๆของพารามิเตอร์สหสัมพันธ์(ρ)และประยุกต์แนวคิดความเพียงพอสำหรับตัวแบบถดถอยเชิงเส้นอย่างง่าย $y_x = \alpha + \beta x + e_x, e_x \sim N(0, \sigma^2), x \in [-1, 1]$ การถดถอยไม่ใช่เชิงสุ่ม และสำหรับตัวแบบเชิงสุ่มปรกติทางเดียว

คำสำคัญ: เกณฑ์สารสนเทศของอากาอิเคะ, การเปรียบเทียบขนาดตัวอย่างใหญ่, วิธีโคกกำลังสองต่ำสุดของเบอร์คสัน, ประชากรปรกติสอง/สามตัวแปร, ตัวแบบบอซซ์-เจนกินส์, สารก่อกุมิต้านทานคาร์ซิโนเอมไบรโอนิก, วิธีอิลคเตอร์, สารสนเทศของฟิชเชอร์, อิทธิพลตรึง, สารสนเทศ, ตัวแบบเชิงเส้น, ตัวแบบลอจิสติก, ตัวแบบลอก-ลอจิสติก, ความควรจะเป็นสูงสุด, การจัดสรรเอกรูปยั้ง, ตัวแบบไวบูลล์หลายขั้นตอน, การตัดสินใจเกณฑ์พหุคูณ, ตัวแบบอิทธิพลผสมไม่เชิงเส้น, ตัวชี้วัดสารอาหาร, ตัวแบบอิทธิพลสุ่มทางเดียว, การออกแบบที่เหมาะสม, ราคา, อิทธิพลสุ่ม, การวัดซ้ำ, เกณฑ์ชวาร์ชเบเชียน, การทดลองพอเพียง, พอเพียง, ความแปรปรวนรบกวน