ชื่อวิทยานิพนธ์ วิธีการสร้างตัวแบบสำหรับตัวแปรตามที่มีค่าความแปรปรวนสูงกว่าค่าเฉลี่ย โดยประยุกต์ใช้กับอัตราอุบัติการณ์ของโรคและความหนาแน่นของสัตว์หน้าดินขนาดใหญ่

ผู้เขียน นางสาวนุชนาถ คงช่วย

สาขาวิชา วิธีวิทยาการวิจัย

ปีการศึกษา 2552

บทคัดย่อ

การวิเคราะห์ถดถอย (Regression analysis) รวมทั้งตัวแบบต่างๆ ซึ่งเป็นส่วนขยายจากการวิเคราะห์การถดถอยได้พัฒนาไปสู่ตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized linear model: GLMs) ได้ถูกนำไปใช้อย่างแพร่หลายในการศึกษาทางวิทยาศาสตร์ชีวภาพ ส่วน Canonical correspondence analysis (CCA) เป็นวิธีที่ได้รับความนิยมใช้อย่างกว้างขวางในกลุ่มนักนิเวศวิทยา แต่วิธีดังกล่าวได้รับความนิยมน้อยในกลุ่มนักวิทยาศาสตร์สาขาอื่น ๆ จึงทำให้เกิดคำถามว่าทำไมนักวิทยาศาสตร์ทางการแพทย์ซึ่งศึกษาค้นคว้าเกี่ยวกับโรคที่เกิดในมนุษย์ และนักวิทยาศาสตร์ทางนิเวศวิทยา ศึกษาเกี่ยวกับความชุกชุมของสัตว์และพืช จึงนิยมใช้วิธีการทางสถิติที่แตกต่างกัน ในส่วนที่มีโครงสร้างข้อมูล (Data structure) ที่สอดคล้องกัน มีลักษณะเหมือนกัน

วิทยานิพนธ์นี้ ประยุกต์วิธีการสร้างตัวแบบของตัวแปรตามที่มีค่าความแปรปรวนสูงกว่าค่าเฉลี่ยโดยใช้ข้อมูลทางด้านระบาดวิทยาและนิเวศวิทยา สำหรับข้อมูลระบาดวิทยาประกอบด้วย (1) อุบัติการณ์โรคปอดบวมของเด็กในจังหวัดสุราษฎร์ธานี ระหว่างปี พ.ศ. 2542 – 2550 และ (2) อุบัติการณ์โรควัณโรคใน 14 จังหวัดภาคใต้ของประเทศไทย ระหว่างปี พ.ศ. 2542 – 2547 ทั้งสองการศึกษาที่มีตัวแปรอิสระเหมือนกัน คือลักษณะทางด้านประชากร ได้แก่ สถานที่อยู่อาศัย ฤดูกาล อายุ และเพศ วัตถุประสงค์ของสองการศึกษานี้เพื่อหาความสัมพันธ์ระหว่างปัจจัยทางด้านประชากรกับอุบัติการณ์ของโรค นอกจากนี้ ผู้วิจัยยังใช้ข้อมูลนิเวศวิทยาเพื่อหาความสัมพันธ์ระหว่างความหนาแน่นของสัตว์หน้าดินขนาดใหญ่ 24 วงศ์กับปัจจัยด้านสิ่งแวดล้อม ที่เก็บรวบรวมข้อมูลมาจากทะเลสาบสงขลาตอนกลาง ในช่วงเดือนเมษายน 2541 ถึงกุมภาพันธ์ 2542

ตัวแบบการถดถอยทวินามนิเสธ (Negative binomial regression model) และตัวแบบการถดถอยเชิงเส้นลอการิทึม (log-transformed linear regression model) ประยุกต์ใช้กับการศึกษา

ข้อมูลระบาดวิทยา ผลการศึกษาพบว่าตัวแบบการถดถอยเชิงเส้นลอการิทึมเป็นวิธีที่เหมาะสม อีกทั้งและยังวิเคราะห์ได้ง่ายและเข้าใจได้ง่ายกว่าตัวแบบการถดถอยทวินามนิเสธ

สำหรับวิธีการวิเคราะห์ข้อมูลนิเวศวิทยานั้น เริ่มต้นจากแปลงข้อมูลความหนาแน่นของสัตว์หน้าดินขนาดใหญ่โดยใช้ลอการิทึมฐานธรรมชาติ แล้วเปรียบเทียบตัวแบบการถดถอยเชิงเส้นแบบพหุที่มีตัวแปรตามมากกว่าหนึ่ง (Multivariate multiple linear regression model: MMR) กับ CCA ผลการศึกษาพบว่า CCA มีประโยชน์ในการแสดงภาพความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามในสองมิติ (bi plot) แต่วิธีนี้ยังมีข้อด้อยกว่า MMR โดย bi plot มีข้อจำกัดสำหรับแสดงความสัมพันธ์นสองมิติที่ไม่สามารถแสดงความสัมพันธ์ที่มีนัยสำคัญทางสถิติได้คลอบคลุมในภาพเดียว ยังต้องแสดงกราฟในมิติอื่นๆ ประกอบ นอกจากนี้ MMR ให้ค่าผลลัพธ์ของความสัมพันธ์แต่ไม่มีการแสดงค่าใน CCA

โดยสรุป จากผลการศึกษาที่พบว่า ตัวแบบการถดถอยเชิงเส้นของลอการิทึมเป็นวิธีที่เหมาะสมสำหรับการศึกษาอุบัติการณ์ของโรคปอดบวมและวัณโรค ดังนั้นผลการศึกษานี้มีประโยชน์อย่างยิ่งที่จะนำไปประยุกต์ใช้กับการศึกษาทางระบาดวิทยาได้ ในทำนองเดียวกันสามารถประยุกต์ใช้ MMR กับตัวอย่างข้อมูลอื่นๆ ทางนิเวศวิทยา และระบาดวิทยาได้ด้วย

Thesis Title	Methods for Modeling Overdispersed Outcomes with Application to Disease Incidence Rates and Macrobenthic Fauna Densities
Author	Miss Noodchanath Kongchouy
Major Program	Research Methodology
Academic Year	2009

ABSTRACT

Regression analysis and its various extensions to generalized linear models (GLMs) have been commonly used in biological studies whereas canonical correspondence analysis (CCA) is a popular method widely used among ecologists. But this method is rarely used by scientists in other disciplines. These considerations prompted us to ask why medical scientists investigating human diseases and ecological scientists investigating animal and plant abundances use such different methods, given that the corresponding data structures are essentially the same.

This thesis applies the methods for modeling overdispered outcomes for epidemiological and ecological data. Two epidemiology datasets comprising children's pneumonia incidence and their demographic determinants (location, season, age, and gender) from Surat Thani province during 1999 – 2007, and another dataset consisting of tuberculosis incidence with similar determinants from 14 provinces in southern Thailand for years 1999 – 2004, are collected. The objective of the two studies was identification of the associations between demographic factors and incidence outcomes. We also used ecological data collected from middle

Songkhla Lake during 1998-1999 for analyzing associations between 24 families of macrobenthic fauna densities and their environmental determinants.

Negative binomial regression model and a log-transformed linear regression model were applied to fit these first two studies providing support for the simpler alternative method.

For an analysis of ecological data, we first transformed the density outcome variables using natural logarithms. We then compared the multivariate multiple linear regression model with the canonical correspondence analysis. We concluded that canonical correspondence analysis is useful for informatively displaying the associations in a two-dimensional biplot, but this method is inferior to multivariate multiple linear regression. The biplot is restricted to two dimensions and can fail to show some associations that are statistically significant but require a further dimension to be seen in such a plot. The multivariate multiple linear regression model also gives fitted values for their outcomes but canonical correspondence analysis does not.

In conclusion, the log-transformed linear regression model was found to be superior to the method preferred by biostatisticians in each of our investigations of disease incidence, so it would be useful to know whether this finding is supported more generally in such epidemiological studies. Similarly, it would be useful to apply the multivariate multiple regression method to additional examples of both ecological and epidemiological studies.