Predict 410 – Regression & Multivariate Analysis – Winter 2015

Syamala Srinivasan, Ph.D.
syamala.srinivasan@northwestern.edu

Course Description

This course develops the foundations of predictive modeling by: introducing the conceptual foundations of regression and multivariate analysis; developing statistical modeling as a process that includes exploratory data analysis, model identification, and model validation; and discussing the difference between the uses of statistical models for statistical inference versus predictive modeling. The high level topics covered in the course include: exploratory data analysis, statistical graphics, linear regression, automated variable selection, principal components analysis, exploratory factor analysis, and cluster analysis. In addition students will be introduced to the SAS statistical software, and its use in data management and statistical modeling.

Required Texts

[1] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. (5th Edition). New York, NY: Wiley [ISBN-13: 978-0470542811]

[2] Fabrigar, L.R. and Wegener, D.T. (2011). *Exploratory Factor Analysis*. New York, NY:Oxford University Press [ISBN-13: 978-0199734177]

[3] Everitt, B.S., Landau, S., Leese M., and Stahl D. (2011). *Cluster Analysis* (5th Edition). New York, NY: Wiley. [ISBN-13: 978-0470749913]

[4] Cody, R. (2011). *SAS Statistics By Example*. Cary, N.C.: SAS Publishing. [ISBN-13 978-1607648000]

[5] Delwiche, L., and Slaughter, S. (2012). *The Little SAS Book: A Primer*. (5th Edition). Cary, NC: SAS Publishing. [ISBN-13: 978-1612903439]

Optional Texts

Our primary text has a solutions manual available. Students are encouraged to use this solutions manual to work problems from the book, both theoretical and applied.

[1] Ryan, A. G. Montgomery, D.C., Peck, E.A., and Vining, G.G. (2013). *Solutions Manual to Introduction to Linear Regression Analysis*. New York, NY: Wiley [ISBN-13: 978-1118471463]

SAS Reference Reading

SAS will be used as the statistical software for both PREDICT 410 and PREDICT 411. In PREDICT 410 we will use SAS to manipulate data, produce statistical graphics, and perform statistical analyses. The two primary SAS reference books for PREDICT 410 are Cody (2011) and Delwiche and Slaughter (2012). The combination of these two books provides an overview and example syntax for many of the SAS capabilities that we will use in this course. Students should review this material before the beginning of the course and also reference it as needed throughout the course. Students should consider these two books to be their primary SAS references for PREDICT 410. Additional SAS support will also be provided in the form of handouts, sample code, and code snippets will be provided for the topics not covered in these two books.

[1] Cody (2011) Chapters 2-4, pp. 19-68
[2] Cody (2011) Chapter 8-9, pp. 111-162

[3] Delwiche and Slaughter (2012) Chapter 3-4, pp. 73-148
[4] Delwiche and Slaughter (2012) Chapter 6, pp. 177-208
[5] Delwiche and Slaughter (2012) Chapter 8-9, pp. 227-280

Required Readings on Library Reserve

These readings will be made available within the course site or through Library Reserves.

[1] Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. (2nd Edition). Los Angeles, CA: Sage Publications, Inc. [ISBN-13: 978-0761930426] – Chapters 2 and 3.

[2] Ratner, B. (2012). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*. (2nd Edition), Boca Raton, FL: CRC Press. [ISBN-13: 978-1439860915] – Chapter 10.

[3] Morrison, D.F. (2004). *Multivariate Statistical Methods*.(4th Edition). Cengage Learning. [ISBN-13: 978-0534387785] – Chapters 6-7.

[4] Everitt, B.S. & Dunn, G. (2001). *Applied Multivariate Data Analysis*. (2nd Edition). New York, NY: Wiley. [ISBN-13: 9780470711170] – Chapters 3, 6, and 12.

Software

SAS Institute, Inc. provides on-cost access to SAS® OnDemand for Academics for students registered in classes like this. Each PREDICT class has its own SAS location within SAS® OnDemand for Academics.

See Guide to Gaining Access to SAS for PREDICT 410 for instructions on setting up your computer to access the SAS Server/Cloud for this course.

Prerequisites - PREDICT 401

Learning Goals

The goals of this course are to:

- Develop statistical modeling as a three step process consisting of: (1) exploratory data analysis, (2) model identification, and (3) model validation.
- Understand how to use automated variable selection as a tool for model identification and as a tool for exploratory data analysis in the presence of a large number of predictor variables or a set of unlabeled predictors.
- Develop a working understanding of the conceptual (theoretical) foundations of linear regression, principal components analysis, factor analysis, and cluster analysis with the objective of being capable of applying these techniques appropriately and validating their results.
- Develop a conceptual and practical understanding of the difference between statistical inference and predictive modeling and how it affects our choices and actions in the statistical modeling process.
- Learn the basics of the SAS Data Step, data manipulation with SAS, and SAS procedures (PROCS) for fitting statistical models.