University of Houston
Bauer College of Business

Marketing Models MARK8335
Spring, 2015

Instructor

Professor Sam K. Hui
Melcher 375J
Email: khui@bauer.uh.edu

Course Overview

The purpose of this doctoral seminar is to provide students with some exposure of empirical modeling in marketing. I also aim to provide a "roadmap" of various modeling issues so that students know where to go for further study/reference. After taking this course, students should have an understanding of the various issues surrounding data analysis and model building in marketing.

The course is divided into three main modules: (i) descriptive analysis of data, (ii) modeling concepts, and (iii) issues in causal inference.

Module I—Data Description: The first step of any empirical project is to explore the data and discover patterns. In the first module, we review several basic statistical techniques to deal with different kinds of data. For each technique, we will review papers that are published in the major journals.

- Linear regression and its statistical properties
- Count data (Poisson regression)
- Binary data (logit, probit)
- Choice data and introduction to choice modeling

Module II—Modeling concepts: After exploring the data, the next step is to develop a model to describe the data. In the second module, we discuss some key concepts in empirical modeling. These concepts are common across many different models.

- Using latent parameters to "explain" observed behavior
- Capturing dynamics in latent parameters
- Capturing "heterogeneity" through shrinkage estimation

Module III—Causal Inference: Making causal statements based on observational data is challenging and involves a whole new set of issues. In the third and final module, we discuss several issues surrounding the area of causal inference. We begin by reviewing the fundamentals of causal inference, then move on to discuss specific issues of endogeneity and sample selection in detail.

Course Requirements

- *Homework assignments (30%)*: There will be six short homework assignments covering the essential ideas of each topic.

- *Midterms (30%):* There will be a short in-class midterm given at the end of each module (3 midterms in total). The goal of these midterms is to encourage you to review the class material and know them by heart before we move on to the next module.

- *Final project (30%):* There will be a short final project that students will present during the last class. See the "Final project" section for details.

- *Class attendance & participation (10%)*: You are expected to attend every class on time and well prepared.

Course Material

- *Textbook*: We will cover a few chapters from the textbook: *Statistical Models: Theory and Practice* by David A. Freedman. You can order the book from Amazon at http://www.amazon.com/Statistical-Models-Practice-David-Freedman/dp/0521743850. I will also assign some other readings in class based on other book chapters and journal articles. Please refer to the Course Schedule for details.

- *Software*: You are required to use the R statistical software to complete the homework assignments.

Course Schedule

#	Date	Topic	Assigned Readings
Module I: Data Description			
1	W Jan 21	Course Overview **[HW#1 out]**	Freedman (2009), Ch.1, 2
2	M Jan 26	Linear Regression	Freedman (2009), Ch. 4
3	W Jan 28	Review (no class meeting)	Herzenstein et al. (2011) Schmitt et al. (2011)
4	M Feb 2	Count data: MLE and Poisson Regression **[HW#1 due; HW#2 out]**	Freedman (2009) p109-113 Agresti (1996) Ch.4
5	W Feb 4	Binary data: Logit and Probit Regression	Agresti (1996) Ch.5
6	M Feb 9	Introduction to Choice Modeling	Louviere et al. (2000) Ch.3, Ch. 6
7	W Feb 11	Review (no class meeting)	Bell et al. (2011) Deng et al. (2010)
8	M Feb 16	Midterm I **[HW#2 due]**	
Module II: Modeling concepts			
9	W Feb 18	Using latent parameters to explain observed behavior I **[HW#3 out]**	Fader and Hardie (2009) Schmittlein et al. (1987)
10	M Feb 23	Using latent parameters to explain observed behavior II	Fader et al. (2005) Fader et al. (2010)
11	W Feb 25	Capturing dynamics in latent parameters	Rabiner (1989)
12	M Mar 2	Review (no class meeting)	Du and Kamakura (2006) Netzer et al. (2008)
13	W Mar 4	Bayesian model: Shrinkage estimation and heterogeneity I **[HW#3 due; HW#4 out]**	Efron and Morris (1977)
14	M Mar 9	Bayesian model: Shrinkage estimation and heterogeneity II	Rossi and Allenby (2003)
15	W Mar 11	Review (no class meeting)	Talukdar et al. (2002) Abe (2009)
16	M Mar 23	Midterm II **[HW#4 due]**	

Module III: Causal Inference			
17	W Mar 25	Basic Concepts of Causal Inference **[HW#5 out]**	Gelman and Hill (2006), Ch.9
18	M Mar 30	Advanced Issues in Causal Inference	Gelman and Hill (2006), Ch.10
19	W Apr 1	Review (no class meeting)	
20	M Apr 6	Endogeneity I **[HW#5 due; HW#6 out]**	Stock and Watson (2011), Ch.9, 12 Freedman (2009), Ch. 8
21	W Apr 8	Endogeneity II	Hui et al. (2013)
22	M Apr 13	Sample selection I	Eliashberg et al. (2014) Krishnamurthy et al. (2015)
23	W Apr 15	Sample selection II	Heckman (1979)
24	M Apr 20	Review (no class meeting)	
25	W Apr 22	Midterm III **[HW#6 due]**	
Final Project			
26	M Apr 27	Preparation for final project presentation	
27	W Apr 29	Preparation for final project presentation	
28	M May 4	Final project presentation	

Final Project and Presentation

Please submit a short paper, 20 pages or less (inclusive of table, figures, references, appendix, everything) on a topic of your interest within the area of quantitative marketing research. You may choose one of the following three options. You will present your project on the last class (5/4/2015).

Option 1: Pick an advanced methodological area that is not covered in class. Learn the relevant literature and write a short "tutorial paper" on the topic. You should also provide some ideas of how the method/technique/topic can be used in marketing research. Specific examples of topics may include, but is not limited to:

- Machine learning methods (e.g., SVM, neural networks, tree-based models, unsupervised learning, nearest-neighbor regression)
- Functional data analysis
- Collection and analysis of neuroscience data (e.g., fMRI data)
- Survival analysis
- Time series models
- Bayesian nonparametrics
- Text mining
- Optimal stopping models
- Spatial / Spatio-temporal models
- etc…

(You can pick any methodological area that is non-trivial and applicable to marketing research).

Option 2: You may also go into more depth in one of the topics we covered in class, by reviewing how the model can be expanded / elaborated, and review how such elaboration can be useful to marketing research. For example (the elaboration of a given model is shown after the → sign):

- Linear regression → regression with variable selection.
- Poisson regression → poisson regression with over-dispersion & its applications
- Logistic/probit regression → other kinds of choice models (e.g., nested-logit, tobit, etc.)
- Probability models → more recent developments on the subject
- Other applications of HMM in marketing
- Hierarchical Bayesian models in marketing & other applications
- Applications of instrument variable approach in marketing
- Applications of propensity scores

(You can pick any topic as long as it goes beyond what we discussed in class, and is applicable to marketing research).

Option 3: You may analyze an actual dataset with an appropriate empirical technique, and write up an empirical paper on the subject. If you take this option, you should discuss the dataset with me first. You will be evaluated by how proficient your empirical analysis is and how well the empirical analysis is tied to the central "story" of the paper.

Detailed Description of Each Session

Module I: Data Description

Session 1--- Course Overview (Jan 21)

Overview: We overview the scope of the course, and discuss the main purpose of modeling. We discuss why empirical data can be "messy", and the associated methodological challenges through a few real-life examples. Finally, we review simple linear regression and how to estimate a regression in R.

Readings:
Freedman, David A. (2009), *Statistical Models: Theory and Practice*, Cambridge University Press, New York, NY. (Chapter 1,2)

Session 2--- Linear Regression (Jan 26)

Overview: We talk about the intuition behind multiple regression, and why we need to "control" for other variables. We then demonstrate how to estimate a multiple regression model in R. We discuss the OLS estimate and its desirable statistical properties, and explain why the Gauss-Markov theorem means. We then overview the normal theory, and the associated t-test and F-test.

Readings:
Freedman, David A. (2009), *Statistical Models: Theory and Practice*, Cambridge University Press, New York, NY. (Chapter 4)

Session 3--- Review (Jan 28)

Overview: This session is for students to review course material and do the homework assignment. We are not going to meet in class. In addition to reviewing the course material up to this point, students should read Herzenstein et al. (2011) and Schmitt et al. (2011) to see how regression is used to provide empirical results in a top journal publication. Note that Schmitt et al. (2011) is the recipient of the MSI/Paul Root Award in the *Journal of Marketing* in 2011.

Readings:
Herzenstein, Michal, Scott Sonenshein, and Utpal M. Dholakia (2011), "Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions," *Journal of Marketing Research*, 48 (Nov), 138-149.

Schmitt, Philipp, Bernd Skiera, and Christophe Van den Bulte (2011), "Referral Programs and Customer Value," *Journal of Marketing*, 75, 46-59.

Session 4--- Count data: MLE and Poisson Regression (Feb 2)

Overview: In this session we discuss another type of data commonly found in marketing research: count data. We begin with an overview of MLE (maximum likelihood estimation) and

introduce the Poisson regression to handle count data. I will demonstrate how to use R to estimate a Poisson regression with a real dataset from my research.

Readings:
Freedman, David A. (2009), *Statistical Models: Theory and Practice*, Cambridge University Press, New York, NY. (Page 109--113)

Agresti, Alan (1996), *An Introduction to Categorical Data Analysis*, Wiley. (Chapter 4)

Session 5--- Binary data: Logit and Probit Regression (Feb 4)

Overview: In this session, we discuss the analysis of binary data (0/1). We start with the logistic regression model and its statistical properties. I will demonstrate how to estimate a logistic regression in R. Next, we explore the parallel between logistic regression and the "latent variable formulation" in random utility theory. We then discuss probit regression.

Readings:
Agresti, Alan (1996), *An Introduction to Categorical Data Analysis*, Wiley. (Chapter 5)

Session 6--- Introduction to Choice Modeling (Feb 9)

Overview: Following up on the logit/probit model, this session provides an introduction to the modeling of consumer choice behavior. We start with random utility theory to arrive at the logit choice model, then explore the use of "choice-type" model in other less-traditional setting. Then, we introduce the nested-logit model and discuss how to estimate such model.

Readings:
Louviere, Jordan J., David A Hensher, and Joffre D. Swait (2000), *Stated Choice Methods: Analysis and Application*, Ch.3 and Ch. 6.

Session 7--- Review (Feb 11)

Overview: This session is for students to review course material and do the homework assignment. We are not going to meet in class. In addition to reviewing the course material up to this point, students should read Bell et al. (2011) and Deng et al. (2011) to see how Poisson regression is used in various different settings.

Readings:
Bell, David R., Daniel Corsten, and George Knox (2011), "From Point of Purchase to Path to Purchase: How Preshopping Factors Drive Unpalnned Buying," *Journal of Marketing*, 31-45.

Deng, Xiaoyan, Sam Hui, and J. Wesley Hutchinson (2010), "Consumer Preferences for Color Combinations: An Empirical Analysis of Similarity-Based Color Relationships," *Journal of Consumer Psychology*, 20(4), 476-484.

Session 8--- Midterm I (Feb 16)

Midterm is closed-book, closed notes. You are allowed to bring a one-page letter-size "cheat sheet".

Module II: Modeling concepts

Session 9--- Using Latent Parameters to Explain Observed Behavior I (Feb 18)

Overview: At the start of Module II, we switch our attention from describing data to building an empirical model of the data. This session discusses one of the main concepts of modeling building: the use of "latent" parameters to explain observed behavioral outcome. Customer-base analysis is used as an example. We start with an introduction of the customer-base analysis setting in Fader and Hardie (2009). Next, we review some key concepts in probability theory, and then discuss the Pareto-NBD model developed by Schmittlein et al. (1987), a widely cited paper in customer analytics.

Readings:
Fader, Peter S., and Bruce G. S. Hardie (2009), "Probability Models for Customer-Base Analysis," *Journal of Interactive Marketing*, 23, 61-69.

Schmittlein, David, G. Morrison, and Richard Colombo (1987), "Counting Your Customers: Who They Are and What Will They Do Next?" *Management Science*, 33 (Jan), 1-24.

Session 10--- Using Latent Parameters to Explain Observed Behavior II (Feb 23)

Overview: Building on the Pareto-NBD model, we will discuss two more papers in the field of customer-based analytics. The first is Fader et al. (2005), which won the Paul Green Award at JMR. The next is Fader et al. (2010) that extends the Pareto-NBD model to a discrete-time setting, which results in the BG/BB model. Finally, I will show you how the BG/BB model is applied to develop an analytic tool in an actual gaming company.

Readings:
Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005), "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42, 415-430.

Fader, Peter S., Bruce G. S. Hardie, and Jen Shang (2010), "Customer-Base Analysis in a Discrete-Time Contractual Setting," *Marketing Science*, 29(6), 1086-1108.

Session 11--- Capturing Dynamics in Latent Parameters (Feb 25)

Overview: In Session #9 and #10, we begin with the setting where latent parameters do not change over time. In this session, we move to the second key modeling concept where the value of latent parameters may change over time, which necessitates the need to capture "dynamics" in latent parameters. We then discuss the concepts of hidden Markov model (HMM), a common approach to capture dynamics in latent parameters.

Readings:
Rabiner, Lawrence R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77(2), 257-286.

Session 12--- Review (Mar 2)

Overview: This session is for students to review course material and do the homework assignment. We are not going to meet in class. In addition to reviewing the course material up to this point, students should read Du and Kamakura (2006) and Netzer et al. (2008), both of which uses the HMM to model marketing data. Note that Netzer et al. (2008) is the recipient of the John Little Award in the *Journal of Marketing Research* in 2008.

Readings:
Du, Rex, and Wagner A. Kamakura (2006), "Household Lifecycles and Life Styles in America," *Journal of Marketing Research*, 43(Feb), 121-132.

Netzer, Oded, James M. Lattin, V. Srinivasan (2008), "A Hidden Markov Model of Customer Relationship Dynamics," *Marketing Science*, 27(2), 185-204.

Session 13 --- Bayesian Model : Shrinkage Estimation and Heterogeneity I (Mar 4)

Overview: We will spend this session (Session #13) and the next (Session #14) to discuss the third key concept of modeling: "allowing for heterogeneity" across consumers. We will start with a discussion of Efron and Morris (1977), which discuss the need for "shrinkage estimation" and why borrowing information will lead to superior estimates and prediction. This motivates the use of hierarchical Bayesian models to capture heterogeneity across consumers.

Readings:
Efron, Bradley, and Carl Morris (1977), "Stein's Paradox in Statistics," *Scientific American*, 236(5), 119-127.

Session 14 --- Bayesian Model : Shrinkage Estimation and Heterogeneity II (Mar 9)

Overview: We provide several examples of how hierarchical Bayesian model is used in marketing models, through the review paper by Rossi and Allenby (2003).

Readings:
Rossi, Peter E., and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22(3), 304-328.

Session 15 --- Review (Mar 11)

Overview: This session is for students to review course material and do the homework assignment. We are not going to meet in class. In addition to reviewing the course material up to this point, students should read Talukdar et al. (2002) and Abe (2009) and understand how a

Hierarchical Bayesian framework is used in both papers. Abe (2009) is a hierarchical Bayesian extension of the Pareto-NBD model discussed in Session #9.

Readings:
Talukdar, Debabrata, K. Sudhir, and Andrew Ainslie (2002), "Investigating New Product Diffusion Across Products and Countries," *Marketing Science*, 21, 97-116.

Abe, Makoto (2009), "Counting Your Customers One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model," *Marketing Science*, 28(3), 541-553.

Session 16--- Midterm II (Mar 23)

Midterm is closed-book, closed notes. You are allowed to bring a one-page letter-size "cheat sheet".

Module III: Causal Inference

Session 17--- Basic Concepts of Causal Inference (Mar 25)

Overview: The third and final module of this class is focused on the issue of causal inference. We begin by stating the fundamental problem of causal inference, and explains the concepts of "ignorability" in detail.

Readings:
Gelman, Andrew, and Jennifer Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press. (Chapter 9: "Causal Inference Using Regression on the Treatment Variable")

Session 18--- Advance Issues in Causal Inference (Mar 30)

Overview: We discuss several more advance issue in causal inference, and provide an overview of several methods to tackle such problems: matching, sub-classification, and regression discontinuity.

Readings:
Gelman, Andrew, and Jennifer Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press. (Chapter 10: "Causal Inference Using More Advanced Models")

Session 19--- Review (Apr 1)

Overview: This session is for students to review course material and do the homework assignment. We are not going to meet in class. Students should review Gelman and Hill (2006), Ch. 9 and 10 thoroughly.

Session 20--- Endogeneity I (Apr 6)

Overview: This session and the next focuses on the problem of endogeneity. We first explain what the problem is and why it matters using a demo in R. We then discuss the common causes of Endogeneity: omitted variables, measurement error, and reverse causality/simultaneity. Then we discuss the standard solution to the endogeneity problem, the instrumental variable (IV) approach. We will show how the method works through an R demo. Finally, we discuss how we can find instruments in real life.

Readings:
Stock, James H., and Mark W. Watson (2011), *Introduction to Econometrics*, Chapter 9, "Assessing Studies Based on Multiple Regression"

Stock, James H., and Mark W. Watson (2011), *Introduction to Econometrics*, Chapter 12 "Instrumental Variables Regression"

Freedman, David A. (2009), *Statistical Models: Theory and Practice*, Cambridge University Press, New York, NY. (Chapter 8: "Simultaneous Equations")

Session 21--- Endogeneity II (Apr 8)

Overview: In this session, I will present one of my recent papers that develops a creative instrumental variable to tackle an endogeneity problem in the grocery setting. Note that Hui et al. (2013) is the recipient of the MSI/Paul Root Award in the *Journal of Marketing* in 2013.

Readings:
Hui, Sam, Jeffrey, Inman, Yanliu Huang, and Jacob Suher (2013), "Estimating the Effect of Travel Distance on Unplanned Spending: Applications to Mobile Promotion Strategies," *Journal of Marketing*, 77 (March), 1-16.

Session 22--- Sample Selection I (Apr 13)

Overview: In this session and the next, we will discuss another major challenge of causal inference: sample selection bias. We will start by discussing what the problem is and why it matters through a R demo. Then, we will discuss two settings in my recent research where the problem of sample selection bias can come up unexpectedly, and what can be done in those cases.

Readings:
Eliashberg, Jehoshua, Sam Hui, and John Z. Zhang (2014), "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, 26(11), 2639-2648.

Krishnamurthy, Parthasarathy, Sam Hui, Narayanan Shivkumar, Chandrasekhar Gowda, and R. Pushpalatha (2015), "Assessing the Effect of Peer Education Outreach on Likelihood and Timing of Health Access among Female Sex Worker using an Extended Cox Model," *Working Paper*.

Session 23--- Sample Selection II (Apr 15)

Overview: We discuss a standard solution to the sample selection problem (Heckman 1979) when additional information on the selection mechanism is available. We will discuss Heckman (1979) and demonstrate how the method works using an R demo.

Readings:
Heckman, James J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153-161.

Session 24--- Review (Apr 20)

Overview: This session is for students to review the course material and do the homework assignment. Given that we went through a lot of material in the last 4 sessions, there is no additional reading for this review session. Students are encouraged to review the material that we went through for "endogeneity" and "sample selection" in preparation for Midterm III.

Session 25--- Midterm III (Apr 22)

Midterm is closed-book, closed notes. You are allowed to bring a one-page letter-size "cheat sheet".

Final Project

Session 26 & 27 --- Project Preparation (Apr 27, Apr 29)

These two sessions are intentionally left blank for students to prepare for project presentations.

Session 28 --- Project Presentations (May 4)

This session is reserved for student presentations.