Course objectives

This sequence of two half-semester courses provides students an understanding and working knowledge of statistical data analysis techniques commonly used in marketing. The focus is on techniques that provide insights into how one variable is predicted (and possibly caused) by other variables.

The courses are designed to complement MKTG 942/943, Research Methods in Marketing (A/B). The latter focuses on (i) linear modeling (linear regression and ANOVA) and on (ii) experimental data. The present courses extend the students’ tool kit in two directions:

1. Analyzing binary data, ordered response data, choice data, count data, truncated or censored data, and duration data using Generalized Linear Models.
2. Identifying and tackling challenges that arise when analyzing non-experimental data.

In short, MKTG940/941 is about “funny Y’s and messy X’s.”

Prerequisites

For MKTG 940: MKTG 942/943, or a graduate / advanced undergraduate level course on regression. For MKTG 941: MKTG 940.

Course format

The class meets once a week, on Mondays 9:00-12:00 PM.

I think of and designed the 940/941 sequence as a single one-semester course. But, since the sequence is administratively split up into two units, it is possible to take MKTG 940 without also taking 941.

There will be several problem sets / homework assignments—typically a weekly assignment for the first 10 weeks. Students will also analyze a data set from a project they are working on or re-analyze the data from a published paper, present the process they went through on the final day of class, and write up a report.
A typical class session will consist of (i) a debrief on the recently submitted homework assignment (if applicable), (ii) a lecture on a given topic, (iii) examples and discussion of practical estimation and interpretation issues.

**List of topics**

Regression-type models for analyzing …
1. Binary data
2. Ordered response data
3. Multinomial / choice data
4. Count data
5. Duration data
6. Other censored/truncated data; Data with selectivity

Identifying and tackling challenges in non-experimental data
1. Influence points, collinearity / ill-conditioning, and missing data
2. Counterfactual / Rubin Causal Model framework
3. Selectivity and other sources of endogeneity
4. Designs and statistical methods to strengthen causal identification
   - Diff-in-Diff; Matching; Regression Discontinuity; Instrumental variables
5. Endogeneity when testing for mediation

**Statistical software**

I will be using SAS in class. Students are welcome to use other statistical packages they are familiar with (e.g., Stata, R, SPSS, JMP) for assignments. All analyses we cover can be performed using SAS, Stata, and R. I am not quite sure about SPSS or JMP, but most analyses definitely can. I promise complete support only for SAS.

**Course materials**

There is no assigned textbook. Class notes, data sets, etc. will be made available on Canvas.

As preparation for the session on analyzing censored, truncated, and self-selected data, I suggest you read the following little book retailing for about $14:


It is also available online through The Penn Library: [http://hdl.library.upenn.edu/1017.12/2246772](http://hdl.library.upenn.edu/1017.12/2246772)
Reference materials: General

There is no assigned (must buy) textbook. The following books provide additional details and background, roughly at the level of the course:


When looking for an accessible primer on a specific topic or technique, consider the “little green books” in the Quantitative Applications in the Social Sciences published by Sage. I found several of them quite useful when I was a student. They are listed here: [http://srmo.sagepub.com/browse?doctype=qass](http://srmo.sagepub.com/browse?doctype=qass)

Reference materials: Software guides

There are several books providing hands-on guidance that you may find useful, also after the course.

For SAS, the following two books by Penn Sociology professor Paul Allison are quite useful:


Similar books exist for Stata and R. Examples for Stata are:


Cameron, A. Colin and Pravin Trivedi. 2010. *Microeconometrics Using Stata, Revised Edition*. Stata Press, College Station, TX.


**Classroom**

Our class meets in 741 JMHH.

**Grading**

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem sets / Assignments</td>
<td>40%</td>
</tr>
<tr>
<td>In-class contributions</td>
<td>10%</td>
</tr>
<tr>
<td>Project Write-up &amp; Presentation*</td>
<td>15%</td>
</tr>
<tr>
<td>Take-Home Final Examination</td>
<td>35%</td>
</tr>
</tbody>
</table>

* The presentation is an opportunity for you to get feedback before finalizing your project. I expect your presentation to be coherent and clear, but I do not grade it for rigor / correctness. Unless your presentation is exceptionally poor or exceptionally good, that 15% of the grade is based on the write-up only.
Plan of Sessions

1. Jan. 15* Introduction – Beyond Classical Linear Regression
   Jan. 20 No class – MLK Day

A. Generalized Linear Models

2. Jan. 27 Beyond Classical Linear Regression & Binary data I
3. Feb. 3 Binary data II
4. Feb. 10 Binary data III
5. Feb. 17 Multinomial data I
6. Feb. 24 Multinomial data II & Ordered response data
7. Mar. 2 Count data
   Mar. 9 No class – Spring Break
8. Mar. 16 Duration data

B. Challenges with Non-Experimental Data

9. Mar. 23 Censored, truncated, and self-selected data
10. Mar. 30 Causal inference from non-experimental data I
11. Apr. 6 Causal inference from non-experimental data II
12. Apr. 13 Causal inference from non-experimental data III
13. Apr. 20 Ill-conditioning, outliers, and influence points
    Repeated/clustered observations and unobserved heterogeneity

C. Topical Applications

14. Apr. 27 Presentation of student analyses

* This is a Wednesday.