

BMGT 881: APPLIED REGRESSION MODELS

Prof. Wolfgang Jank
4322 Van Munching Hall
Tel: (301) 405-1118

E-mail: wjank@rhsmith.umd.edu

Office Hours: *Feel free to drop by anytime you like. To ensure that I am in the office, please send me a quick email or call my office first.*

COURSE OVERVIEW

The nature of business is changing. Due to increasing desktop computing power and companies amassing massive amounts of data, business decisions are becoming more and more data driven. Moreover, not only does more and more data become available, the nature of the data also changes. While in the past, most data arrived in the form of static flat tables, data now becomes more and more dynamic, streaming and augmented with information that was not available previously (geographical information, temporal information, network information). Moreover, data has become more complex and is now available at different layers or hierarchies. For instance, while companies may have individual-level information about their employees, this individual layer is often augmented by company-level information across different industries. Data is also different in that it may combine static and dynamic information. For instance, data from online auctions involve static attributes (such as the product sold) as well as dynamic attributes (such as the incoming bids or the price changes during the auction). Combining static and dynamic information is not trivial and calls for new methods suitable to extract all information that such data carry.

In this class, you will learn different methods and models to address such data challenges. The class starts with basic regression modeling. Basic regression models are extremely powerful and provide a good solution to many data problems. However, you will also learn when basic regression methods break down or, at least, become less desirable. In such situations, alternate models and methods are asked for. Many of these alternate methods build upon the basic principles of regression but refine one (or more) of its aspects to tackle a specific data challenge. There exists a *huge* amount of alternate methods. While it is impossible to cover all possible alternatives in one single class, we will introduce *some* of them. By doing so, you will also be provided with links and references to additional alternate models and methods.

The class is very *hands-on*. Primary focus is on students *using* the methods learned in class and *applying* them to real-world research problems. This includes learning of appropriate software. The class has a strong research flavor in that emphasis is on the key elements of scholarly research: independent research on real research questions using real data, documenting and presenting the findings in the form of scholarly papers and presentations.

INSTRUCTOR:

Wolfgang Jank is associate professor of Decisions, Operations & Information Technologies at the Robert H. Smith School of Business, University of Maryland, and affiliated with the Center for Electronic Markets & Enterprises, the Center for Excellence in Service, and the McGill eSocialLab.. He is interested in applying ideas from statistics and data mining to problems in electronic commerce, marketing, and operations management. Dr. Jank's research has been published in the literature of statistics, data mining, information systems, and marketing. He has authored over fifty refereed articles and book chapters, and presented his work at national and international meetings. Dr. Jank received his Master's degree from the Technical University of Aachen (Germany) and his PhD in Statistics from the University of Florida. After moving to the University of Maryland, he initiated, together with Dr. Shmueli, a new research area on Statistical Challenges in eCommerce. Dr. Jank is member of the American Statistical Society, the Institute of Mathematical Statistics, the European Network for Business and Industrial Statistics, the Association for Computing Machinery and INFORMS. He is past president of the University of Florida's chapter of the statistical honor society Mu Sigma Rho. Prof. Jank has been involved in a variety of consulting projects for private and public organizations, and he is advisory board member for several companies. Prof. Jank is teaching classes in data analytics in various programs (undergraduate, MBA, executive MBA and PhD) at the University of Maryland. He has received numerous awards including the top 15% teaching award for teaching MBA core classes.

- Contact info: wjank@rhsmith.umd.edu (e-mail).
- Website: <http://www.smith.umd.edu/faculty/wjank>

COURSE PRE- REQUISITES:

Students should have a basic understanding of statistics. I will assume that students have mastered a course in introductory statistics and are familiar with basic statistical concepts.

TEXTBOOK AND CLASS MATERIAL

- Required book: *--none--*

I do not require a textbook for this class. However, I encourage you to obtain (purchase or check-out at library) at least one introductory textbook on regression as complementary reading.

I will post class material in the form of lecture notes (PowerPoint slides) and papers from the literature. All material will be posted on Blackboard.

- Some web resources for the review of basic concepts of probability and statistics basics:
By being admitted to the doctoral program, I assume that you are familiar with basics of probability & statistics, especially hypothesis testing and confidence intervals. While we will *briefly review* the main concepts, this review will be rather quick, without going into any details. I would *strongly* recommend that you refresh your knowledge on these basic concepts if you feel that you need to. You can refresh your knowledge by opening your old text book on statistics. Alternatively (or in addition) there are a lot of useful resources on the web; e.g.
<http://www.stat.berkeley.edu/~stark/SticiGui/index.htm>
<http://davidmlane.com/hyperstat/>
<http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html>
<http://www.psychstat.missouristate.edu/sbk00.htm>

SOFTWARE:

We will make extensive use of the software R. R is open source software and available from CRAN (<http://cran.r-project.org/>). CRAN hosts the basic software, add-on packages and a ton of additional reference material. You should spend a good amount of time *at the beginning of class* to check out all the resources available and to familiarize yourself with the software. I will also give a brief introduction to the main concepts during our first meeting.

The following online documents give a very detailed introduction and overview of the software:

- <http://cran.us.r-project.org/doc/manuals/R-intro.pdf>
- <http://cran.us.r-project.org/doc/manuals/R-data.pdf>
- <http://cran.us.r-project.org/doc/contrib/Farnsworth-EconometricsInR.pdf>
- <http://cran.us.r-project.org/doc/contrib/Owen-TheRGuide.pdf>
- <http://cran.r-project.org/doc/contrib/usingR.pdf>

You should go over these documents very carefully within the first weeks to understand the basic principles of R and to get started with the software.

For those of you who would like a paper-back introduction to the software R, I can recommend the following book. But note that the above online resources provide an equally good introduction.

“An R and S-Plus Companion to Applied Regression” by John Fox, Sage Publications, 2002

R is primarily a command-line language. While usage of R is extremely straightforward, you may find a GUI environment even more convenient. The GUI can be obtained from the following link:

<http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>

Some frequently asked questions:

Why do we use R?

R is one of the most powerful statistical software packages. Since it is based on an open source project, it is free! As a consequence, an increasing number of researchers (not only from statistics but also from business areas such as information systems, operations management, management science, marketing, or finance) are using the software. Moreover, since it is an open source project, many researchers frequently contribute to the software’s capabilities by adding new packages. This means that R evolves much faster than traditional (commercial) software packages. Many statistical methods and models that address tomorrow’s data-challenges can already be found in R today. In the first part of the class, we will make use of R’s capabilities for basic regression analyses. While many commercial software packages offer similar capabilities for basic regression, R’s advantage lies in its flexibility, particularly for creating insightful and powerful graphs and analyses of the results. In the second part of the class, we will introduce more advanced methods. Many of these advanced methods are not yet available in standard (commercial) software packages.

Is this an R course?

No, this is not an R course. While I recommend that you use R (especially for the second part of this class), you are not obligated to. Moreover, all regression concepts will be explained independent of any software. That is, while I will post sample R code to implement the concepts, the concepts equally apply to other software solutions (such as SAS or SPSS).

COURSE TECHNOLOGY:

We will use modern clicker-technology for collecting feedback and checking progress (see e.g. http://www.news.com/New-for-back-to-school-Clickers/2100-1041_3-5819171.html or <http://www.oit.umd.edu/ITforUM/2005/Winter/clickers.html>). This technology will allow me to get your feedback in real-time. It will also allow you to perform reality checks relative to the entire class.

COURSE OBJECTIVES

The course objectives can be divided up into *direct* and *indirect* objectives. Direct objectives are directly related to the course topic of regression modeling and statistics & data analysis in general. Indirect objectives are related to general research skills useful for any PhD student.

The main *direct* course objectives are

1. To learn about a wide variety of *regression techniques*; to use and apply regression in real-world business applications; to understand when to use what technique; to understand the limitations of a particular technique. Regression, much like statistics in general, is best appreciated via hands-on application. We will spend much time analyzing data from real-world problems. Much of the class will be driven by motivating the need for a particular technique and explaining its general principles rather than getting bogged-down in all the details and mathematics. This is not to say that these details are unimportant; however, as an introductory class, more emphasis will be spent on creating enthusiasm and appreciation for statistical thought rather than providing you with all methodological detail. In that sense, this class may be different from other statistics classes you have taken in the past. Depending on where your research will lead you, you may want to take another class on regression later on in your career, one that focuses more on the mathematical details which will be under-emphasized here.
2. To be able to read and understand methodological work in *journal papers*. Every week, we will go over one (or more) journal papers that cover some aspect of regression. This will be done by student presentations together with in-class discussion. The idea behind this is to give students an appreciation of the individual aspects of regression in research settings. The focus will be on understanding what goal is to be accomplished by a particular paper, and how it uses regression to achieve that goal. It will *not* be the responsibility of the student to be able to replicate every single detail of statistical methodology; rather, the goal is to understand the *motivation* for the particular regression method and to understand how this method helps to accomplish the goal of the paper.

The main *indirect* course objectives are

1. To improve students' *oral presentation* skills. Every class period, a student will give a short presentation based on one (or more) scholarly papers. The idea is for students to obtain experience in giving oral presentations. One can never have enough experience giving presentations. Oral presentations are key in a research environment (e.g. at colloquia or conferences) and one key factor in your ability of landing a job is your job talk. Preparedness, good structure, conciseness, clarity, and personality are only some of the aspects that are important. This class will give you an opportunity to practice these aspects on a weekly basis.
2. To expose students to reading and understanding *scholarly papers*. In this class, we will be reading at least one paper per week. Reading papers is different from reading a textbook. It requires practice to understand the general structure of scholarly papers, to understand what information can be found where, to understand the big picture of the paper and to be able to extract all important details. *Reading* papers is also an important first step to *writing* your own research papers. We will be reading papers that

focus on the use different aspects of regression in a particular business application. Our (your!) goal will be to generalize from the specific application, to understand the motivation and need of regression within the given data environment.

3. To jump-start students' *research* in the form of scholarly papers and presentations. There will be no exams in this class. Instead, students will produce a semester paper and a semester presentation. The emphasis in the paper is on scholarly writing. That is, the paper should contain the basic components of scholarly research: introduction (with a motivation for the problem to be researched); basic literature review; description of the statistical models and methodology used; description of the data used; presentation and discussion of the results; conclusions (including potential extensions of the research). Special emphasis should be placed on displaying data and displaying the results in an insightful way. This should be done in the form of well-designed tables and graphs. Part of the grade for the paper and presentation will be determined by how well information is conveyed using statistical techniques. The paper will be based on several databases provided by the instructor. These databases will be accompanied by a set of research questions. Each student will pick one research question for their semester paper and investigate that question based on the techniques learned in this class. The end-of-semester presentation will be based on the semester paper.

STUDENT DELIVERABLES

1. **READING AND CLASS PARTICIPATION:** Every week, there will be assigned readings via handouts and papers. Every student is expected to complete these reading before class and contribute to the class discussion based on these readings. Notice that the weekly papers will be presented by only one student. However, this does not mean that only that student reads the papers. All other students are also expected to read and subsequently contribute to the discussion. Students that are not well prepared and hence cannot contribute to the discussion will be penalized via the participation grade.
2. **WEEKLY STUDENT PRESENTATIONS:** At the beginning of every class period, one student will give a short (15-20 min) presentation on one or more assigned papers. The goal of this presentation is to summarize the basic ideas of a paper from a statistical/methodological point of view. That is, the goal is not to merely summarize the substantive part of a paper. Rather, the idea is to generalize away from the substantive application, to explain the empirical goal of the paper and to explain why regression is important to achieve that goal. All students are required to contribute to the discussion during and after the presentation. The goals of this deliverable are a) to strengthen oral presentation skills (especially of technical topics); b) to be able to generalize away from one particular application and see the "bigger picture;" and c) to improve the learning process via initial self-study followed by class discussion.
3. **HOMEWORK ASSIGNMENTS:** Two homework projects will be assigned. Each project comprises ideas discussed during several class periods. The first project will focus on basic regression ideas; the second project will focus on more advanced regression ideas. One goal of the projects is to build software skills and to implement the concepts learned in class. Another goal is to strengthen written communication skills. Each homework deliverable should read like a short paper. That is, turning-in merely a collection of graphs and software print-outs is **completely unacceptable and will result in a score of zero!** Each deliverable should address each of the following items, organized in separate sections:
 - a. Summary/ overview: briefly introduce the problem and summarize the steps you have taken to solve it.

- b. Data: Using summary statistics, graphs, charts and tables, explore and discuss the data. Point to the aspect of the data that are most important to solve the problem. Point out data anomalies or difficulties and suggest solutions to overcome them.
- c. Model: Describe the regression model(s) that you use to solve the problem. Motivate why you use it. It should become clear how you feed the original data into your model. Are any data manipulations necessary to apply the model?
- d. Results: Discuss the results from applying the model to the data. What economic insight can be drawn from your analysis and how is this helpful to solve the problem?
- e. Conclusions/ limitations: Discuss limitations of your analysis and possible shortcomings.

In each of these sections, you are expected to write full sentences and coherent paragraphs; using only keywords is unacceptable. Also, you are expected to provide graphs and tables that illustrate your data, model and analysis. Each graph or table should be well-formatted and should have a caption! Style matters! Remember, **only include a table/graph, if you also discuss it in the text!** Tables and graphs are never self-explanatory and require a short description of a) what has been done to create the table/graph, and b) what can be learned from that table/graph. Recall that the final product should look and read like a snapshot of a scholarly paper.

Important: While there are only 2 homework assignments, each assignment is rather comprehensive and requires several hours of work. That is, do not wait until the very last moment to solve the assignment. Get started as soon as the homework is posted. As each homework assignment relates to several weeks worth of class material, make a habit out of working on your homework a little bit every week. In the past, students have made the mistake to wait for the very last moment to solve the homework assignment – **do not make the same mistake!**

4. END-OF-SEMESTER PAPER: Every student will prepare a paper on a selected topic. I will make available several rich databases together with a set of different research questions. Each student will pick a research question to work on. The final paper should consist of 10 to 15 pages (11pt, double-spaced, 1 inch margin). Papers have to be consistent with academic standards and have to include introduction, positioning, description of the data, description of the statistical model and techniques used, presentation and discussion of the results, and conclusion. A short list of references is also required. The paper should put special emphasis on displaying data and results using appropriate tables and graphs. The goal of this deliverable is to obtain a potential starting point for a student's research paper.
5. END-OF-SEMESTER PRESENTATION: The end-of-semester papers will be presented during the last week of classes. In the past, faculty from throughout the Smith School have attended these presentations so well-structured and organized presentations are highly desirable.

ACADEMIC INTEGRITY

You must scrupulously abide by the University's Code of Academic Integrity (see <http://www.inform.umd.edu/CampusInfo/Departments/JPO/AcInteg/>). Among other things, the Code prohibits students from cheating on exams, plagiarizing papers, and submitting fabricated documents.

The University of Maryland Honor Pledge reads:

“I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination.”

Unless you are specifically advised to the contrary, this pledge statement should be handwritten and signed on the front cover of all documents submitted for evaluation in this course.

GRADING POLICY

Your final grade for the course is based on your performance outlined above. The weights given to each of these components are as follows.

Weekly Student Presentations	15%
Homework (2X)	40%
Class Participation	15%
End-of-Semester Paper	15%
End-of-Semester Presentation	15%

100%

TENTATIVE SCHEDULE (SUBJECT TO CHANGE)

Week	Date	Topic	Paper Reading & Student Presentation (see Blackboard for weekly papers)	Deliverables Due
1	01/28	Introduction; The software <i>R</i> ; Review of basic statistical concepts: Confidence Intervals and Hypothesis Testing;	----	
2	02/04	Regression basics: Model fitting via Least Squares; Model interpretation; Model evaluation	Paper/Prez 1	
3	02/11	Regression basics: Statistical Inference for Regression; Selection of important predictors; Model assumptions	Paper/Prez 2	
4	02/18	Deviations from the linear model and modeling alternatives (Dummy Variables; Interaction terms)	Paper/Prez 3	
5	02/25	Deviations from the linear model and modeling alternatives (Nonlinear Transformations)	Paper/Prez 4	
6	03/04	Unusual Data, Violation of Model Assumptions and Residual Analysis	Paper/Prez 5	
7	03/11	Collinearity, Principal Components and Variable Selection	Paper/Prez 6	HW1
8	03/18	<i>Spring Break – No Class</i>	----	
9	03/25	Methods for network data: Social Network Analysis	Paper/Prez 7	
10	04/01	Flexible methods for large data sets: Nonparametric and Semiparametric Regression	Paper/Prez 8	
11	04/08	Methods for Time- and Spatially-Dependent Data	Paper/Prez 9	
12	04/15	Methods for combinations of cross-sectional and (time- or spatially-) dependent data: Functional Data Analysis	Paper/Prez 10	
13	04/22	Methods for non-continuous response data: Logit and Choice Models	Paper/Prez 11	
14	04/29	Methods for hierarchical data: Linear Mixed Models, Hierarchical Models	Paper/Prez 12	HW2
15	05/06	END OF SEMESTER PRESENTATIONS	----	Semester Paper