NORTHWESTERN UNIVERSITY

SCHOOL OF PROFESSIONAL STUDIES

PREDICT 410: Regression and Multivariate Analysis Winter 2016

Syamala Srinivasan, Ph.D.

syamala.srinivasan@northwestern.edu

Skype Screen name: syamala.srinivasan

Office Hours: Mondays 7:00 – 9:00 pm CST (If you want to talk to me, you are free to call on Skype).

Teaching Assistant: Nancy Gilbert-Pierce

NancyGilbertPierce2014@u.northwestern.edu

If you send me an email, please put your section number in the subject line (Predict 410-XX).

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] Everitt, B. (2009). *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences*. Boca Raton, FL: CRC Press [ISBN-13: 978-1439807699]

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

[4] 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]

Students wishing to have a more verbose description of linear regression are encouraged to buy Pardoe's book. Pardoe provides a nice verbal description of linear regression topics at the MBA level. Note that this text is entirely focused on the modeling aspects of linear regression and has almost no coverage of the statistical fundamentals. It makes an excellent complement the LRA, but it is not a complete replacement for the LRA book.

[2] Pardoe, I. (2012). *Applied Regression Modeling*. (2nd Edition). New York, NY: Wiley [ISBN-13: 978-1118097281]

Page numbers from Pardoe's book are integrated into the syllabus as recommended reading.

Students wishing to have a more complete alternative to the LRA book than Pardoe's book, or students wishing to have an R presentation of linear regression for additional study, can consider Sheather's regression book. Sheather's book is at the same level as the LRA book. Sheather's book is a nice presentation of linear regression. However, since it focuses on an R presentation of regression it is not appropriate as a primary text for Predict 410. Note that the LRA book provides examples in both SAS and R.

[3] Sheather, S. (2009). A Modern Approach to Regression with R. Springer [ISBN-13 978-1441918727]

Page numbers from Sheather's book are not integrated in the syllabus as recommended reading. As an R based text this book might cause confusion for anyone who is not already an R user, hence it is not an 'officially' recommended book, like Pardoe's book. Any student interested in using this book as a supplement should match chapter titles.

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.
The library course reserves will show this reading as a journal article from the ***Journal of Targeting, Measurement, and Analysis for Marketing*****.**

[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

The Northwestern University School of Professional Studies provides access to the Social Sciences Computing Cluster (SSCC). We will use a SAS web application called SAS Studio on this environment. Students will be instructed on how to obtain a SSCC account, install the SSCC VPN, and access the SAS Studio web application.

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.

Evaluation

The student's final grade will be determined from a total of 650 possible points as follows:

• Participation	(100 possible points, 10 points each session)
• Proctored Final Exam	(50 possible points, proctored exam)
• Take-Home Final Exam	(50 possible points, unproctored exam)
• Assignments	(350 possible points from homework assignments)
• Quizzes	(100 possible points, 5 quizzes for 20 points each)

Final Exam: The final exam will be administered in two parts – a Proctored part that will require ProctorU and an unproctored part. Both exams will be administered through the course site, and both exams must be completed in continuous, but independent, exam sessions scheduled at the student's convenience.

Assignments: Predict 410 will have eight assignments for students to complete. All the assignments will be graded by the instructor for the total of 350 points.

The graded assignments will be:

Assignment #1	20 points
Assignment #2	25 points
Assignment #3	60 points
Assignment #4	50 points
Assignment #5	80 points
Assignment #6	80 points
Assignment #7	20 points
Assignment #8	15 points

Quizzes: Each quiz must be completed by 11:55 p.m. Sunday night (Central Time). All quizzes are open book and open notes.

Grading Scale

A	= 93–100%	(604–650 points)
A-	= 90–92%	(585–603 points)
B+	= 87–89%	(565–584 points)
B	= 83–86%	(539–564 points)
B-	= 80–82%	(520–538 points)
C+	= 77–79%	(500–519 points)
C	= 73–76%	(474–499 points)
C-	= 70–72%	(455–473 points)
F	= 00–69%	(000–454 points)

Discussion Board Etiquette

The purpose of the discussion boards is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency is not unimportant, but content of the message is paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism.

Proctored Assessment

There is a proctored assessment requirement for this class. Please see the Assignments section in the course site for more information. The final exam must be proctored. Students are encouraged to use the ProctorU software.

Attendance

This course will not meet at a particular time each week. All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this class. Please note that any scheduled synchronous or "live" meetings are considered supplemental and optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Late Work

Students must provide written notification of late work 24 hours prior to the deadline. One grace day is allowed for those who provide late work notification. Only one grace day without reduction of points is allowed. A 25% reduction is applied to the grade for every 12 hours late. No negative points are applied.

Learning Groups

Student study groups will be utilized in this course as a means to foster a collaborative learning environment. The study groups are facilitated through the use of Adobe Connect, much in the same manner as our sync sessions are held. An Adobe Connect link will be provided by the instructor via the course site.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit <www.scs.northwestern.edu/student/issues/academic_integrity.cfm>.

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism, by visiting <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online.

Other Processes and Policies

Please refer to your SCS student handbook at <www.scs.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies.

Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion.
When changes are made, students will be notified via an announcement in the course site.

Note: All courses operate on a Monday to Sunday schedule.

Session 1 – Complete By Sun 1/10

Topic: Exploratory Data Analysis and Simple Linear Regression

Learning Objectives

After this session, the student will be able to:

- Understand the importance and the role of exploratory data analysis in statistical modeling.
- Recognize the difference between exploratory data analysis and data management.
- Use the appropriate data summaries and statistical graphics for exploratory data analysis.
- Perform an exploratory data analysis for the simple linear regression model.
- Fit and interpret a simple linear regression model.
- Perform a goodness-of-fit analysis to verify the model assumptions for the simple linear regression model.

Assigned Reading

[1] Fox (2008) Chapters 2-3, pp. 11-49
[2] LRA Chapters 1–2, pp. 1–66

Optional:
[3] Pardoe (2012) Chapters 1-2, pp. 1-82
[4] Everitt (2009) Chapters 1-3, pp. 1-80

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – None

Assignments – Assignment #1 (20 Points)

Assignment #1: Getting to Know Your Data is due Sunday at 11:55 p.m. (Central Time).

Sync Session – First sync session will be held in Week #1 in scheduled time slot.

Session 2 – Complete By Sun 1/17

Topic: Multiple Linear Regression

Learning Objectives

After this session, the student will be able to:

- Fit and interpret a multiple linear regression model.
- Compute and interpret the statistical tests associated with the multiple linear regression model.
- Understand the analysis of variance table and the associated metrics and tests of significance for the multiple linear regression model.
- Interpret R-Squared and Adjusted R-Squared and use them for model comparison.
- Understand the appropriateness of using a fitted regression model to predict out-of-sample.
- Understand the difference in the computation and the interpretation of a confidence interval on a fitted value and a prediction interval.
- Perform a goodness-of-fit analysis to verify the model assumptions of the multiple linear regression model.

Assigned Reading

[1] LRA Chapter 3, pp. 67–128

Optional:
[2] Pardoe (2012) Chapter 3, pp. 83-136
[3] Everitt (2009) Chapter 3, pp. 81-102

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – Quiz #1 (20 Points)

Assignments – Assignment #2 (25 points)

Assignment #2: Regression Model Building is due Sunday at 11:55 p.m. (Central Time).

Sync Session - None

Session 3: – Complete By Sun 1/24

Topic: Model Validation

Learning Objectives

After this session, the student will be able to:

- Use residual analyses to assess the goodness-of-fit of a fitted regression model.
- Define the statistical concept of an outlier, understand how to detect outliers, and understand how outliers can affect the regression fit.
- Define the statistical concept of leverage, understand how leverage is computed, understand how to use leverage estimates to detect outliers, and understand how leverage affects parameter estimation and residual computation in linear regression.
- Validate a regression model for the purposes of statistical inference.
- Validate a regression model for the purpose of predictive modeling.
- Validate a regression model for specific application use.
- Differentiate between applications that require a statistical model validation and applications that require an operational or business validation.

Assigned Reading

[1] LRA Chapter 4, pp. 129–170
[2] LRA Chapter 6, pp. 211-222
[3] LRA Chapter 11, pp. 372-388

Handouts

[1] Best Practice of Modeling Process in a Business Environment

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – Quiz #2 (20 Points)

Assignments – Assignment #3 (60 Points)

Assignment #3: Data Analysis and Regression is due Sunday at 11:55 p.m. (Central Time).

Sync Session - None

Session 4 – Complete By Sun 1/31

Topic: Variable Transformations

Learning Objectives

After this session, the student will be able to:

- Differentiate between cases where variable transformations are needed a priori versus cases where variable transformations are needed empirically to improve the model fit of a fitted regression model.
- Apply the Box-Cox family of transformations for transformations to normality.
- Use indicator variables to include categorical variables as predictor variables in a regression model.
- Use indicator variables to discretize a continuous predictor variable.
- Use indicator variables to create complex interactions and sophisticated model specifications.
- Interpret indicator variables in the context of a specified regression model.

Assigned Reading

[1] LRA Chapter 5.1-5.4, pp. 171-187
[2] LRA Chapter 8, pp. 260-284

Optional:
[3] Pardoe (2012) Chapter 4, pp. 137-188

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – None

Assignments – Assignment #4 (50 Points)

Assignment #4: Statistical Inference in Linear Regression is due Sunday at 11:55 p.m. (Central Time).

Sync Session - Second sync session will be held in Week #4 in scheduled time slot.

Session 5 – Complete By Sun 2/7

Topic: Automated Variable Selection

Learning Objectives

After this session, the student will be able to:

- Compute a nested F-test and understand how it is used by the forward, backward and stepwise variable selection algorithms to select an optimal subset of the predictor variables.
- Describe the pros and cons of the stepwise variable selection algorithm.
- Use different statistical metrics in automated variable selection algorithms to affect the model selection.
- Understand how penalized measures such as Mallow's Cp, AIC, and BIC are defined and how to use them in automated variable selection to provide a trade-off between model fit and model complexity.
- Use automated variable selection as an exploratory data analysis tool.
- Use automated variable selection as part of the statistical modeling process as a means of model identification.

Assigned Reading

[1] LRA Chapter 10, pp. 327-371

[2] Ratner (2012) Chapter 10, pp. 177-194 – Available through the library reserves. This reading will show up in the library course reserves as a journal article from the *Journal of Targeting, Measurement, and Analysis for Marketing*.

Optional:
[3] Pardoe (2012) Chapter 5, pp. 189-242

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – Quiz #3 (20 Points)

Assignments – Assignment #5 (80 Points)

Assignment #5: Automated Variable Selection, Multicollinearity, and Predictive Modeling is due Sunday at 11:55 p.m. (Central Time).

Sync Session - None

Session 6 – Complete By Sun 2/14

Topic: Multicollinearity and Principal Components Analysis

Learning Objectives

After this session, the student will be able to:

- Define multicollinearity and describe how it affects regression estimates and inference.
- Use the Variance Inflation Factor (VIF) as a model diagnostic for multicollinearity.
- Take remedial actions to correct or minimize multicollinearity and its effects.
- Perform a principal components analysis and determine how many principal components to keep.
- Understand how principal components are computed and the roles of eigenvalues and eigenvectors in their computation and use.
- Use principal components analysis as a tool for dimension reduction.
- Use principal components analysis as a tool to correct for multicollinearity.

Assigned Reading

[1] LRA Chapter 9, pp. 285-326
[2] Everitt (2009), Chapters 9-10, pp. 169-210

Optional:
[3] Morrison (2004), Chapter 6, pp. 264-316

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – None

Assignments – Assignment #6 (80 Points)

Assignment #6: Principal Components in Predictive Modeling is due Sunday at 11:55 p.m. (Central Time).

Sync Session - None

Session 7 – Complete By Sun 2/21

Topic: Exploratory Factor Analysis

Learning Objectives

After this session, the student will be able to:

- Define Thurstone’s common factor model and understand the concept of a simple factor structure.
- Define factor analysis as a statistical model and understand its statistical assumptions.
- Understand the different methods of estimation for factor analysis and how to estimate them in SAS.
- Fit, interpret, and validate a factor analysis.
- Apply factor rotations to increase factor interpretation.
- Use the output from the SAS procedure PROC FACTOR to decide how many common factors to estimate.
- Discuss the limitations of factor analysis.
- Understand the conceptual differences between exploratory factor analysis and principal components analysis.

Assigned Reading

[1] Everitt (2009) Chapter 11, pp. 211-238

Optional:
[2] Morrison (2004), Chapter 7, pp. 317-370

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students’ comments is graded.

Quiz – Quiz #4 (20 Points)

Assignments – Assignment #7 (20 points)

Assignment #7: Factor Analysis is due Sunday at 11:55 p.m. (Central Time).

Sync Session - Third sync session will be held in Week #7 in scheduled time slot.

Session 8 – Complete By Sun 2/28

Topic: Cluster Analysis

Learning Objectives

After this session, the student will be able to:

- Use statistical graphics to visualize clusters.
- Understand the various similarity measures, how they can affect cluster formulation, and when one measure may be preferred other another.
- Describe the differences between hierarchical and non-hierarchical clustering techniques.
- Select the number of clusters based on clustering metrics.
- Use cluster analysis to perform population segmentation.
- Discuss how segmentation can be used in predictive modeling and how it can affect the results of a predictive model.
- Discuss the limitations and practical caveats of cluster analysis.

Assigned Reading

[1] Everitt (2009) Chapter 12, pp. 239-260

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – None

Assignments – Assignment #8 (15 Points)

Assignment #8: Cluster Analysis is due Sunday at 11:55 p.m. (Central Time).

Sync Session - None

Session 9 – Complete By Sun 3/6

Topic: Multivariate Data Analysis

Learning Objectives

After this session, the student will be able to:

- Recognize principal components, factor analysis, and cluster analysis as a class of statistical problems called 'unsupervised learning' problems.
- Use principal components, factor analysis, and cluster analysis to perform segmentation.
- Understand how the dimension of the data affects cluster formulation, or the curse of dimensionality.
- Use principal components in conjunction with cluster analysis.
- Use factor analysis in conjunction with cluster analysis.

Assigned Reading

[1] Everitt and Dunn (2001) Chapter 3, pp. 48-73
[2] Everitt and Dunn (2001) Chapter 6, pp. 125-160
[3] Everitt and Dunn (2001) Chapter 12, pp. 271-290

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – Quiz #5 (20 Points)

Assignments – None

Sync Session - Final sync session will be held in Week #9 in scheduled time slot.

Session 10: Exam Week – Complete By Sun 3/13

Learning Objectives

After this session, the student will be able to:

- No new learning objectives will be introduced.

Assigned Reading - None

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded.

Quiz – None

Assignments – Final Exam (100 Points)

Proctored Final Exam	(50 points)
Take-Home Final Exam	(50 points)

Both components of the Final Exam are due **Sunday, March 13, 2016** at 11:55 p.m. (Central Time).

Both the proctored portion of the final exam and the unproctored portion of the final exam will be open from **March 7, 2016** (12:00 AM) to **March 13, 2016** (11:55 PM). The proctored portion of the final exam will be proctored by ProctorU and must be taken in one continuous two hour sitting. The unproctored portion of the final exam has one hour time limit for one continuous sitting.

Sync Session - None